Journal of Nature, Science & Technology 4 (2021) 1-5
Nature, Science & Technology
www.acapublishing.com

Research Article

Ensemble Based Neural Network for the Classification of MURA Dataset

Mithun Ghosh[1*], Sahil Hassan[2], Prama Debnath[3]
[1]*Department of Systems and Industrial Engineering, University of Arizona, Arizona, USA.*
[2]*Department of Electrical and Computer Engineering, University of Arizona, Arizona, USA.*
[3]*Department of Computer Engineering American International University-Bangladesh, Dhaka, Bangladesh.*

Keywords

Neural Network and Convolution,
Neural Network,
Image classification,
Kappa coefficient,
Adaboost learning.

Abstract

The Musculoskeletal Radiographs (MURA) dataset, proposed by Stamford Machine Learning (ML) group, has 40,561 images of bone X-rays from 14,863 studies. The X-ray images belong to seven body areas of the upper extremity namely wrist, elbow, finger, humerus, forearm, hand, and shoulder. Radiologists have classified the data into two classes, namely normal and abnormal. Six board-certified Stanford radiologists labeled the data samples using most votes, which is considered the gold standard. The 169 layers deep model, introduced by the Stamford ML group, works well on a par with the gold standard except for the humerus radiographs, despite humerus data labeled with high accuracy. We propose to develop a comparatively shallower version of a neural network and a convolutional network with 10 hidden layers each in an Adaboost framework in the humerus data and the model performance is on par or sometimes superior to the Stamford ML group model. We evaluate the performance of our model using the validation error and Cohen's kappa coefficients. We have shown that our modeling framework is much faster in terms of the model training time and as accurate compared to the 169 layers of deep neural network introduced by the Stamford ML group. Also, with increased resources, the performance of our model will increase.

1.Introduction

The current era of big data has sped up the development of machine learning-based models. It is becoming increasingly crucial that these models can handle large datasets without losing significant information. Out of the several application fields of machine learning models, the medical sector is one of the most growing and demanding sectors. It requires models that can handle a sizeable amount of data with high accuracy. Among the several machine learning approaches, the neural network is a state-of-the-art algorithm to handle large datasets (Deng et al. (2009)[[1.]) with high accuracy. This makes it a suitable choice for medical applications with lots of data.

With the era of big data, along with the capability of modeling unstructured data, the popularity of machine learning models is booming in medical science. The usage of large data in the modeling will bring new hidden information into the decision-making process. Because of the large data handling capabilities, the neural network is becoming the prime model to handle image data in medical science. The main purpose of the study of the radiograph classification due to the importance of early detection of the bones condition. The radiographic study is one such branch, where the correct classification of the bone's condition can prevent serious damages in the long term. According to WHO (World Health Organization), one of the critical impediments on any individual is musculoskeletal-related fractures. Woolf and Pfleger (2003) [10.] discussed the burden of the major musculoskeletal conditions. These conditions affect around 1.7 million people around the world. Based on the classification outcome, the patient may need to go further diagnosis and treatment. But most of the time it is difficult even for the naked eye to detect any abnormality in the X-ray images which we can notice in Fig. 1. Mistreatment because of misclassification is critical in any situation for that patient. Because of this imperative situation, an accurate machine learning model must be needed that can flawlessly detect the bone's condition in a faster, detailed, and more interpretable way.

A Stamford Machine learning (ML) group (Rajpurkar et al.(2017)[8.]) introduced a large group of data for musculoskeletal radiographs and named it MURA. This dataset can help to train and create an accurate model for classification in radiographic studies. It contains 14,863 studies with 40,561 images. Each study in the data contains multiple views of the targeted body part, which are labeled by radiographers as normal or abnormal. These are high-resolution images that make the size of the whole dataset around 3.3 gigabytes. The enormous size of this dataset and the crucial accuracy requirements of the radiographic study sector require neural network-based approaches.

The Stamford ML group proposed a deep neural network model of 169 dense layers to reliably classify the MURA dataset. This model classifies most of the X-ray images of different body parts with good agreement with the radiologist classifiers (gold standard). However, it shows low agreement with the gold standard for the finger and humerus dataset. In the case of the finger dataset, Rajpurkar et al. (2018)[8.] showed that the existing finger dataset labels had a low value of Cohen's kappa coefficient. This causes the model to perform poorly. However, for humerus data, despite highly accurate labels, their model performs poorly. The computational complexity of their deep neural network model requires a Graphical Processing Unit (GPU) to execute. To address these two issues, our work presents an ensemble model of 10 layered neural networks. Our model can run on a general-purpose computer without a GPU. But to get the full advantage of the model, we require some high-performing computers (HPC) because of the huge size of the image files. We also build an auto-encoder as an ensemble framework for the neural networks and add another model architecture namely, Convolutional Neural Network(CNN). We have carried out the model training on the humerus dataset only.

*Corresponding Author: mithunghosh@email.arizona.edu
(Mithun Ghosh Orcid: 0000-0002-3599-4805)
Received 06 April 2021 Revised 20 May 2021 Accepted 20 May 2021
Journal of Nature, Science & Technology 4 (2021) 1-5
https://doi.org/10.36937/janset.2021.004.001

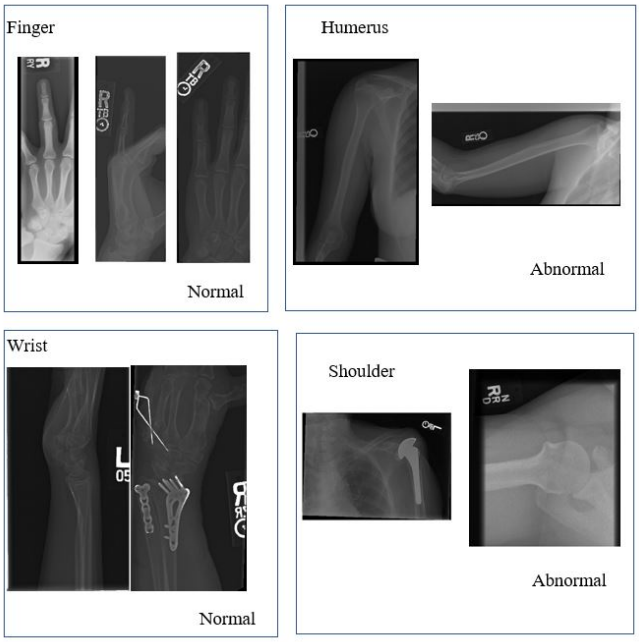

Figure 1. MURA dataset conducts musculoskeletal studies which contain 14863 images of the radiography of the upper extremity. In each of the studies, radiologists manually labeled multiple views. The right side of this figure explains some normally labeled images of the elbow and wrist, respectively wherein the left side describes some abnormal images of the humerus and shoulder, respectively.

We organize the paper in the following way with the section 2 describes the details of the MURA dataset, section 3 describes the proposed ensemble model, section 4 contains the model interpretation, section 5 compares the performance of the proposed model with the Stamford ML group's model, section 6 describes the related works, and Section 7 discusses the overall findings.

2. MURA Dataset

The task in the MURA dataset is to find the binary class of {0,1}. The output of each study contains one or more views and we denote the expected output as 0 (normal) or 1(abnormal). A summary of the study data is given on the official MURA dataset website [2.]. We can see the performance of various authors' models in different categories of the human body in [2.]. Some of these models perform well for different categories. But For upper extremity categories: humerus and finger, almost all the models perform badly. This may be since the images in these categories are not so clear. Also, the number of samples is not sufficiently high in these categories. We can look at the top right image in Fig. 1, which is not so clear to be easily classified by a model. Thus, for this reason, we concentrate our modeling approach on the humerus data. On the left side of Fig. 2, we can also see that number of studied patients is very few for both the training and validation dataset in the humerus study.

2.1. Data Collection

As previously mentioned, data is collected from the MURA dataset's official website [2.]. A large number of patients were studied with different parts of fractures in their bodies. The total number of radiographs image is 14,863 that is collected from 12,713 patients with 40,561 multi-view X-ray images. We can classify these images as one of the seven standard upper extremity radiographic study types. Stanford Hospital board-certified radiologists marked these images as normal or abnormal based on the radiographs. The labeling was conducted during the assessment of DICOM (Digital Imaging and Communications in Medicine) images on a medical-grade display of at least 3 megapixels PACS (Picture Archiving and Communication System) with max luminance of 400 cd/m^2 and min luminance 1 cd/m^2 with a pixel size of 0.2. Also, the native resolution is 1500×2000 pixels. We divide the dataset into training and validation sets. Training set includes 11,184 patients, 13,457 studies, 36,808 images, and validation set includes 783 patients, 1,199 studies, 3,197 images. All these datasets are mutually exclusive and no crossover between these datasets.

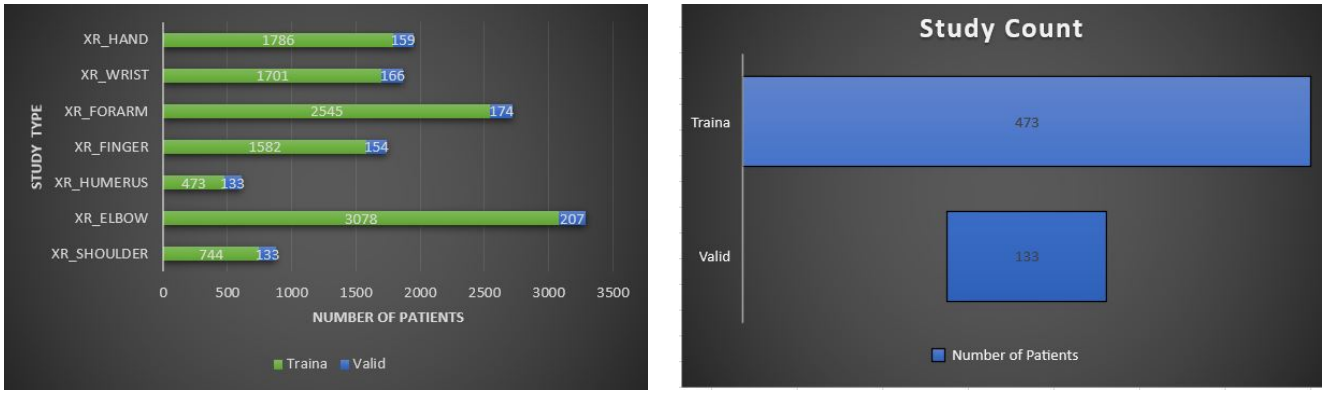

Figure 2. Left: Statistics of the data in each of the seven categories of the studies. XR_S means the X-ray images in the category of S. In training set XR_WRIST has the maximum number of patients, followed by XR_FINGER, XR_HUMERUS, XR_SHOULDER, XR_HAND, XR_ELBOW and XR_FOREARM. X_FOREARM with 606 patients has got the least number. We can see similar pattern in validation set, XR_WRIST has the maximum, followed by XR_FINGER, XR_SHOULDER, XR_HUMEROUS, XR_HAND, XR_ELBOW, XR_FOREARM. Here XR_FINGER defines radiographs of Finger upper extremity. Right: Data Statistics for humerus Data

2.2. Validation Data Collection

Extra labels of the validation data are obtained from board-certified Stanford radiologists. This data is composed of 207 musculoskeletal experiments to assess models to get a resilient assessment of radiologist performance. Radiologists respectively examined and identified each study in the validation set individually. With the help of PACS in the clinical reading room environment, they restored the validation data in a DICOM file. The average professional experience of these radiologists is 8.3 years, varying from 2 to 25 years. The

radiologists had no direct exposure to any clinical data. A standardized data entry program preserves the manual classifications by the radiologists.

2.3. Abnormality Analysis

For the abnormalities check in the dataset, we investigate radiologist reports of manually labeling 100 abnormal studies. In the findings, we see that 53 studies were labeled with fractures, 48 with hardware, 35 with degenerative joint diseases, and 29 with other abnormalities, including lesions and subluxations.

3. Model

The proposed ensemble model comprises a smaller, relatively shallower neural network and a convolutional neural network. The proposed neural networks are trained on the train data from humerus X-ray images of MURA using the Adaboost algorithm (Freund and Schapire [1999] [4.]). Section 3.1 explains the structures and training of each of these neural network architectures. Section 3.2 explains the overall ensemble training and prediction methods.

3.1. Network Architecture and Training

We show the structural details of each smaller deep neural network in Fig. 3. Each neural network has 10000 nodes in the input layer and one node in the output layer. We consider 10 hidden layers apart from input and output layers and reduce the node number of each layer as the network grows deeper. We build this network framework in the python Keras module. Now, the structural details of each CNN are shown in Fig. 4. Each of the CNN comprises a kernel window size of 4×4 with stride one and max-pooling of 2×2. This auto-encoder network compresses the size of the image from 100×100 to 9×9 with minimum reconstruction error. After that, the layers are flattened and fully connected to a dense layer with 500 nodes and a rectified linear activation function and then another dense layer with 20 nodes, and finally the two output layers.

To optimize the network, each neural network uses the ADADELTA optimizer (Zeiler (2012)[12.]). This optimizer is a popular approach to train NN with the added advantage of the per-dimension learning rate. Also, our models appear robust to noisy gradient information. We tuned our models with different model architecture choices, various data modalities and found the optimal values of the hyperparameters.

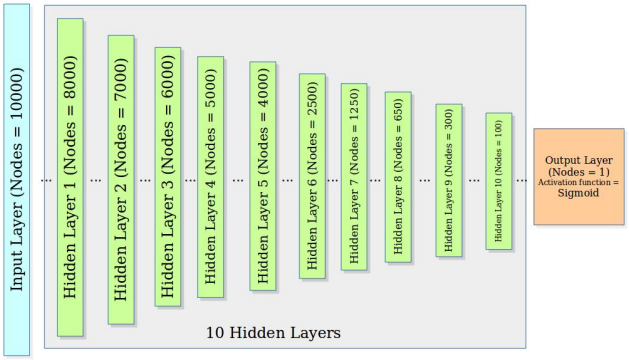

Figure 3. Structure of the small neural network used as a weak classifier in the ensemble model.

We consider minimizing the binary cross-entropy loss. For image X of study type T in the training set, the loss is:

$$L(X, y) = w_{T,1} y \log p(Y = 1|X) - w_{T,0}(1 - y)\log p(Y = 0|X) \quad (1)$$

Here, y and $p(Y = i|X)$ are the label and probability that the network assigns to the label i, respectively. Also, $w_{T,1} = \frac{|N_T|}{|A_T|+|N_T|}$ and $w_{T,0} = \frac{|A_T|}{|A_T|+|N_T|}$, where N_T and A_T define normal and abnormal images of the particular study T in the training set, respectively.

The hidden layer nodes use Rectified Linear Units (ReLU) function as an activation function. ReLU is easier to calculate and enables faster convergence during training. The output layer uses the Sigmoid function for activation. The threshold for the output node is set at 0.5. The ReLU and Sigmoid functions are shown in Eq. (2) and (3) respectively.

$$f(x) = max(0, x) \quad (2)$$

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \quad (3)$$

Among the several model parameters of the neural networks, we keep the training epoch number as a variable to study the effect of epochs on the classifier's performance on the training and validation dataset. There are overall 1272 training images on the humerus. The batch size is kept at 30 during training. We vary the epoch sizes from 10 to 40.

3.2. Ensemble Training and Prediction

In the previous sections, we described the architecture of neural networks. We use the Sampling-based Adaboost algorithm to train these neural networks with emphasis on the wrong classifications made by a previously trained classifier. The ensemble model contains 5 small neural networks, which implies that 5 Adaboost iterations are used to train the model. We describe the architecture of the CNN in Fig. 5. For the CNN, due to lack of resources, we just fitted it with five classifiers and ten epochs.

Algorithm 1 Training steps for neural network ensemble model using Adaboost Algorithm

1: **Input:** $S := \{x_i, y_i\}_{i=1}^{N}$, *Learning rounds* T, and *hypothesis class* H
2: **Initialize:** *Distribution* $D_1(i) = \frac{1}{n}$, $S := normalize(S)$
3: **for** $t = 1, 2, ..., T$ **do**
4: $h_t = argmin_{h \in H} \hat{err}(h, S, D_t)$
5: $\epsilon_t = \sum D_t(i)[h(x_i) \neq y_i]$
6: $\alpha_t = \frac{1}{2} log(\frac{1-\epsilon_t}{\epsilon_t})$
7: $D_{t+1}(i) = \frac{D_t(i)}{Z_t} exp(-\alpha_t y_i h_t(x_i))$
8: **end for**

Figure 4. Algorithm of the AdaBoost framework.

The overall flow of the ensemble training is shown in Fig. 4. Initially, all the training images and their labels are imported from directories. The training images contain different amounts of pixels. We resize all the images into 100×100 pixels and flatten them to 1×10000 sizes of pixels. After that, we normalize these pixel values between $[0,1]$ and run the iterative training procedure of the Adaboost algorithm, with the initial training data distribution (D) equal for each sample. In each iteration, a small NN or CNN is trained on training data sampled as per the updated distribution
by the previous iteration. Upon completion of training, the weighted error ϵ_t is calculated (step 5). Based on ϵ_t, the weight of the trained neural network, α_t is calculated (step 6). Then, the distribution D is updated as per the equation described in step 7 of the algorithm. Here, Z_t is a normalization term. Finally, the training dataset is sampled as per the updated D for training purposes in the next iteration.

We conducted the prediction using this ensemble neural network model as per the following equation

$$H(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x)) \quad (4)$$

4. Model Interpretation

As mentioned before, we aim to focus on the most vulnerable upper extremity (humerus) prediction accuracy by previous methods in the MURA dataset. We calculated both training error and validation error based on training and validation data. The error calculation formula:

$$Error = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n} \quad (5)$$

where y_i defines the ith true observed value, $\hat{y}_i$ defines corresponding predicted value from our model and n defines the length of the observation.

The kappa coefficient (κ) is more robust than the simple percentage agreement calculation and measures an inter-rater agreement for categorical items. The probability of the arrangement occurring by chance is taken into consideration by κ. The kappa of Cohen measures the agreement between two raters, each classifying N items into mutually exclusive categories. κ is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o can be defined as the relative observed agreement among raters and p_e is the hypothetical chance agreement probability, using the observed data to calculate the probabilities of each observer seeing each category randomly. Also, κ =1 represents the full agreement between raters and κ =0, which implies that there is no effective agreement between the two raters or the agreement is worse than random. In the following section, we describe our model performance using these statistics.

5. Radiologist vs Model Performance

To determine the performance evaluation of our proposed model, we use Cohen's kappa coefficient (κ) for both training and validation data. Table 1 presents the κ values of classifications made by the radiologists with the best performance, model by Stamford ML group, and model proposed in this work. The results contain model performance for different training epochs. We calculate all the mentioned κ values in this table by comparing them with the gold standard. Table 1 results show that, for the classification of the training data, the model performs somewhat reliable.

Table 1. We compare the performance(κ value) of our proposed model(NN) with the best radiologist performance and Stamford ML group model performance, on the humerus dataset of MURA. The comparison is done in terms of Cohen's kappa (κ)performance.

Datatype	R-3	SMLM	NN E=20	NN E=30	NN E=40	CNN E=10
V	0.93	0.60	0.425	0.45	0.51	0.62
Train	-	-	0.619	0.65	0.76	0.78

V: validation, E: Epoch number, R-3: Radiologist-3, SMLM: Stamford Machine Learning Model NN: Neural Network, CNN: Convolutional Neuron Network.

With increasing training epochs, the κ value increases, showing increasing the reliability of the model's performance. For predictions on validation data, the κ value is on par compared to the radiologist and Stamford ML group model with a larger number of the epoch. However, with increasing training epochs, κ based on the validation data increases. This shows, with larger training epochs, our model can be made reliable enough to compete with the other two methods. However, increasing the number of epochs increases training time significantly, which is why we are restricted to present results for a significantly higher number of epochs. We suggest using a High-Performance Computing (HPC) server to train this model for higher epochs. For training epoch 20, the model yields training and validation performance of κ values as 62% and 42%, respectively. For training epoch 30, training and validation κ values are 65% and 45%. Finally, for 40 epochs, for training, the κ value is 76% and for validation, the κ value is 51%. The study of training and validation errors show that increasing training epochs result in decreasing training and validation errors.

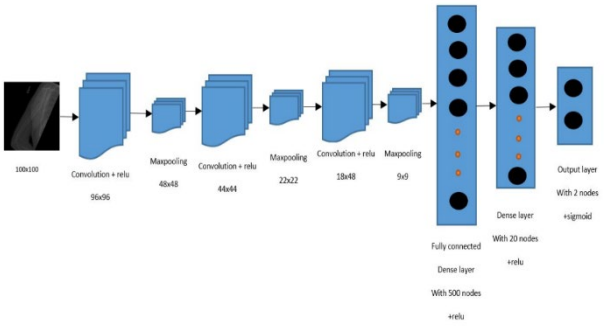

Figure 5. Structure of the shallow convolutional neural network which is used as a weak classifier in the ensemble framework

This again brings us to the conclusion of training the ensemble model for more epochs to get better performance. We may increase the number of smaller neural networks to get a more reliable classification performance. For the CNN with five classifiers, we get 78% training and 62% validation errors. As we can see from the table that CNN performs better with a comparatively lower epoch number. For CNN, with the increase of the number of classifiers, we believe we will get a better κ value. As mentioned above, because of the lack of resources, we could not perform a much larger number of classifiers. We fixed epoch size to 10 in the CNN model.

6. Related Work

Table 2. Overview of publicly available medical radiographic image datasets

Dataset	Study Type	Label	Images
MURA	Musculoskeletal (Upper Extremity)	Abnormality	40561
Pediatric Bone Age (AIMI)	Musculoskeletal (Hand)	Bone Age	14,236
O.E.1 (OAI)	Musculoskeletal (Knee)	K&L Grade	8,892
Digital Hand Atlas (Gertych et al., (2007) [5.])	Musculoskeletal (Left Hand)	Bone Age	1,390
ChestX-ray14	Chest	Multiple Pathologies	112,120
Openl	Chest	Multiple Pathologies	7,470
MC	Chest	Abnormality	138
Shenzen	Chest	Tuberculosis	662
JSRT	Chest	Pulmonary Nodule	247
DDSM	Mammogram	Breast Cancer	10,239

Our model is robust in the sense that we can apply in many large available datasets in many categories such as speech recognition (Hannun et al. (2014)[6.]), question answering (Rajpurkar et al. (2016)[7.]) heart arrhythmias (Rajpurkar et al. (2017) [8.]), etc. Nowadays, because of the large development in technology, we get a large amount of data in many areas. Large data means more information and our model should be able to handle these areas as well. It's not so easy to find openly available radiographs dataset although there is some continuous effort to make the domain of medical datasets available openly. Previously collected datasets are smaller than MURA in size. Table 2 gives a publicly available dataset in radiographs images. There are truly few publicly accessible datasets of musculoskeletal radiographs. The Stanford Artificial Intelligence Program in Medicine and Imaging hosts a dataset comprising illustrated skeletal age (AIMI) pediatric hand radiographs. Gertych et al. (2007) [5.] described the various ages labeled with bone age radiology renderings of children and names as the Digital Hand-Atlas. K&L grade of osteoarthritis (OAI) is labeled as the knee radiographs which were initialed by the Osteoarthritis. Each of these datasets contains less than 15,000 images. Some of the seminal work has been mentioned in the MURA

dataset official website [2.], where we can see the rank and the performance of these models. A family of embedding functions was used as an ensemble method to give improved results by Xuan et al. (2018)[11.].

7. Discussion

Detecting the anomaly in radiographs at an early stage is crucial for the patient. We can notice from the results of Rajpurkar et al. (2018) [9.] that even experienced radiologists may sometimes misclassify some critical detections. Human classification is costly, more time-consuming, and requires more effort. These reasons have made the machine learning-based classifier model a reliable alternative. Although several established models perform relatively well on the MURA dataset, for some upper extremities they are not reliable enough. These models require compute-intensive complex neural network-based models that are difficult to train. We attempted to address these issues in our proposed neural network models by incorporating them into an ensemble framework. The obtained results suggest that the proposed models work with good reliability. We further noticed an increasing upward trend in model reliability with the increasing number of training epochs. Based on these results, we strongly conclude that, with the increase in computational resources, our model can be one of the most reliable candidates in MURA dataset classification for humerus. In the future, a Bayesian framework can facilitate the model performance with uncertainty quantification, as described in Ghosh et al.(2020)[3.].

Nomenclature

NN : Neural Network
CNN : Convolutional Neural Network

Declaration of Conflict of Interests

The authors declare that there is no conflict of interest. They have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1.] Deng, J., Dong, W., & Socher, R. jia Li, L., Li, K., Fei-fei, L.. Imagenet: A large-scale hierarchical image database 13 (2009).
[2.] https://stanfordmlgroup.github.io/competitions/mura/
[3.] Ghosh, M., Li, Y., Zeng, L., Zhang, Z., & Zhou, Q. , Modeling multivariate profiles using Gaussian process-controlled B-splines. IISE Transactions, (2012) 1-12.
[4.] Freund, Y., Schapire, R., & Abe, N.. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence, 14 (1999) 771-780.
[5.] Gertych, A., Zhang, A., Sayre, J., Pospiech- Kurkowska, S., & Huang, H. K.. Bone age assessment of children using a digital hand atlas. Computerized medical imaging and graphics, 31 (2007), 322-331.
[6.] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Ng, A. Y.. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014).
[7.] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P.. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016).
[8.] Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., & Ng, A. Y.. Cardiologist-level arrhythmia detection with convolutional neural networks. arXiv preprint arXiv:1707.01836 (2017).
[9.] Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Ng, A. Y.. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957 (2017).
[10.] Woolf, A. D., & Pfleger, B.. Burden of major musculoskeletal conditions. Bulletin of the world health organization, 81 (2003), 646-656.
[11.] Xuan, H., Souvenir, R., & Pless, R.. Deep randomized ensembles for metric learning. In Proceedings of the European Conference on Computer Vision (ECCV) (2018) (pp. 723-734).
[12.] Zeiler, M. D., Adadelta: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.

How to Cite This Article

Ghosh, M., Hasan, S., and Debnath, P., Ensemble Based Neural Network for the Classification of MURA Dataset, Journal of Nature, Science & Technology, 3(2021), 25-61.
https://doi.org/10.36937/janset.2021.004.001