Proceedings of the 2019 5th International Conference on Advances in Electrical Engineering (ICAEE), 26-28 September, Dhaka, Bangladesh

Deep CNN-Based Ensemble CADx Model for Musculoskeletal Abnormality Detection from Radiographs

Tusher Chandra Mondol[1], Hasib Iqbal[1], and MMA Hashem[1]

[1]Department of Computer Science and Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh
tusher.mondol2013@gmail.com, hasibiqbal207@gmail.com, mma.hashem@gmail.com

***Abstract*—Musculoskeletal Disorders (MSDs) are excoriations and afflictions that assail body movement of human. In present days diagnosis of musculoskeletal conditions are dependent on radiographs. Sometimes doctors or radiologist can make an error that can mislead the diagnosis of abnormalities. So, we have been motivated to develop a novel Computer-Aided Diagnosis(CADx) system based on Deep Convolutional Neural Network (Deep CNN) that will help the doctors to identify musculoskeletal abnormalities through radiographs. We have used VGG-19, ResNet architecture to build a model for four types of study (Elbow, Wrist, Finger, and Humerus). 5-fold cross-validation method is also applied to evaluate our models. Then we have applied ensemble techniques to improve the model's performance. Finally, based on some performance evaluating metrics the best one is selected for each of the study types and in aggregate. Our proposed technique tested on a benchmark radiographic dataset named 'MURA', and the final result is compared to other prominent techniques. For Elbow, Finger, Humerus, Wrist study, model performance was consecutively 86.45%, 82.13%, 87.15%, and 87.86%. Experimental consequences show that our proposed method is a condign strategy to resolve musculoskeletal abnormalities detection.**

Keywords: Deep CNN, Musculoskeletal Abnormality, CADx, VGG-19, ResNet, 5-fold Cross-validation, Transfer Learning, Ensemble.

I. Introduction

Musculoskeletal abnormalities invade the muscles, tendons, discs, ligaments, bones, nerves, blood vessels etcetera. Global Burden of Disease (GBD) conducted a study in 2016, found that musculoskeletal abnormalities were the second highest exploiter to global disability, and lower back pain. Around 20% - 30% people worldwide live with tormenting musculoskeletal abnormalities [1]. Risk of enkindling musculoskeletal abnormalities has stimulated by age, family history, activity level, practicing poor gesture at work, etc. Especially a computer engineer is in a high risk of developing such abnormalities because he or she has to engage with a computer in the same fashion every day.

Proper treatments can cure musculoskeletal abnormalities. These treatments planned on the cornerstone of diagnosis. Diagnosis of musculoskeletal abnormalities is dependent on radiographs. After evaluating these radiographs, radiologists make treatment plans. Sometimes doctors can also be misguided or confused. Because in most cases, radiological elucidation is highly impacted by the clinical affairs or circumstances of the patient, pertinent preterit history and quondam medical imaging. A recently conducted report enumerated that around one billion radiologic examinations are perpetrated worldwide annually, most of which are elucidated by the radiologists [2]. 13% major and 21% minor discrepancy rate found in a 2007 study of the influence of experienced neuro-radiologist second reading of CT and MR studies initially interpreted by general radiologists [3]. A study from Massachusetts General Hospital conducted in 2010 found 26% major inter-observer and 32% intra-observer discrepancy rates [4]. Kim and Mansfield published a radiological errors categorization system in 2014 on the base of a dataset containing 1269 errors, which shows that under-reading was one of the most common clinically significant error category [5] [6]. So, we have been motivated to develop a decision support system for radiologists with the aim of preventing problems from getting worse as a result of failure to see the signs of trouble.

This paper aims to develop a Computer-Aided Diagnosis (CADx) system that can classify musculoskeletal radiographs as abnormal or normal. To do so, we choose image based CADx as it is more powerful compared to feature based CADx [7]. First of all, we have collected a benchmark dataset 'MURA' [8] which contains radiographic images of elbow, forearm, hand, humerus, wrist, finger, and shoulder. Among them, we have worked on four of this upper extremity (elbow, finger, humerus, and wrist). We have passed the collected data through some pre-processing techniques. Then we have used those data to construct a model. We have used two deep CNN architecture (VGG-19, ResNet) to construct several models. The input of the model is one or more views of a certain study. For VGG-19 a probability of abnormality is assigned on each view, by a 19-layer convolutional neural network. For ResNet, the probability of abnormality is assigned on each view, by a 50-layer convolutional neural network. These probabilities are then averaged to predict the final probability of being abnormal of a study. After that, we have performed ensemble techniques. Finally, we have compared several classification models concerning some performance evaluation metrics to assess the best classifier. Experimental results show that our method performs well on musculoskeletal abnormalities detection,

978-1-7281-4934-9/19/$31.00 ©2019 IEEE

gaining better accuracy ranging from 82.13% to 87.86% for different study types. The proposed system architecture is illustrated in Figure 1.

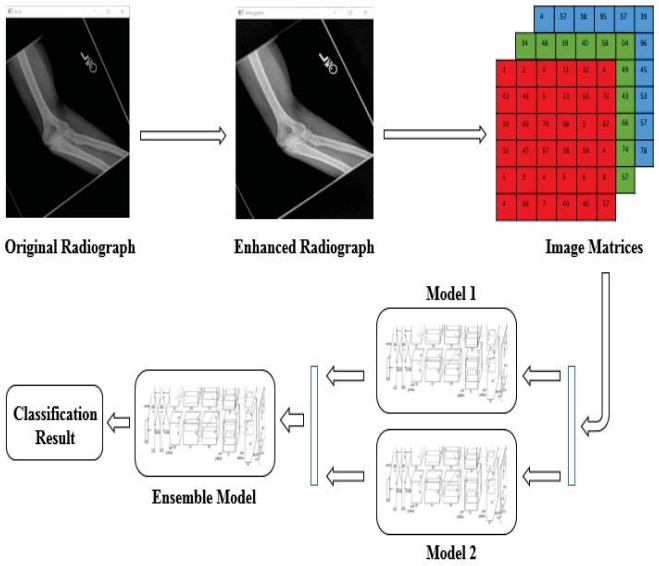


Fig. 1: System Overview

We summarize our contributions to this work as follows:

- Applying some prominent pre-processing techniques on radiographic image data.
- Applying a set of benchmark architecture to quantify all performance measuring parameters of our classifier model to ensure better performance.
- Ensemble of several models to improve the model's performance.
- Experimental analysis and comparison with other prominent classification techniques.

The remaining part of the paper is organized as follows: Section II demonstrated the related works that have been done in this field. The methodology of the system with the subsection of system design, block diagram, etc. are investigated in Section III. Result analysis and discussion is described in Section IV & the conclusion of the paper is drawn in Section V.

II. Related Works

Pranav Rajpurkar et al. [9] used a 169-layer convolutional neural network to predict the probability of abnormality for each image in a study of musculoskeletal abnormalities detection. DenseNet architecture used in the network. Then they replaced the fully connected final layer with a single output layer. After that, they applied a sigmoid nonlinearity. They have normalized each image of the dataset to get the similar standard deviation and mean of images in the ImageNet training set. Variable-sized images are then scaled to 320 X 320. During training, data is augmented by using rotations of up to 30 degrees and random lateral inversions. The initialized the network's weights with weights from a model pre-trained on ImageNet [10]. End-to-end training is used to train the network using Adam optimizer. Finally with the lowest validation losses authors ensembled the five models. But in some study types, their model's performance was poor especially for Wrist, Humerus.

Jose George et al. [11] proposed a technique for detection of Temporal bone abnormalities. Their dataset was collected from the diagnosis of the ear in High Resolution Computed tomography (HRCT) images. To improve contrast authors histogram equalized each image and performed median filtering to remove the noise and outliers. Area of interest is selected using an adaptive mask. Authors extracted two features for both ears' area of interest. One kind of the feature was texture features, extracted from Gray Level Co-occurrence Matrix (GLCM) and the other one was geometric features obtained from the region of the temporal bone. The texture features include Contrast, Maximum Probability, Energy, Inverse Element Difference Moment and Entropy. 15 features which were used for classification were extracted by calculating GLCM in 0o, 45o and 90o directions. Wavelet SVM was used for the classification.

N.Umadevi et al. [12] proposed an ensemble classification system for detecting fracture in human bone by X-ray images. At the first phase, a hybrid de-noising method with ICA that was coupled with wavelets enhances radiographs. In the second phase, segmentation extracted the bone structure using an active contour model. To estimate the initial seeds, region growing algorithm was used. Texture and shape features that best disclose the peculiarity of the segmented image were extracted. Then using binary classifiers SVM, KNN and BPNN they classified the presence of a bone fracture.

III. Proposed Methodology

In this section data collection, data pre-processing, model training and testing, perform ensemble and assess classifiers with respect to diversity process are explained briefly. Figure 1 illustrates the workflow of overall system.

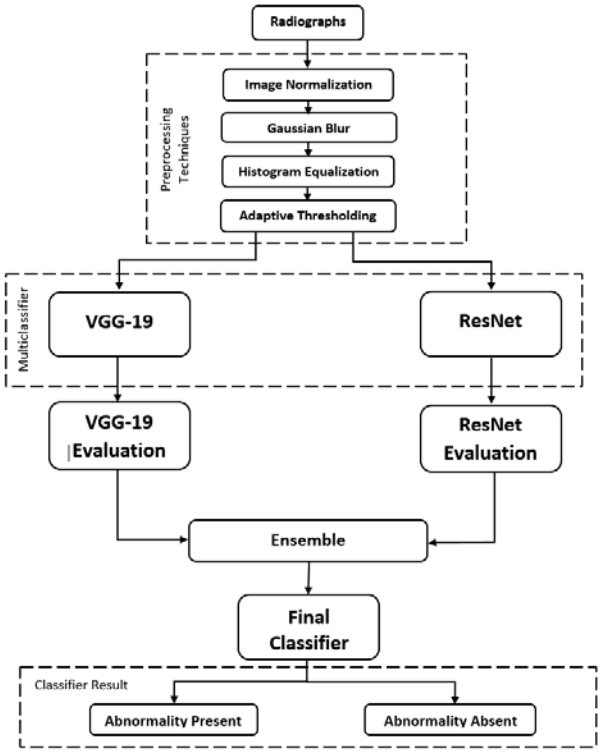


Fig. 2: Flowchart of Abnormality Detection Technique

A. Data Collection

We have collected our dataset "MURA" from Stanford University ML group [8] which is a large and benchmarked dataset of musculoskeletal radiographs comprising 14,863 studies from 12,173 patients. We have worked with 7260 studies with a total of 22,938 multi-view images. Each image belongs to one of four standard upper limb radiographic study types: elbow, finger, humerus, and wrist. Table II discloses the distribution of studies.

TABLE I: Distribution of Dataset We Used

Study	Train		Validation		Total
	Normal	Abnormal	Normal	Abnormal	
Elbow	1094	660	92	66	1912
Finger	1280	655	92	83	2110
Humerus	321	271	68	67	727
Wrist	2134	1326	140	97	3697
Total No. of Studies	4829	2912	392	313	8446

B. Data Preprocessing

Before feeding the data to the neural network we have applied some preprocessing techniques that enhance the radiograph images. First, a radiograph is normalized in the range between (55 – 255). Then the normalized radiograph is smoothed through Gaussian Blur using a 5X5 kernel. After that histogram equalization is done. And finally, adaptive thresholding is done by calculating the mean of a 3X3 kernel for each pixel then updating the pixel value with the mean value. Besides these, we have also tried Canny Edge Detection but we have not got any expected betterment. Finally, an enhanced radiograph is achieved.

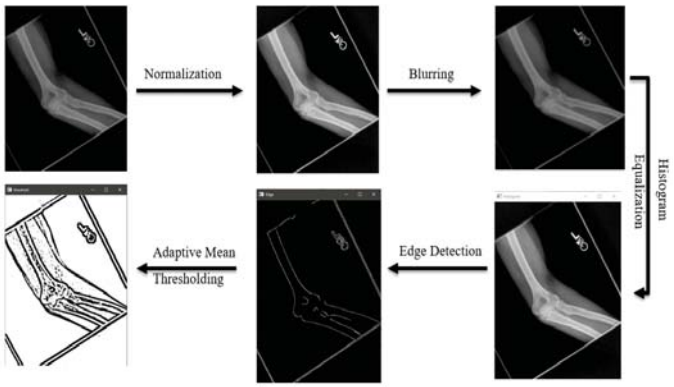


Fig. 3: Data Preprocessing Effects

C. Model Construction

Before the training period, we studied a bunch of state of the art architecture in the field of convolutional neural network & finally decide to train model based on VGG and ResNET architecture. Both architectures have been trained on the ImageNet dataset, which contains more than 22,000 object categories and 15 million high resolution training images [13]. VGG architecture is a shallow network that is capable of capturing small features but skips big pattern majority of the time. On the other hand, ResNET architecture is quite a deep network that is designed consisting of the residual block which aims to learn big pattern as well as the small one. Predefined training data is fitted individually to both models in the training phase.

In the model training process, we studied transfer learning where a model trained on one task is re-purposed on a second related task. Deep Neural Networks have great modeling or predictive abilities. But many parameters generally needs more samples or training data. The implication that deep learning may not be as effective on the small or mid-size dataset. The appearance of transfer leaning quite reduces this problem. Transferable weights are effective in the sense that they already tuned for extracting features where initializing with random weights may or may not able to grab a weight that will able to extract feature. Thus the model training process starts with transferable weight. We decided to use the pre-trained model approach to transfer learning. To apply the transfer learning concept, we have collected pre-trained weight of Imagenet and initialize weights of it before starting the training process.

In times of VGG architecture, we have tuned the first five-layer and seventeen, eighteen number layer for the betterment of the model. We have frozen the weights of the first five-layer before training so that it just carries the transferable weights throughout the training time. In the original architecture, 17th 18th No. layer both consists of 4096 neurons. As Imagenet is a big dataset compared to us, we planned to reduce the number of neurons to reduce the complexity. We simulated with a various number of neurons finally choose to use 1024 neurons in those two layers. After all this fine-tuning, compilation of model in both architectures is done using SGD optimizer with a learning rate of 0.0001 and momentum of 0.9. The final step of training is to fit training data with the compiled model. We run the defined model for 100 epochs with the early stopping of patience 10. Training of model is done various times with different epochs, and at last, the decision is made to approach like above. We have applied the 5 fold cross-validation technique during the training period. This method is used to determine the stability of the trained model. In this scenario, all the dataset used for training as well as testing. For each type, we have finally selected our base model of each architecture in terms of Area Under ROC Curve (AUROC). For which model AUROC is the highest that model is selected.

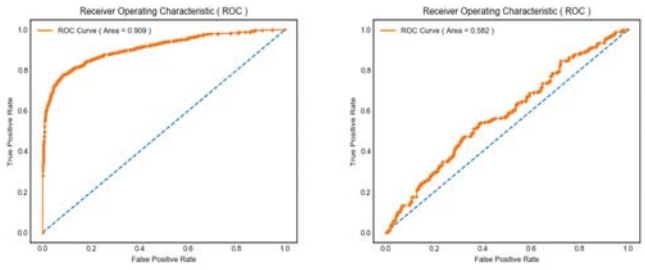


Fig. 4: ROC curve of VGG-19 and ResNet respectively for Wrist

D. Model Ensemble

After selecting base models, an ensemble technique applied to base models of each class with the target of Achieving

better accuracy. Following equation is used in determining the weight of each model for a specific domain.

$$W_i = \frac{C_i}{\sum_{i=1}^{n} c_i} \tag{1}$$

After that, probability of abnormality is determined using following formula:

$$PredictionProbability = \sum_{i=1}^{n} P_i * W_i \tag{2}$$

Where

P_i = Probability Assigned for a Radiograph by i^{th} classifier

W_i = Weight of i^{th} Classifier

If (Prediction Probability > .5) then the radiograph will be classified as Abnormal. After the training phase, testing of the model is used to measure how well the model performs at making predictions on that testing set.

E. Performance Measuring Metrics

Evaluating the neural network model is an essential part. A model may give satiated results when evaluated using one metric but concurrently may give anomalous results when evaluated against another metric. So, to assess classifiers with respect to diversity, our model has been evaluated in terms of multiple performance measuring indices. First, a confusion matrix for actual condition and the predicted condition is formed consisting of TP, TN, FP, and FN to assess the parameter. Significance of these terms is narrated below.

- TP = True Positive (Identified Correctly)
- TN = True Negative (Identified Incorrectly)
- FP = False Positive (Rejected Correctly)
- FN = False Negative (Rejected Incorrectly)

Evaluation metrics are stated below:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

$$Specificity = \frac{TN}{FP + TN} \tag{6}$$

$$F_\beta Measure = \frac{(1 + \beta^2)PR}{\beta^2 P + R} \tag{7}$$

$$MissRate = 1 - \frac{TP}{TP + FN} \tag{8}$$

$$FallOut = 1 - \frac{TN}{FP + TN} \tag{9}$$

$$MCC = \frac{TPXTN - FPXFN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

$$Cohen'sKappa = \frac{Accuracy - Expected}{1 - Expected} \tag{11}$$

IV. Results Analysis & Discussion

A. Experimental Results

We have applied two different types of convolutional neural network architecture into four types of XRAY data. We have computed some performance evaluation metrics for each architecture of each study. Based on this we choose the best architecture. Their corresponding results are discussed in the corresponding subsection.

1) Elbow:

Confusion matrix of different deep CNN model for 'Elbow' classification is illustrated bellow:

TABLE II: Cofusion Matrix Parameters of Elbow

	True Positive	True Negative	False Positive	False Negative
VGG-19	354	552	93	80
ResNet	173	122	57	113
Ensembled	192	210	20	43

To evaluate the stability of our model we have applied 5-fold cross-validation during training of VGG-19 model. Accuracy vs. Epochs graph is given below:

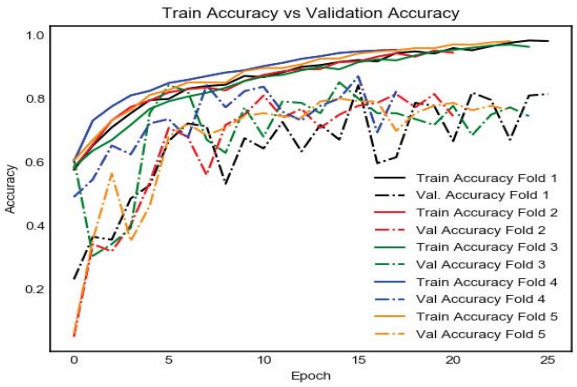
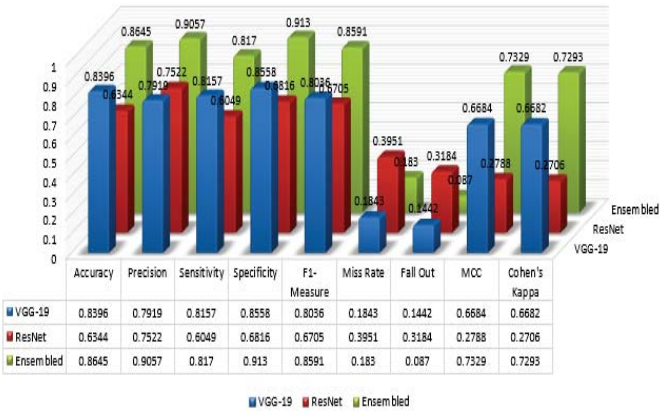


Fig. 5: Accuracy vs. Epochs Graph



	Accuracy	Precision	Sensitivity	Specificity	F1-Measure	Miss Rate	Fall Out	MCC	Cohen's Kappa
VGG-19	0.8396	0.7919	0.8157	0.8558	0.8036	0.1843	0.1442	0.6684	0.6682
ResNet	0.6344	0.7522	0.6049	0.6816	0.6705	0.3951	0.3184	0.2788	0.2706
Ensembled	0.8645	0.9057	0.817	0.913	0.8591	0.183	0.087	0.7329	0.7293

Fig. 6: Performance Metrics of Elbow Classification

For 'Elbow' classification the accuracy of our Ensemble method is 86.45%, F_1 Measure is 0.8591, and Miss Rate is 0.183. On every aspect Ensemble model is better than both VGG-19 and ResNet.

2) Finger:

Confusion matrix of different deep CNN model for 'Finger' classification is illustrated bellow:

TABLE III: Confusion Matrix Parameters of Finger

	True Positive	True Negative	False Positive	False Negative
VGG-19	175	192	72	22
ResNet	113	193	134	21
Ensembled	155	130	48	14

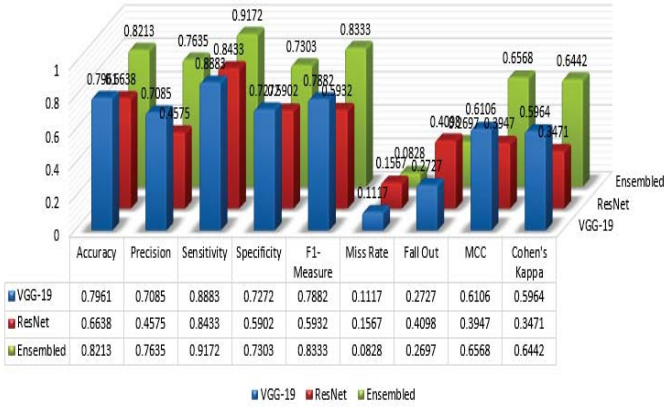


Fig. 7: Performance Metrics of Finger Classification

For 'Finger' classification the accuracy of our Ensemble method is 82.13%, F_1 Measure is 0.8333, and Miss Rate 0.0828. On every aspect Ensemble model is better than both VGG-19 and ResNet.

3) Humerus:

Confusion matrix of different deep CNN model for 'Humerus' classification is illustrated bellow:

TABLE IV: Confusion Matrix Parameters of Humerus

	True Positive	True Negative	False Positive	False Negative
VGG-19	116	144	32	20
ResNet	72	91	68	57
Ensembled	128	123	17	20

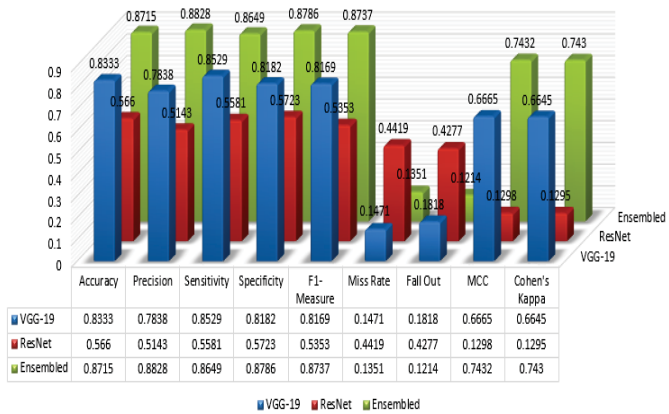


Fig. 8: Performance Metrics of Humerus Classification

For 'Humerus' classification the accuracy of our Ensemble method is 87.15%, F_1 Measure is 0.8737, and Miss Rate is 0.1351. On every aspect Ensemble model is better than both VGG-19 and ResNet.

4) Wrist:

Confusion matrix of different deep CNN model for 'Wrist' classification is illustrated bellow:

TABLE V: Confusion Matrix Parameters of Wrist

	True Positive	True Negative	False Positive	False Negative
VGG-19	634	1128	222	97
ResNet	103	275	192	89
Ensembled	335	244	51	29

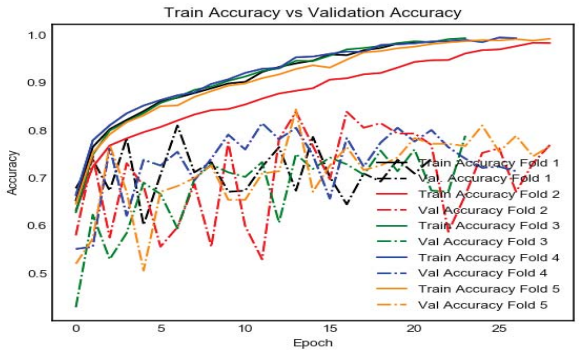


Fig. 9: Accuracy vs. Epochs Graph

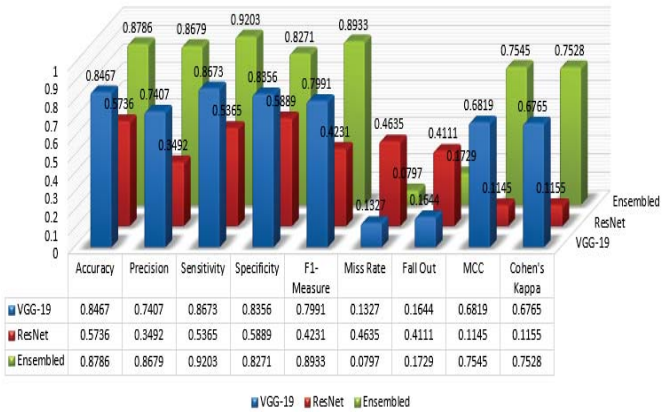


Fig. 10: Performance Metrics of Wrist Classification

For 'Wrist' classification the accuracy of our Ensemble method is 87.86%, F_1 Measure is 0.8933, and Miss Rate is 0.0797. On every aspect Ensemble model is better than both VGG-19 and ResNet.

From the above representation, it can be inferred that our proposed ensemble model works comparatively well on almost all (Elbow, Finger, Humerus and Wrist) classification task.But there still exists some scope of improvement, so we seek more notable research in this area. Eventually our results and analysis should provide important baseline and steerage for future work .[1]

[1]Code for this work is available at https://github.com/hasibiqbal207/XRAY-Abnormality-Detection-CADx-Model

B. Discussion

We compare board-certified radiologists, our reference paper MURA and our model on the Cohen's kappa statistic.

	Radiologist 1	Radiologist 2	Radiologist 3	MURA	**Our Model**
Elbow	.850	.710	.719	.710	**.729**
Finger	.304	.403	.410	.389	**.644**
Humerus	.867	.733	.933	.600	**.743**
Wrist	.791	.931	.931	.931	**.753**
Total	.703	.694	.748	.658	**.717**

Fig. 11: Comparison Between Radiologists, MURA and Our Model Based on Cohen's Kappa Statistic

On 'Elbow' study, model performance (.729) is higher than the MURA (.710) and worst radiologist performance (.710) but lower than the best radiologist performance (.850). On 'Finger' studies model performance is the global best (.644). On 'Humerus' study, model performance (.743) is higher than the MURA (.600) and worst radiologist performance (.733) but lower than the best radiologist performance (.933). On 'Wrist' model performance (.753) is lower than the worst radiologist performance (.791). Overall, the model performance (.717) is better than MURA (.658) and two other radiologist's performance but lower than best radiologist's performance (.748).

V. Conclusion

In today's revolution oriented medical environment, Picture Archiving and Communication System (PACS), Computer-Aided Diagnosis (CADx) system play a significant role in a wide spectrum of services and applications. The most significant quality target of this type of system is high speed and feasible efficiency in disease detection. Our research here contains some of the new explorations of how to pre-process image data and lastly how to build up a classifier model using these data for musculoskeletal abnormality detection using radiographs. We gave an overall design. We seek that this research will be reviewed to promulgate the advances, prospects, limitations, and challenges for the improvement of such a decision support system that will help the doctors to take the therapeutic medical decision quickly and precisely. For further improvement of our study we would like to recommend the following things :

- Contrast Limited Adaptive Histogram Equalization (CLAHE) should be used instead of Histogram Equalization (HE) to enhance the radiographs.
- More research should be done to improve performance and make the model more feasible.
- Whole classification task should be over the cloud.
- A user-friendly mobile application should be developed for the end-users.

References

[1] Chollet and Francois. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the global burden of disease study 2016. pages 1211–59. Manning Publications Co., 2016.

[2] Abujudeh HH. Bruno MA, Walker EA. Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. page 1668–1676. RSNA, 2015.

[3] Worthington M Rennie I McKinstry CS. Briggs GM, Flynn PA. The role of specialist neuroradiology second opinion reporting: is there added value? page 791–795. Europe PMC, 2008.

[4] Hani H. Abujudeh, Giles W. Boland, Rathachai Kaewlai, and G. Scott Gazelle James H. Thrall Pavel Rabiner, Elkarn F. Halpern. Abdominal and pelvic computed tomography (ct) interpretation: discrepancy rates among experienced radiologists. pages 1952–7. Springer, 2008.

[5] AJR Am J Roentgenol Kim YW, Mansfield LT. Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors. pages 465–70, 2014.

[6] Smith M. Charles C. Thomas. Error and variation in diagnostic radiography. Springfield, 1967.

[7] Shoji Kido, Yasusi Hirano, and Noriaki Hashimoto. Detection and classification of lung abnormalities by use of convolutional neural network (cnn) and regions with cnn features (r-cnn). In *International Workshop on Advanced Image Technology*. IEEE, 2018.

[8] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. MURA Dataset: Towards Radiologist-Level Abnormality Detection in Musculoskeletal Radiographs. https://stanfordmlgroup.github.io/competitions/mura/, 2018. [Online; accessed n.d].

[9] Aarti Bagul Daisy Ding Tony Duan Hershel Mehta Brandon Yang Kaylie Zhu Dillon Laird Robyn L. Ball Curtis Langlotz Katie Shpanskaya Matthew P. Lungren Andrew Y. Ng Pranav Rajpurkar, Jeremy Irvin. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. 2018.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database.

[11] Subin T. K† Jose George and K. Rajeev. Detection of temporal bone abnormalities using hybrid wavelet support vector machine classification.

[12] N. Umadevi ; S.N. Geethalakshmi. Multiple classification system for fracture detection in human bone x-ray images. In *2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12)*. IEEE, 2012.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. ACM, 2017.