Evaluation of Parameter Fine-Tuning with Transfer Learning for Osteoporosis Classification in Knee Radiograph

Usman Bello Abubakar[1]
Computer Science
Baze University
Abuja, Nigeria

Moussa Mahamat Boukar[2]
Computer Science
Nile University of Nigeria
Abuja, Nigeria

Steve Adeshina[3]
Computer Engineering
Nile University of Nigeria
Abuja, Nigeria

***Abstract*—Osteoporosis is a bone disease that raises the risk of fracture due to the density of the bone mineral being low and the decline of the structure of bone tissue. Among other techniques, such as Dual-Energy X-ray Absorptiometry (DXA), 2D x-ray pictures of the bone can be used to detect osteoporosis. This study aims to evaluate deep convolutional neural networks (CNNs), applied with transfer learning techniques, to categorize specific osteoporosis features in knee radiographs. For objective labeling, we obtained a selection of patient knee x-ray images. The study makes use of the Visual Geometry Group Deep (VGG-16), and VGG-16 with fine-tuning. In this work, the deployed CNNs were assessed using state-of-the-art metrics such as accuracy, sensitivity, and specificity. The evaluation shows that fine-tuning enhanced the VGG-16 CNN's effectiveness for detecting osteoporosis in radiographs of the knee. The accuracy of the VGG-16 with parameter fine-tuning was 88% overall, while the accuracy of the VGG-16 without parameter fine-tuning was 80%.**

***Keywords*—Osteoporosis; transfer learning models; convolutional neural network; fine-tuning**

I. Introduction

Osteoporosis is a severe illness common in about 9% of citizens, above 50 years, in the United States [1] and about 200 million women worldwide. One in three people in developed nations may experience an osteoporotic compression fracture (OCF) [1]. The likelihood of recurrent fractures greatly increases after the initial fracture [2] [3] [4]. Even one OCF is linked to a greater death rate and a lower quality of life [5].

Osteoporosis, which is defined as porous bone, is a condition in which the mass of the bone is low and the bone tissues have undergone microarchitectural deterioration. Osteoporosis increases fracture risk of the wrist, hip, and spine, among other bones, and lowers bone mineral density (BMD). Additionally, osteoporosis alters the quantity and type of proteins in bones. Osteoporotic fractures are described as those that happen at a site where there is low BMD and are more likely to happen beyond the age of roughly 50 [6] [7].

Every individual irrespective of gender and race could be affected by the disease and as the population ages, its prevalence would also increase. Among specialists, it is known as a silent bone disease because its symptoms are not spotted before a fracture and thus, pose threats to a patient by inducing other secondary bone problems like arthritis and the likes [8]. In the skeletal system, there is a continuous activity of bone tissues been lost by resorption, and also bone tissues have been rebuilt back by formation. The system is said to be at a bone loss when bone tissue formation is less than bone tissue resorption [9].

It has long been believed that deep learning is effective at learning feature categorization from medical images [10]. Deep Learning (DL) classifiers utilize high-dimensional features to improve the performance of DL networks in object detection and image classification. Machine Learning (ML) techniques, in contrast to DL techniques, rely on explicitly categorized features [11].

Deep CNNs have been proved to be efficient tools for categorizing images, but they are difficult to employ with medical radiographic image data since they require a large amount of training data. Transfer learning is recognized as an efficient method in training deep CNNs when the dataset is small to prevent overfitting [12].

We use a dataset of knee radiographs (or knee X-rays) to apply and assess deep transfer learning algorithms for classifying osteoporosis. This work objectively assessed the impact of parameter fine-tuning on a transfer learning deep CNN model's performance for identifying knee radiograph pictures based on the BMD value (T-score).

II. Related Work

Authors in [13] performed a comparison of classification systems for osteoporosis prediction using feature selection based on wrappers. As classification methods, multilayer feed-forward neural network (MFNN), Naive Bayes, and logistic regression were employed. Single Nucleotide Polymorphisms (SNPs), age, menopause, and BMI of Taiwanese women were all included in the dataset utilized for the study.

The three classifiers, utilizing SNP, were tested using a 10-fold cross-validation method both with feature selection and without feature selection. Without using wrapper-based feature selection, the Area under Curve (AUC) for the MFNN was 0.489. The AUC for naive Bayes was 0.462, and the AUC for logistic regression prediction was 0.485 [13].

The performance metric for classifiers utilizing a wrapper-based strategy yielded an AUC of 0.631 for MFNN, AUC of 0.569 naïve Bayes, and AUC of 0.620 for logistic regression models [13]. The experimental results demonstrated that the MFNN model with the wrapper-based technique was the most accurate predictive model for predicting disease susceptibility in Taiwanese women based on the complicated interplay between osteoporosis and SNPs. The findings reveal that the proposed technology can help patients and clinicians make better decisions based on clinical data such as SNP genotyping data [13].

The study proposed by [14] investigates whether adding clinical information improves diagnosis when compared to images alone when using deep learning. 1131 images from patients who had skeletal bone mineral density testing and hip radiography at the same general hospital between 2014 and 2019 were gotten. From hip radiographs, five convolutional neural networks (CNN) models were employed to assess osteoporosis [14]. Adding clinical values increased accuracy, sensitivity, and specificity.

Using only hip radiograph images, without clinical covariates, GoogleNet and EfficientNet b3 models displayed the highest levels of model performance. EfficientNet b3 demonstrated the best accuracy, sensitivity, and other metric core among the five ensemble models when patient factors were taken into account [14]. Increasing clinical covariates increased the accuracy of the deep learning models [14].

The authors in [15] revealed that dental panoramic radiographs can be used to accurately diagnose osteoporosis using CNNs. Additionally, integrating patient factors in common clinical contexts enhanced all predictions' performance measures in comparison to using the image-only mode. The study hypothesized that advanced inference, which is possible by deep learning, which, in turn, simultaneously takes important information about clinical factors into account that cannot be determined from dental panoramic X-ray images alone, led to an increase in diagnosis precision [15].

Various implementations of EfficientNet and ResNet were employed in the study by the authors. The most accurate ResNet and EfficientNet techniques, respectively, were ResNet-152 and EfficientNet-B7. However, EfficientNet-b7 obtained better results than other CNN models [15]. Gradient-weighted Class Activation Mapping (Grad-CAM) was used to visualize learning. ResNet concentrated on the cortical bone at the base of the jaw. Contrarily, EfficientNet concentrated on the area above the cortical bone as well as the cortical bone at the bottom border of the jaw.

The authors in [16] developed a cutting-edge, reliable bone disease prediction model based on recognized risk factors. Then it was feasible to discover the early risk factors for determining the beginnings of bone disorders using Pre-training and fine-tuning. The most significant risk factors are coupled with model parameters during the pre-training phase to calculate contrastive divergence, which minimizes record size.

Using the ground truth values "g1" and "g2," where "g1" stood for osteoporosis and "g2" for a rate of bone loss, the outcomes of the preceding phase were compared [16]. The model was produced using a Deep Belief Network (DBN), and it was then contrasted with models made both before and after essential feature identification. The study's conclusions indicated that adding pertinent variables might improve the predictive model's performance.

The authors in [17] built a model to predict the risk of osteoporosis using supervised machine learning. The study made public the variables that experts considered while determining the risk of osteoporosis. Developing a predictive model for the identification of people in Nigeria who are at risk for osteoporosis was the study's main objective. The supervised machine learning techniques Nave Bayes (NB) classifier and Multi-layer Perceptron were utilized to develop the predictive model for osteoporosis risk (MLP). The identification and data collection from patients in Nigerian hospitals found that there were 20 risk markers, including CD4 count levels classified as low, moderate, and high risk [17]. According to their finding, NB got 71.4% accuracy while the MLP had the best got 100%.

There has been a scarcity in the use of DL to interpret and predict osteoporosis from a knee radiograph. This research aims at filling this gap in the existing knowledge that points to the need for further understanding and investigation of osteoporosis prediction using DL from knee radiographic images.

III. METHODS

A. Research Design

This research tries to classify osteoporosis in knee radiographs. To replicate the osteoporosis diagnostic range in the DXA approach, we employed a segmented dataset. In addition, the Keras Deep Learning (DL) packages were employed for data normalization and augmentation. The diagnosis of osteoporosis from knee radiographs was performed using the VGG-16 transfer learning deep neural network. We examined the accuracy of the osteoporosis prognostic diagnostic using the transfer learning model with and without parameter fine-tuning using cutting-edge performance metrics.

B. Dataset

The dataset, published in August 2021, was gotten from Mendeley data uploaded by [18]. The dataset images were statistically augmented (i.e. increased) using data augmentation in python. Fig. 1 shows two images from the dataset indicating osteoporosis cases and normal cases.

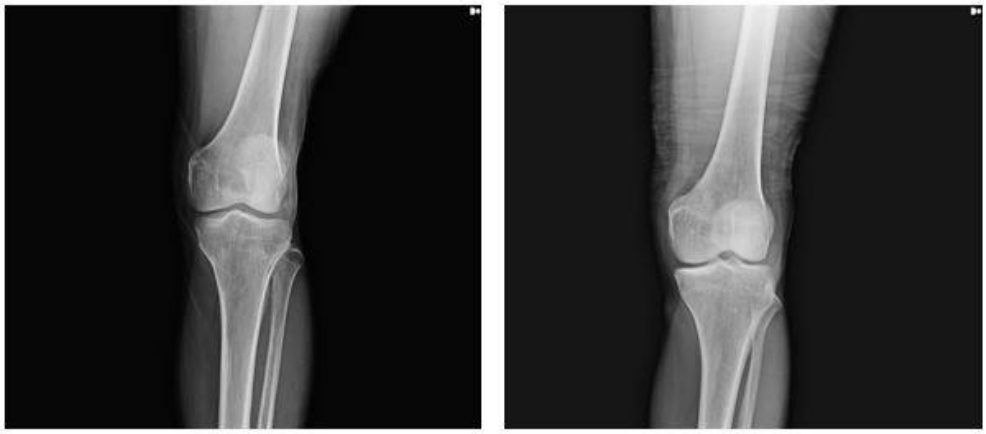

Fig. 1. Osteoporosis Case and Normal Case [18].

The dataset, after statical augmentation using python augmentation functions, comprises 323 normal knee radiograph images and 323 osteoporotic knee radiograph images of patients. Table I shows the splitting of image data into train, test, and validation data.

TABLE I. IMAGE DISTRIBUTION

Class	Total	Training	Testing
Normal (0)	323	259	65
Osteoporosis (1)	323	259	65

C. Grayscale Conversion

The dataset consists of images in Red Green Blue (RGB) format. A three-dimensional byte array (i.e., RGB image) stores a color value for each pixel. RGB format increases the complexity of training the model. Grayscale (i.e., black and white images) are preferred as they simplify computational complexity.

The modality of our research is based on knee x-ray data and thus, in an x-ray, color is irrelevant to diagnosis. Due to this reason, and the fact that grayscale images are easier to train a deep learning network, the images were converted from RGB to grayscale using the OpenCV python library.

D. Data Normalization

It is the process of converting image data pixels to a predetermined range : (0, 1) or (-1, 1). The pixel values in most images range from 0 to 255. Training a deep neural network with large integer values can interfere with or slow down the learning process. Therefore, picture normalization is a recommended practice: pixel values range between 0 and 1.

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

E. Data Augmentation

When working with deep learning models, it is paramount to ensure that the model gets a sufficient amount of training data. Data augmentation is the application of various changes to original images, resulting in several altered copies of the same image. Each replica, however, differs from the others in some ways due to the augmentation procedures used.

For this study, augmentation was done using Keras ImageDataGenerator in python. Itemized below are some of the techniques applied:

1) Standardization
2) Rotation
3) Shifts
4) Brightness changes, among others

The Keras ImageDataGenerator class is intended to give real-time data augmentation, which is said to be its key advantage. Every epoch, the model is given fresh versions of the images due to the ImageDataGenerator python class.

F. Transfer Learning Model Used (VGG-16)

In this work, two CNN study groups were used: VGG16 and the parameter fine-tuning model from VGG16. The difference between the two implementations is that the latter used parameter fine-tuning while the former did not. This was performed by unfreezing a couple of the original model's top levels and training the newly added classifier layers alongside the base model's final layers. The schematic diagram for the two transfer learning models used in this work is depicted in the block diagrams in Fig. 2 and Fig. 3.

The VGG16 architecture was chosen since it had been widely adopted and considered cutting-edge in image classification applications trained on a large dataset [10] [19].

G. Training the Model

Five folds were randomly selected from the training dataset of the chosen images. This prevented bias or overfitting while performing a five-fold cross-validation on the model training. The dataset was split into independent training and validation sets within each fold using an 80 to 20 split. A validation set that was completely different from the other training folds was chosen to assess the training state throughout training. Once one model training phase was complete, the other independent fold was utilized as a validation set, and the previous validation set was recycled as part of the training set to evaluate the model training. Fig. 4 shows a five-fold cross-validation done in this study.

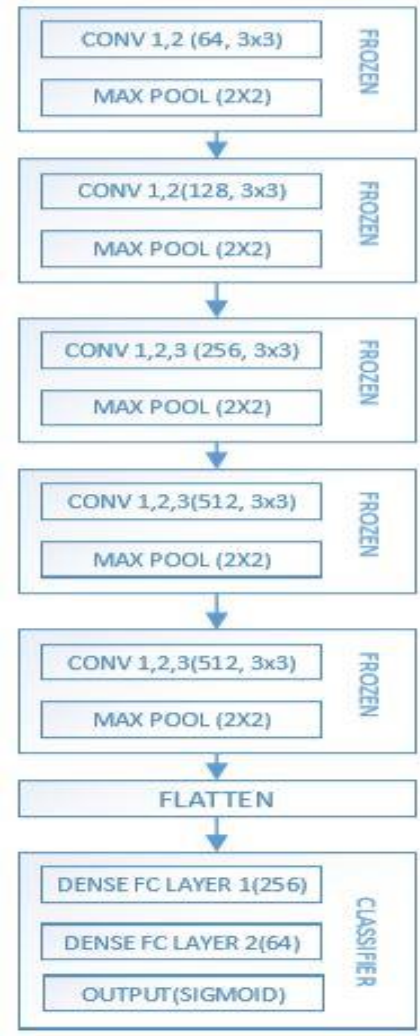

Fig. 2. VGG-16 without Parameter Fine-Tuning.

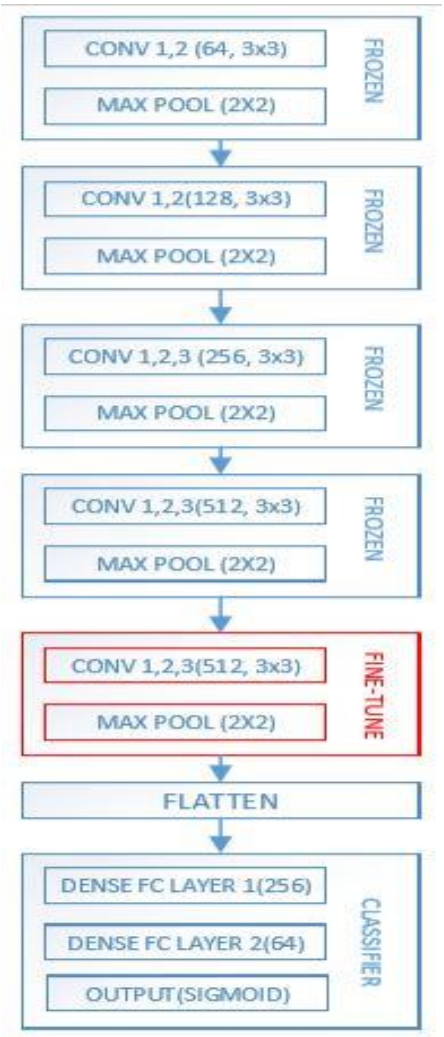

Fig. 3. VGG-16 with Parameter Fine-Tuning.

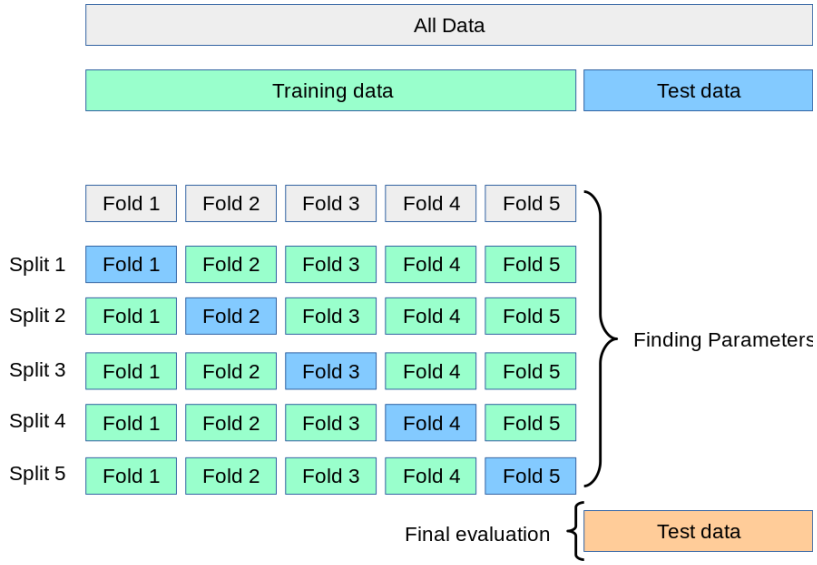

Fig. 4. Overview of 5-Fold Cross Validation.

This process of cross-validation was repeated for the VGG-16 without parameter fine-tuning and for the VGG-16 with parameter fine-tuning. The Google colabs Graphics Processing Unit (GPU) was used to train and test all models. The Keras library and TensorFlow were used throughout the process of applying the transfer learning deep learning models.

IV. RESULTS

A. Performance Metrics

The following metrics were established for each model to fully assess its performance: (1) sensitivity, (2) specificity, (3) accuracy, (4) precision, and (5) F1-score. The formula for the specified metrics is expressed below.

$$\text{Sensitivity} = \frac{true\ positive}{true\ positive + false\ negative} \quad (1)$$

$$\text{Specificity} = \frac{true\ negative}{true\ negative + false\ negative} \quad (2)$$

$$\text{Accuracy} = \frac{true\ negative + true\ positive}{all\ cases} \quad (3)$$

$$\text{Precision} = \frac{true\ positive}{true\ positive + false\ positive} \quad (4)$$

$$\text{F1} = 2 \times \frac{precision \times recall}{precision + recall} \quad (5)$$

B. Confusion Matrix

A method for summarizing a classification algorithm's performance is the confusion matrix (CM). In addition to giving insight into the mistakes the classifier is making, it also reveals the specific mistakes that are occurring. The confusion matrix helps to overcome the limitation of using classification accuracy alone. Fig. 5 and Fig. 6 show the confusion matrix for the VGG-16 model without parameter fine-tuning and the VGG-16 model with parameter fine-tuning respectively.

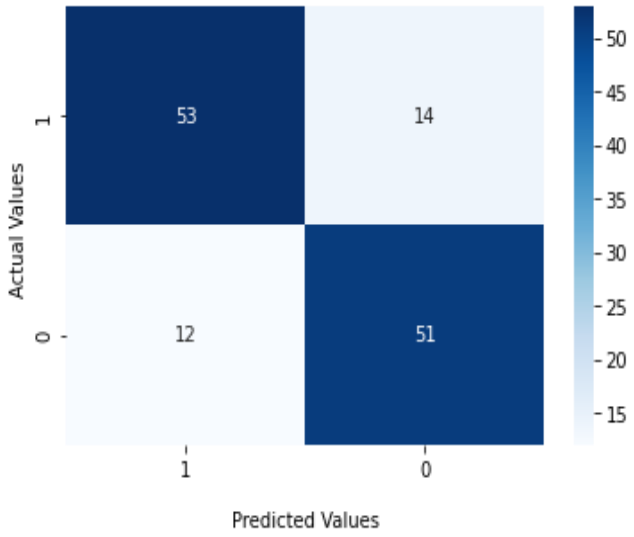

Fig. 5. Confusion Matrix for VGG-16 without Parameter Fine-Tuning.

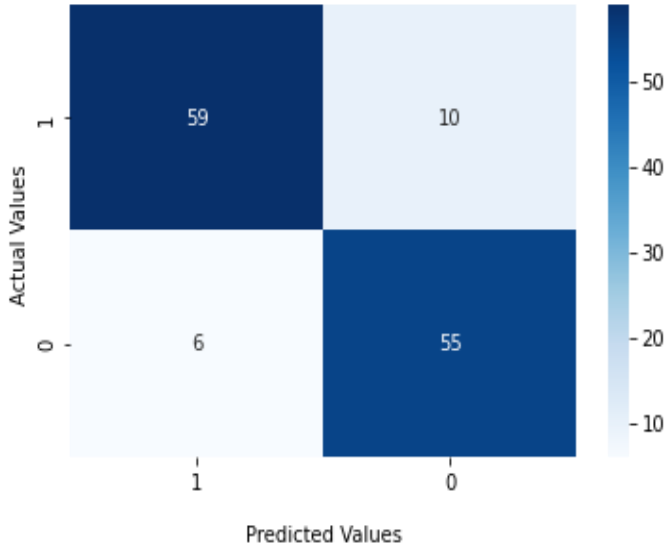

Fig. 6. Confusion Matrix for VGG-16 with Parameter Fine-Tuning.

C. Prediction Performance

The osteoporosis patient knee x-rays dataset has been tested using the CNN models utilized in this work. The dataset was divided into training and testing portions in an 80:20 ratio for all transfer learning models. The overall accuracy obtained for the two classifiers on the dataset is summarized in Table II. Each model underwent 50 epochs of training. For all models, as the loss metric, binary_crossentropy was used as the dataset target has two classes (i.e., binary classification problem). RMSprop is the chosen optimizer, and its learning rate is 0. 001.

The Keras evaluate function was invoked on the compiled model with the test data as an argument to evaluate the accuracy of the models. Table III provides a comparison of our work with similar works. Fig. 7 shows a chart visually depicting the performance difference between the two implementations of the VGG-16 transfer learning model.

D. Algorithm Justification

The justification for choosing VGG-16 architecture was that it had been widely adopted and recognized as state-of-the-art in both general and medical image classification tasks but has not readily been applied to osteoporosis classification from patient knee radiographs. Additionally, VGG-16 has been trained on large-scale datasets, so that a transfer learning approach could be adopted for large-scale image recognition.

The reason for using parameter fine-tuning is that research shows it boosts the performance of a deep learning model over random initialization [20].

E. Dataset Justification

The reason for choosing the knee radiograph dataset is because deep learning research on osteoporosis classification using knee x-ray is still relatively scarce.

F. Limitations of the Study

A deep learning model requires massive amounts of data to be efficient. The number of training observations in the dataset was not large enough and hence poses a limitation to the study. However, data augmentation was applied to mitigate such limitations.

G. Recommendation and Future Work

The perception based on the findings stipulates that overfitting in transfer learning due to few data samples can be avoided using certain techniques: cross-validation, data augmentation, and parameter fine-tuning. Findings also show that parameter fine-tuning in transfer learning can be used to significantly increase the accuracy, sensitivity, specificity, precision, and F1 of a deep learning model.

Osteoporosis is caused not just by low bone mineral density, but also by other factors such as age, gender, weight, height, and so on. These are clinically important risk factors for osteoporosis. For future work, we would like to extend our methods by adding patient variables such as age, and gender, amongst others, as clinical covariates to create an ensemble model with the transfer learning models

TABLE II. RESULTS OBTAINED

	Ac	Se/Re	Sp	Pr	F1
VGG-16 without Fine-Tuning	0.80	0.82	0.81	0.79	0.80
VGG-16 with Fine-Tuning	0.88	0.91	0.90	0.86	0.88

*AC: ACCURACY, SE: SENSITIVITY, RE: RECALL, SP: SPECIFICITY, PR: PRECISION

TABLE III. COMPARISON WITH OTHER WORKS

	Classifier	Accuracy	Sensitivity/Recall	Specificity
Our Paper	VGG-16	0.80	0.82	0.81
Our Paper	VGG-16: Fine-Tuning	0.88	0.91	0.90
[14]	ResNet-18	0.79	0.86	0.86
[12]	CNN with 3 layers	0.66	0.68	0.65
[15]	ResNet-50	0.83	0.75	0.90

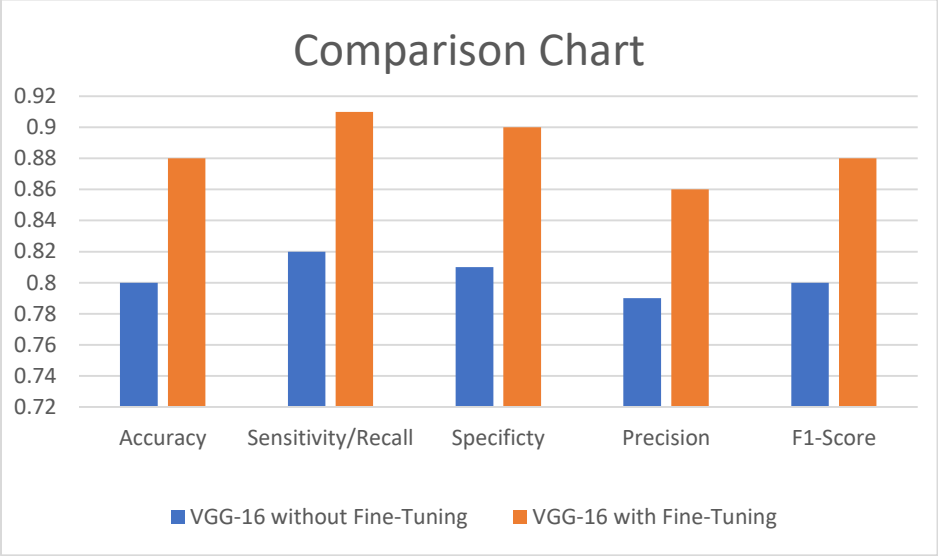

Fig. 7. Comparison Chart.

V. CONCLUSION

In circumstances when there is a small training dataset, this study demonstrates the efficacy of deep CNN tuning and transfer learning for detecting osteoporosis in knee x-ray images. On networks that have already been trained for the categorization of osteoporosis, we have used the VGG-16 transfer learning technique. According to the experimental findings, the fine-tuning technique enabled transfer learning to obtain an overall accuracy of 88%, which was higher than that of 80% achieved by transfer learning without fine-tuning.

The results show that parameter fine-tuning in transfer learning can be used to significantly increase the accuracy, sensitivity, specificity, precision, and F1 of a deep learning model. For future work, we would like to extend our methods by creating an ensemble approach of adding patient clinical covariates to classify osteoporosis with VGG-16 from knee radiograph.

This research was broken into several parts: Introduction, related works, methods, results, and conclusion. The method section provided details as to how the dataset was acquired, the augmentation techniques used, the grayscale conversion of images from RGB to grayscale, the cross-validation split used, and the transfer learning model applied. The results section depicted some state-of-the-art deep learning evaluation metrics used to evaluate the transfer learning variations of the VGG-16 model used.

REFERENCES

[1] N. C. Wright et al., "The recent prevalence of osteoporosis and low bone mass in the United States based on bone mineral density at the femoral neck or lumbar spine," Journal of Bone and Mineral Research, vol. 29, no. 11, pp. 2520–2526, Nov. 2014, doi: 10.1002/jbmr.2269.

[2] A. B. Hodsman, W. D. Leslie, J. F. Tsang, and G. D. Gamble, "10-year probability of recurrent fractures following wrist and other osteoporotic fractures in a large clinical cohort: an analysis from the Manitoba Bone Density Program," Arch Intern Med, vol. 168, no. 20, pp. 2261–2267, Nov. 2008, doi: 10.1001/ARCHINTE.168.20.2261.

[3] S. Roux et al., "The World Health Organization Fracture Risk Assessment Tool (FRAX) underestimates incident and recurrent fractures in consecutive patients with fragility fractures," J Clin Endocrinol Metab, vol. 99, no. 7, pp. 2400–2408, 2014, doi: 10.1210/JC.2013-4507.

[4] C. M. Robinson, M. Royds, A. Abraham, M. M. McQueen, C. M. Court-Brown, and J. Christie, "Refractures in patients at least forty-five years old. a prospective analysis of twenty-two thousand and sixty patients," J Bone Joint Surg Am, vol. 84, no. 9, pp. 1528–1533, 2002, doi: 10.2106/00004623-200209000-00004.

[5] J. R. Center, T. v. Nguyen, D. Schneider, P. N. Sambrook, and J. A. Eisman, "Mortality after all major types of osteoporotic fracture in men and women: an observational study," Lancet, vol. 353, no. 9156, pp. 878–882, Mar. 1999, doi: 10.1016/S0140-6736(98)09075-8.

[6] J. A. Kanis, A. Oden, O. Johnell, C. de Laet, B. Jonsson, and A. K. Oglesby, "The components of excess mortality after hip fracture," Bone, vol. 32, no. 5, pp. 468–473, 2003, doi: 10.1016/S8756-3282(03)00061-9.

[7] O. Johnell and J. A. Kanis, "An estimate of the worldwide prevalence and disability associated with osteoporotic fractures," Osteoporosis International, vol. 17, no. 12, pp. 1726–1733, Dec. 2006, doi: 10.1007/S00198-006-0172-4.

[8] T. Sozen, L. Ozisik, and N. Calik Basaran, "An overview and management of osteoporosis," European Journal of Rheumatology, vol. 4, no. 1, pp. 46–56, Mar. 2017, doi: 10.5152/EURJRHEUM.2016.048.

[9] B. L. Riggs et al., "Changes in bone mineral density of the proximal femur and spine with aging. Differences between the postmenopausal and senile osteoporosis syndromes," Journal of Clinical Investigation, vol. 70, no. 4, pp. 716–723, 1982, doi: 10.1172/JCI110667.

[10] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/J.MEDIA.2017.07.005.

[11] U. Bello Abubakar, M. Mahamat Boukar, and S. Dane, "Review of Swarm Fuzzy Classifier and a Convolutional Neural Network with VGG-16 Pre-Trained Model on Dental Panoramic Radiograph for Osteoporosis Classification", Accessed: Jul. 26, 2022. [Online]. Available: www.jrmds.in

[12] K. S. Lee, S. K. Jung, J. J. Ryu, S. W. Shin, and J. Choi, "Evaluation of Transfer Learning with Deep Convolutional Neural Networks for Screening Osteoporosis in Dental Panoramic Radiographs," Journal of Clinical Medicine, vol. 9, no. 2, Feb. 2020, doi: 10.3390/JCM9020392.

[13] H. W. Chang, Y. H. Chiu, H. Y. Kao, C. H. Yang, and W. H. Ho, "Comparison of classification algorithms with wrapper-based feature selection for predicting osteoporosis outcome based on genetic factors in a Taiwanese women population," International Journal of Endocrinology, vol. 2013, 2013, doi: 10.1155/2013/850735.

[14] N. Yamamoto et al., "Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates," Biomolecules, vol. 10, no. 11, pp. 1–13, Nov. 2020, doi: 10.3390/BIOM10111534.

[15] S. Sukegawa et al., "Identification of osteoporosis using ensemble deep learning model with panoramic radiographs and clinical covariates," Scientific Reports 2022 12:1, vol. 12, no. 1, pp. 1–10, Apr. 2022, doi: 10.1038/s41598-022-10150-x.

[16] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial Intelligence, vol. 97, no. 1–2, pp. 273–324, 1997, doi: 10.1016/S0004-3702(97)00043-X.

[17] E. N. Chidozie et al., "Osteoporosis Risk Predictive Model Using Supervised Machine Learning Algorithms," http://www.sciencepublishinggroup.com, vol. 5, no. 6, p. 78, Jan. 2018, doi: 10.11648/J.SR.20170506.11.

[18] I. Majeed Wani and S. Arora, "Knee X-ray Osteoporosis Database," vol. 2, 2021, doi: 10.17632/FXJM8FB6MW.2.

[19] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/S11263-015-0816-Y.

[20] "How transferable are features in deep neural networks? | Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2." https://dl.acm.org/doi/10.5555/2969033.2969197 (accessed Jan. 10, 2022).