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

Deep Convolutional Neural Network-Based Detection of Bone Abnormalities in Musculoskeletal Radiographs

Prakash U M[1*], Arivazhagan N[2]

Submitted: 18/01/2024 **Revised**: 27/02/2024 **Accepted**: 04/03/2024

Abstract: Musculoskeletal abnormalities typically rely on radiographic examinations for diagnosis, but even experienced radiologists can miss abnormalities, underscoring the need for improved detection methods. This paper presents a novel approach utilizing Deep Convolutional Neural Networks (Deep CNN) for computer-aided bone abnormality detection. Leveraging the Stanford MURA dataset featuring radiological images of seven upper extremity types, we employ pre-processing techniques such as Histogram Equalization (HE) and Contrastive Limited Adaptive Histogram Equalization (CLAHE) to enhance image quality before inputting them into our proposed 5_{SET} (5S) model. This model accurately classifies images into the seven upper extremity categories and identifies normal or abnormal conditions. Our results demonstrate a remarkable overall accuracy of 92.10%, with a precision of 94% for specific extremities and a Cohen's Kappa score of 91.5%. This proposed model highlights the efficacy of combining preprocessing techniques with Deep CNN for high-precision bone abnormality detection.

Keywords: *Deep CNN, Musculoskeletal Abnormality, Data Augmentation, Radiographs, MURA, ADAM optimizer and Computer Vision*

1. Introduction

Global Burden of Disease, also known as GBD, found that abnormalities of musculoskeletal diseases were the second highest contributor to disability and lower back pain across the world in a study in 2016. The study concluded that approximately 20% to 30% of people worldwide live tormented with musculoskeletal abnormalities that go undetected sometimes [1]. Almost all of our daily work is done using the upper extremities of our body. An abnormality in the bone of an upper extremity may result in a painful, dependent life. The person is unable to perform simple daily chores on his own. The detection of the abnormality in bone depends upon the evaluation of radiographs by radiologists. After evaluating these radiographs, they make adequate treatment plans for the patients. Sometimes, radiologists or doctors become confused or misguided. This happens because, in most cases, radiology results are highly impacted by the circumstances of the patient and clinical affairs and quondam medical imaging. Under-reading was one of the most significant clinical errors, performed on a dataset consisting of 1269 errors [2][3]. A recent report enumerated that approximately one billion radiological studies are perpetrated worldwide every year, most of which are explained by radiologists [4].

1.1. Musculoskeletal Systems and Statistics

A major inter-observer discrepancy rate of 26% and an intra-observer discrepancy rate of 32% were found in a study conducted by Massachusetts General Hospital [5]. A minor discrepancy rate of 13% major and 21% minor was found in a study in 2007, which shows the influence of highly experienced neuro-radiologists who had their second reading of MR and CT studies interpreted by radiologists initially [6]. The workload of radiologists is escalating with more images, increased volume of cases, greater complexity, and decreased time to detect abnormality by radiologists, leading to burnout of radiologists. Patients having musculoskeletal problems are most likely to visit small-scale primary care centers. Almost 105 million people visited the primary medical centers in ambulance and hospital emergency departments and outpatients in the United States of America from 2009 to 2010. They all were for diseases related to the connective tissue and musculoskeletal system. Of these visits, approximately 39 million were supposed to visit the primary care centers, around 32.4 million went to surgical specialists, and about 17 million visited medical specialists. At the time of offering radiology services in a primary care center, a reduction of access issues, including a decrease in the time required to diagnose and treat, was found. This led to poor quality because of the insufficient skills and training of the radiologists. [39] Increasing radiologist workloads and escalating primary care centers of radiology make it much more relevant to explore the application and scope of Artificial Intelligence and intense learning to assist in diagnosis to radiologists and physicians of primary care centers to improve the quality of care of patients. In this paper, we have used Deep Learning for computer-aided diagnosis using image feature extraction [7][38].

1.2. Motivation

Deep learning algorithms are inspired by the magnificent human brain and are based on a structure of neural networks that are trained on a dataset to learn various discriminative features [17][18][19][20]. Injuries and diseases to the bone are some of the major factors that contribute to causing bone abnormalities. To be specific, conditions of Musculoskeletal disorders are known to affect a little more than 1.7 billion people around the world. They are considered one of the most common causes of long-term severe pain and bone disability. There are more than 30 million visits to the emergency department annually and this is

**1 Department of Computing Technologies, School of Computing, SRM Institute of Science and Technology, Chennai, India.*
Email: prakashm3@srmist.edu.in
2 Department of Computational Intelligence, School of Computing, SRM Institute of Science and Technology, Chennai, India.
**Corresponding author: prakashm3@srmist.edu.in*

increasing day by day. Hence, we need to minimize the error rate caused by the Radiologist and analyze faster [9]. The clinician must get a compiled medical diagnosis to provide an effective treatment [10]. For this purpose, we need to implement an AI solution.

1.3. Contribution

The models which are thus trained are then used to detect and diagnose. In our model, we have used Deep CNN to classify the seven upper extremities and predict the abnormality in bone using radiological images. Deep CNN takes an image from the dataset and assigns learnable biases and weights to the various aspects of the image to differentiate one image from the other. The CNN model provides a high computation rate and is very efficient. The dataset used in the model is the MURA dataset, a benchmark dataset produced by Stanford [8]. It contains radiological images of seven different types of upper. The radiological images were not clear enough for feature extraction by the model, thereby giving a lower accuracy. To increase the quality of images, we applied Histogram Equalization (HE) and Contrastive Limited Adaptive Histogram Equalization (CLAHE). These processes enhanced and improved the clarity of the image. The resulting images were of high quality, which enabled the model to extract the required features, giving a high accuracy of 92.10 percent. Our model consists of five main layers: the Convolutional Layer, Activation Layer, Batch Normalization Layer, Maxpooling Layer, and Dropout Layer. When the image is fed into the model, it classifies the image into one of the seven classes of the MURA dataset and then predicts whether the image has a normal bone or an abnormal bone. As we have used Deep CNN, the model keeps on learning and providing better results. The result consists of fourteen classes of outcomes predicting abnormality in the seven classes of the MURA dataset.

The proposed model aids in detecting abnormalities present in the upper extremity bones. The dataset that we used in our model consisted of various types of upper extremity bone abnormalities, including fractures, degenerative joint diseases, bones having hardware, and other monsters, including subluxations and lesions. A model that can detect all or more bone abnormalities mentioned above can reduce the work of radiologists to a great extent. Moreover, it will also reduce errors that occur by manually detecting bone abnormality. Minor details that may go unnoticed by the human eye can easily be detected by the model, which continues to learn. The model can reduce the delays caused in reporting by X-rays, and urgent cases can then be diagnosed as soon as possible.

1.4. Organization of the paper

The paper is organized as the following sections: Section 2 describes the related works, Section 3 describes the System Approach and Implementation, Section 4 describes the System Analysis and Performance and Section 5 describes the conclusion.

2. Related Works

Jose George et al., [11] proposed a Hybrid Wavelet technique for detecting Temporal bone abnormalities. High High-resolution computed tomography (HRCT) images were used as the dataset for the diagnosis of the ear. Histogram Equalization and Median Filtering were performed on each image to increase the contrast and remove the noise and outliers from the HRCT images. An adaptive mask was used to select the Region of Interest (ROI). The texture features were extracted from the Gray Level Co-occurrence Matrix (GLCM) and the geometric features were removed from the region of the temporal bone. Maximum Probability, Inverse Element, Entropy, Contrast, Energy, and Difference Moment were included in the texture features. Calculation of GLCM in 0 degrees, 45 degrees, and 90-degree directions was used to extract the 15 features, which in turn was used for classification. Wavelet Support Vector Machine was the technique that was used for classification.

Tusher Chandra et al., [12] proposed a technique for detecting Bone Abnormalities in the MURA dataset using the Deep CNN based on the CADx (Computer Aided Diagnosis) model. The dataset of four of the upper extremities from the MURA dataset was used. At first, the images were normalized and smoothed using Gaussian Blur. Histogram Equalization and Adaptive Thresholding followed this process. The architectures used in the model are VGG (Visual Geometry Group) and RestNET (Residual Network). Both of these architectures were trained on the ImageNet dataset [13]. The pre-trained weight of ImageNet was collected to apply the transfer learning concept. To compile the model in both architectures, an SGD optimizer was used. The model was run for 100 epochs with early patience of 10. The area under the ROC curve (AUROC) was used to select the base model of the two architectures. This process was done for all the upper extremity datasets separately. To further increase the accuracy, the ensemble technique was applied to the model. When the results were compared, the performance of the Ensemble model was better than the VGG and RestNET1. The performance of the model was also compared to MURA and 3 other radiologists, where the model had a better result than MURA with two other radiologists. But in this paper, only four of the upper extremities in the MURA dataset were used. Also, the individual accuracy of these four upper extremities was not satisfied.

N.Umadevi and S.N.Geethalakshmi [14] proposed a method was created to find bone fractures in X-ray images. It involved enhancing the images and then identifying fractures. Enhancements reduced noise using techniques like Independent Component Analysis (ICA) and wavelets. The fracture identification used active contour modeling and a region-growing algorithm to pinpoint key areas. To emphasize important details, texture and shape features were extracted. Finally, binary classifiers like Back Propagation Neural Network (BPNN), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) were used to detect bone fractures, aiming to improve accuracy.

Róża Dzierżak et al. [15] proposed a transfer learning technique to predict the probability of abnormality in bone using a VGG 19.

Pranav Rajpurkar et al. proposed a technique to predict the probability of abnormality in a study of musculoskeletal abnormalities detection using a 169-layer convolutional neural network [21]. The architecture used in the model was DenseNet. The fully connected final layer was replaced with a single output layer. Sigmoid Nonlinearity was applied to the images. To get a similar mean and standard deviation in the ImageNet training set, normalization was performed in each image of the dataset. The images were scaled to a size of 320×320. After that, data augmentation was performed using lateral inversions and rotations of up to 30 degrees. This was followed by the initialization of weights performed by using weights of a pre-trained model on Imagenet [16]. Adam Optimizer was used for the end-to-end training of the model. Then, the five models were ensembled having the lowest validation loss. In this paper, the

performances of the model in some classes were not good, especially in Humerus and Wrist classes.

3. System Approach and Implementation

A Deep Convolutional Neural Network was used on the benchmark MURA dataset provided by Stanford for the classification of different parts of the upper extremities as well as to detect bone abnormality in musculoskeletal radiographs [37]. The data is preprocessed using suitable image enhancement algorithms and then the processed data is used to train the Convolutional Neural Network [22] [23].

3.1. Proposed Methodology

Fig. 1 shows the complete working of system architecture. The technique of Histogram Equalization is used for preprocessing the images. The images were resized to a fixed size of 400×400 [32-36]. The enhanced images were fed into the neural network which is a multiclassifier model. The model classifies the image into the seven types of upper extremities as well as predicts whether it is normal or abnormal. This classification and prediction results in fourteen outcomes.

3.2. Software and Libraries Used

The entire algorithm was developed in Jupyter Notebook using Python 3. Keras was used as the Neural Network API. The Python Image Library (PIL) was used to load the images in the dataset. The library cv2 was used for the extraction of the images. Techniques like Histogram Equalization and CLAHE were applied and the images were resized using the cv2 library. The total number of images for testing and training is less, to overcome this issue, image Augmentation techniques are used [24-26].

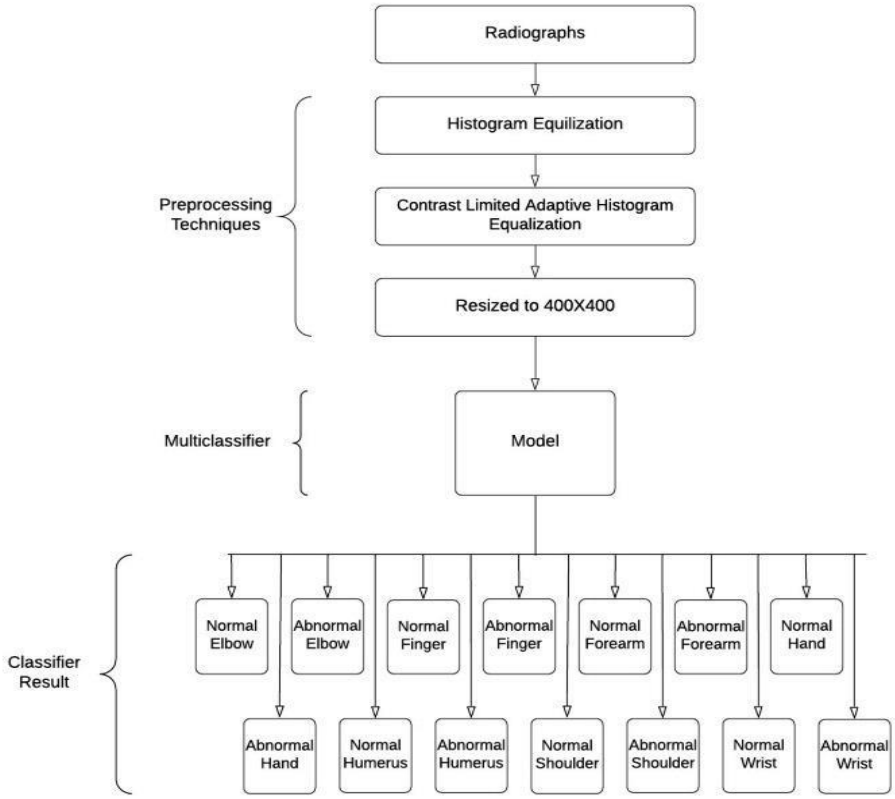

Fig. 1. System Architecture

The NumPy library was used to convert all the images into a simple NumPy array so that they could be formatted and used for data extraction. The dataset also contained a CSV file containing information about each image location and also whether the image was normal or abnormal. So, to read and use the data, Pandas library was used. Its pattern rules follow standard Unix path expansion rules. Moreover, it is predicted that it is faster than other methods. The Seaborn library was used for the creation of the Confusion Matrix and the sklearn library was used to calculate various performance metrics such as Precision, Cohen's Kappa, and Accuracy metrics.

3.3. Data Collection

The MURA dataset was collected by a view board of an institution and is made available publicly for research on Bone Abnormalities. MURA is a dataset provided by Stanford of HIPAA-compliant musculoskeletal radiographs which are de-identified and consist of 14,863 radiographic studies from approximately 12,173 patients that constitute a total of about 40,005 radiographic images which are multi-view in nature and are from the Picture Archive and Communication System (PACS) from the Stanford Hospital. Each of the radiographic studies belongs to one of the seven types of standard upper extremities: the elbow, forearm, humerus, wrist, finger, hand, and shoulder. Each radiographic study has been labeled manually as abnormal or normal by radiologists board-certified from Stanford Hospital.

3.4. Data Preprocessing

Before feeding the data into our neural network, it needs various pre-processing techniques. Training and testing datasets were combined and the K fold cross-validation technique was applied, which selects the different distributions of data samples. The

dataset for the humerus and forearm was less as compared to the dataset of other types of upper extremities. Hence, the combined dataset will result in more images [27-31].

Table 1 Distribution of MURA Dataset

Study	Training		Testing		Total
	Abnormal	Normal	Abnormal	Normal	
ELBOW	2006	2925	230	235	5396
FINGER	1968	3138	247	214	5567
HUMERUS	599	673	140	148	1560
WRIST	3987	5765	295	364	10411
HAND	1484	4059	189	271	6003
FOREARM	661	1164	151	150	2126
SHOULDER	4168	4211	278	285	8942
TOTAL	14873	21935	1530	1667	40005

Table 1, Shows that to preprocess the dataset, we divided the dataset having seven subclasses positive (abnormal) and negative (normal). This results in a dataset consisting of fourteen classes which will be used for the classification of normal and abnormal bone structures along with the different types of upper extremities.

Table 2 Distribution of Combined Images

Study	Abnormal	Normal
ELBOW	2236	3160
FINGER	2215	3352
FOREARM	812	1314
HAND	1673	4330
HUMERUS	739	821
SHOULDER	4446	4496
WRIST	4282	6129

Even after combining the training and validation images, the ratio of the number of images is not good enough to feed into the neural network. Hence, there is a need for Data Augmentation. We used Python Image Library also known as PIL for this purpose. PIL is used to open, manipulate, and save many different image file formats. PIL was used to read the images of the MURA dataset one by one. Open Source Computer Vision Library, also known as OpenCV, was used for feature extraction and to improve the quality of the images. Histogram equalization (HE) and Contrastive Limited Adaptive Histogram Equalization (CLAHE) were used to get a more enhanced image. Fig. 2 shows that Histogram Equalization is a technique used for computer image processing to give the image an improved contrast. In Contrastive Limited Adaptive Histogram Equalization, a transformation function is derived from each neighbourhood when applying the contrast limiting procedure. CLAHE prevents the over-amplification of noise that is risen by the adaptive histogram equalization. The images in our dataset are of different sizes. For the neural networks to perform better, we need images that are of the same size. Thus, we converted all the images into a standard size of 400×400. So the images will not lose their quality and the convolutional neural network will perform better giving higher accuracy.

Fig. 2. Transformation of the image using HE and CLAHE

Table 2 reveals fewer Humerus and forearm images compared to other categories, requiring equal class distribution. Data Augmentation, demonstrated in Fig. 3, involves generating new training data from existing samples. we expanded the dataset by flipping and rotating, resulting in 60,803 images with nearly equal class representation. Following data augmentation to achieve a balanced class distribution, we partition the dataset into an 80:20 split, with 80% designated for training and 20% for testing. The 20% allocation for testing comprises images that have not been previously seen by our model, ensuring an unbiased evaluation. Table 3 shows the number of images in each class after dividing them into training and validation datasets. Altogether 48649 images were used for training the model and 12154 images were the ones that the model had never seen and hence were used for testing.

Fig. 3. Augmentation of data (a) Original image, (b) Flipped image, (c) Rotated image

3.5. Network Architecture and Design

Deep Learning outshines traditional Machine Learning as it automates feature extraction from images, eliminating manual effort. During training, it independently identifies and combines features, expediting the learning process. Our research on Deep CNN architectures led to a superior model for classifying upper extremity segments and predicting bone abnormalities.

Table 3 Distribution of Dataset Used

STUDY	TRAINING	TESTING
ELBOW NEGATIVE	3700	925
ELBOW POSITIVE	3445	861
FINGER NEGATIVE	3471	867
FINGER POSITIVE	3015	753
FOREARM NEGATIVE	3732	932
FOREARM POSITIVE	3328	832
HAND NEGATIVE	4047	1011
HAND POSITIVE	2788	696
HUMERUS NEGATIVE	3339	834
HUMERUS POSITIVE	3279	819
SHOULDER NEGATIVE	3368	841
SHOULDER POSITIVE	3335	833
WRIST NEGATIVE	4612	1153
WRIST POSITIVE	3190	797
TOTAL	48649	12154

Fig. 4 Shows, In our proposed model, we are using 5SET (5S), which starts with a convolutional layer as they are the main building blocks of a neural network and are used for feature extraction from the image. Each convolutional layer is followed by an activation function which has the purpose of introducing non-linearity into the output of each neuron. The Activation Function which we have used is Rectified Linear Unit (ReLU). The following equation depicts the working of the ReLU Activation Function.

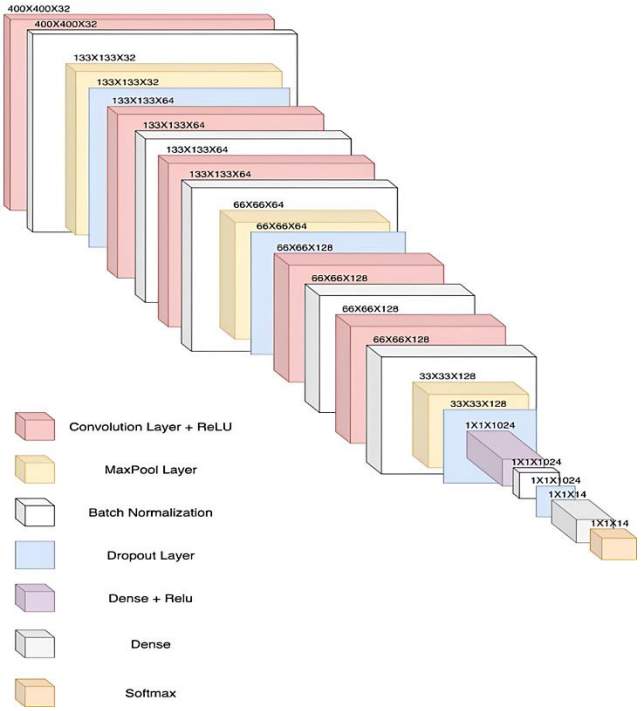

Fig. 4. 5 S Model Architecture

$$f(x) = \max(0, x) \quad (1)$$

Then comes the batch normalization layer which is responsible for improving a neural network's speed, performance, and stability. It normalizes the input by adjusting and scaling the output of the activation layer. This is followed by a max pooling layer represented in (Eq. (1)). A pooling layer is one more building block of the convolutional neural network which progressively minimizes or maximizes the spatial size of the representation and thereby reduces the computation and the number of parameters used. It operates on each of the feature maps independently.

When feature extraction is done, max pooling is the most preferred type of pooling. This layer is followed by a dropout layer which is used to prevent the model from overfitting as it randomly selects neurons according to a set percentage and does not consider the output of those neurons or considers them as 0. After this layer, we have a fully connected layer (FC) which converts the entire matrix into a single 1D vector as it is followed by 2 dense layers which require input in a 1D vector. The purpose of the dense layer is to feed all the outputs from the previous layer into its neurons and every single neuron provides output to the next layer. The first Dense Layer is followed by an Activation function, Rectified Linear Unit (ReLU).

The second dense layer provides the final output to classify the image into one of the fourteen classes and is followed by an Activation Function, Softmax. The following equation (Eq. (2)) depicts the working of the Softmax Activation Function.

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_i}} \; for \; i = 1, .., K \; and \; z = (z_1, \ldots, z_K) \epsilon R^K \quad (2)$$

We compiled the model using the loss function as categorical cross-entropy and three optimizers with different learning rates were used to get the best accuracy. The first optimizer used was Adam Optimizer, which is an adaptive learning rate method, that computes learning rates for all the parameters individually. After that, the Adagrad optimizer was used, which is a gradient-based optimizer that adapts the learning rate to the parameters, performing larger updates for parameters associated with features that are not frequent and smaller updates for parameters that have features occurring frequently. At last, Adadelta optimizers were used, which is an extension of Adagrad and reduce its aggressive, monotonically decreasing learning rate to achieve more accuracy. Our model has 143,036,302 parameters, out of which 143,033,422 are trainable parameters, and 2880 are non-trainable parameters.

3.6. Algorithm

Our Algorithm is divided into three major phases. In the first phase of our algorithm, the preprocessing of data is done. After that, we move to the second phase of our model, which is model creation, and finally, we move to the third and last phase of our algorithm, which is the training and testing of the model.

Phase 1: Preprocessing

Step 1: Separate the MURA dataset into 7 folders with 2 subfolders each (Assume data is organized as needed)

Step 2: Augment data to maintain a common data distribution across each class

Step 3: Data Preprocessing and Transformation

For each image in the dataset:

Load image using imread function

Convert image from RGB to Gray

Apply Histogram Equalization

Apply CLAHE to the image

Save the preprocessed image to a folder

Step 4: Change directory structure into 14 folders for Keras Image Data Generator

After following these steps, the first phase of our algorithm was completed successfully and we moved on to the second phase of our algorithm, where we imported the libraries, defined the Convolutional Neural Network, and selected the appropriate loss function.

Phase 2: Model Creation

Step 5: Import necessary libraries and layers

Step 6: Initialize the model

```
model = Sequential()
model.add(Conv2D(32,(3, 3), input_shape=(400, 400, 1)))
model.add(ReLU())
model.add(BatchNormalization())
model.add(MaxPooling2D())
model.add(Dropout(0.25))

model.add(Conv2D(64, (3, 3)))
model.add(ReLU())
model.add(BatchNormalization())
model.add(Conv2D(64, (3, 3)))
model.add(ReLU())
model.add(BatchNormalization())
model.add(MaxPooling2D())
model.add(Dropout(0.25))

model.add(Conv2D(128, (3, 3)))
model.add(ReLU())
model.add(BatchNormalization())
model.add(Conv2D(128, (3, 3)))
model.add(ReLU())
model.add(BatchNormalization())
model.add(MaxPooling2D())
model.add(Dropout(0.25))
model.add(Flatten())
model.add(Dense(1024))
model.add(ReLU())
model.add(BatchNormalization())
```

model.add(Dropout(0.5))
model.add(Dense(14, activation='softmax'))
Step 7: Compile the model
model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
When model creation was successfully done, we moved on to our final phase where we split the data into training and testing, trained the model on the training images, and tested the model on the testing images.
Phase 3: Training and Testing
Step 8: Using keras Image Data Generator for Splitting 60803 images into 48649 for training and 12154 for Validation
Step 9: Training the model with Adam as first Optimizer
Step 10: After reaching peak accuracy with Adam, optimizer is changed to Adagrad
Step 11: when reached the peak accuracy with Adagrad, Optimizer is again changed to Adadelta
Step 12: Save the model and Load for Testing
Step 13: Testing the model with 12154 Validation Images and getting the result

4. System Analysis and Performance

The existing optimizer architecture uses upto 4 upper extremities, used for the classification of upper extremities as well as for the prediction of bone abnormality but the accuracy were not satisfactory. The proposed 5s model was most suitable and produces higher accuracy. We used different optimizers such as Adam, Adagrad, and Adadelta, which aided in increasing the accuracy of the model. For the enhancement of radiograph images, we used Histogram Equalization but it was not enough because the features required for extraction and learning were not completely visible to our model thereby decreasing its accuracy. So, we needed to enhance the image more which is why we used CLAHE along with Histogram Equalization which enhanced the images to an extent that the features required for extraction were fairly visible to the model and it was able to extract and learn more features correctly. The enhanced images made it easier for the model to learn and predict accurately. For the training of our model, 48649 images were used and 12154 images were used for testing. The model classified these images into the seven classes of upper extremities and predicted their abnormality as well. As a result of this, together fourteen types of outcomes were predicted, which were the normality and abnormality of each class. We got much better accuracy for each of the outcomes.

4.1. Analysis of Algorithm

To increase the performance of our model, we have used three optimizers. At first, we used Adam Optimizer. It resulted in a training accuracy of 22% and a validation accuracy of 23%. Fig. 5 shows the accuracy and loss of the Adam Optimizer.

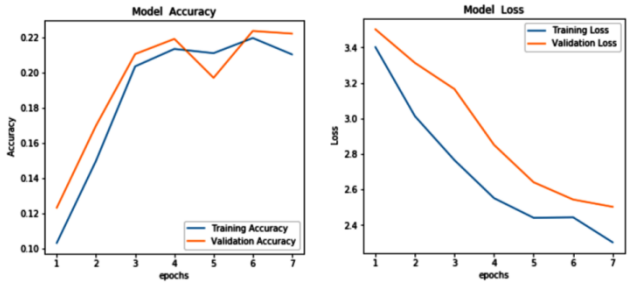

Fig. 5. (a) Accuracy of Adam Optimizer (b) Loss of Adam Optimizer

After that, we used Adagrad Optimizer. It resulted in a training accuracy of 68% and validation accuracy of 70%. Fig. 6 shows the accuracy and loss of the Adagrad Optimizer.

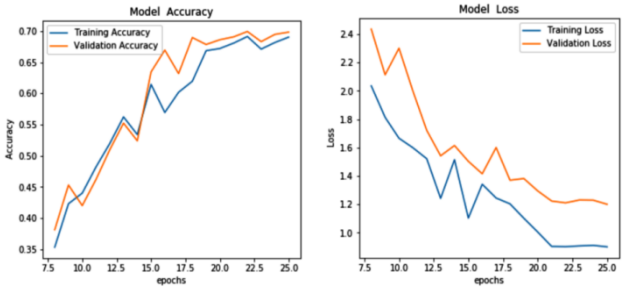

Fig. 6. (a) Accuracy of Adagrad Optimizer (b) Loss of Adagrad Optimizer

After that, we used Adadelta Optimizer. Table 4 shows the result of training accuracy of 92% and a validation accuracy of 92.10%. Fig. 7 shows the accuracy and loss of the Adadelta Optimizer.

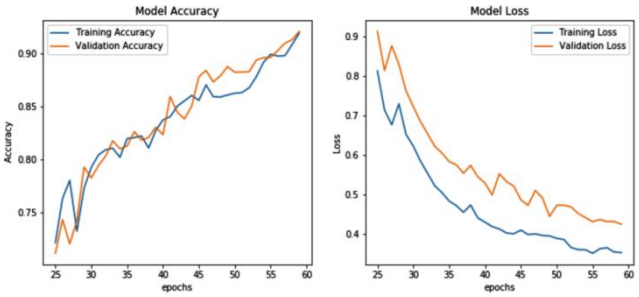

Fig. 7. (a) Accuracy of Adadelta Optimizer (b) Loss of Adadelta Optimizer

Table 4 Accuracy of the Model

OPTIMIZER	ACCURACY	
	TRAINING	VALIDATION
ADAM	22%	23%
ADAGRAD	68%	70%
ADADELTA	92%	92.10%

The proposed model classified the images into seven classes of upper extremities and further predicted them for abnormality. For each category, the result is evaluated using a confusion matrix. A confusion matrix is formed for the actual condition and the predicted condition, which consists of parameters like TP, FN, FP, and TN for the assessment, where TP means True Positive (Identified Correctly), TN means True Negative (Identified Incorrectly), FP means False Positive (Rejected Correctly), FN means False Negative (Rejected Incorrectly). We use these parameters to derive the required results from the output received from the model. These results will later help in deriving the performance of our designed model. The evaluation according to these parameters (TP, TN, FP, FN) helps us to get significant results.
These metrics are referred from the data mining algorithms and includes various evaluation metrics used to assess the performance. The evaluation metrics include Accuracy (Eq. (3)), Precision (Eq. (4)), Sensitivity (Eq. (5)), Specificity (Spec) (Eq. (6)), MissRate (Eq. (7)), Fallout (Eq. (8)), Cohen's Kappa Score (Eq. (9)), F1-score (Eq. (10)).

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + FP + TN + FN}} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Sensitivity/Recall = \frac{TP}{TP + FN} \quad (5)$$

$$Specificity = \frac{TN}{FP + TN} \quad (6)$$

$$MissRate = 1 - \frac{TP}{TP + FN} \quad (7)$$

$$Fallout = 1 - \frac{TN}{FP + TN} \quad (8)$$

$$Cohen's\ Kappa = \frac{Accuracy - Expected}{1 - Expected} \quad (9)$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (10)$$

The result of each category of upper extremity has been described below.

1) ELBOW*:* Table 5 shows the Confusion Matrix for the Elbow class.

For the Elbow class, the Precision for Normal is 87% and for Abnormal, it is 96%. We also calculated the Recall and F1 Score which are 97% and 92% for the Normal Elbow, and 93% and 94% for the Abnormal Elbow. The Miss Rate for Normal is 3.4% and for Abnormal, it is 6.7%.

Table 5 Confusion Matrix Parameters of ELBOW

CLASSIFICATION	TP	TN	FP	FN
NORMAL	893	11099	130	32
ABNORMAL	803	11256	37	58

Table 6 Confusion Matrix Parameters of Finger

CLASSIFICATION	TP	TN	FP	FN
NORMAL	832	11182	105	35
ABNORMAL	668	11377	24	85

2) FINGER*:* Table 6 shows the Confusion Matrix for the Finger class.

For the Finger class, the Precision for Normal is 89% and for Abnormal, it is 97%. We also calculated the Recall and F1 Score which is 96% and 92% for the Normal Finger, and 89% and 92% for the Abnormal Finger. The Miss Rate for Normal is 4% and for Abnormal, it is 11.2%.

3) FOREARM: Table 7 shows the Confusion Matrix for the Forearm class.

Table 7 Confusion Matrix Parameters of Forearm

CLASSIFICATION	TP	TN	FP	FN
NORMAL	848	11159	63	84
ABNORMAL	729	11282	40	103

For the Forearm class, the Precision for Normal is 93%, and for Abnormal, it is 95%. We also calculated the Recall and F1 Score, which are 91% and 92% for the Normal Forearm and 88% and 91% for the Abnormal Forearm. The Miss Rate for Normal is 9% and for Abnormal, it is 12.4%.

4) HAND*:* Table 8 shows the Confusion Matrix for the Hand class.

Table 8 Confusion Matrix Parameters of HAND

CLASSIFICATION	TP	TN	FP	FN
NORMAL	946	11049	94	65
ABNORMAL	619	11412	46	77

For the Hand class, the Precision for Normal is 91% and for Abnormal, it is 93%. We also calculated the Recall and F1 Score which are 94% and 92% for the Normal Hand, and 89% and 91% for the Abnormal Hand. The Miss Rate for Normal is 6.4% and for Abnormal, it is 11%.

5) HUMERUS*:* Table 9 shows the Confusion Matrix for the Humerus class.

For the Humerus class, the Precision for Normal is 92% and for Abnormal, it is 94%. We also calculated the Recall and F1 Score which is 93% and 93% for the Normal Humerus, and 89% and 91% for Abnormal Humerus. The Miss Rate for Normal is 6.8% and for Abnormal, it is 11.2%.

Table 9 Confusion Matrix Parameters of HUMERUS

CLASSIFICATION	TP	TN	FP	FN
NORMAL	777	11256	64	57
ABNORMAL	727	11286	49	92

6) SHOULDER*:* Table 10 shows the Confusion Matrix for Shoulder class.

Table 10 Confusion Matrix Parameters of SHOULDER

CLASSIFICATION	TP	TN	FP	FN
NORMAL	800	11242	71	41
ABNORMAL	765	11283	38	68

For the Shoulder class, the Precision for Normal is 92% and for Abnormal, it is 95%. We also calculated the Recall and F1 Score which are 95% and 93% for the Normal Shoulder, and 92% and 94% for the Abnormal Shoulder. The Miss Rate for Normal is 4.9% and for Abnormal, it is 8.1%.

7) WRIST*:* Table 11 shows the Confusion Matrix for the Wrist class.

Table 11 Confusion Matrix Parameters of WRIST

CLASSIFICATION	TP	TN	FP	FN
NORMAL	1087	10867	134	66
ABNORMAL	708	11300	57	89

For the Wrist class, the Precision for Normal is 89%, and for Abnormal, it is 93%. We also calculated the Recall and F1 Score of 94% and 92% for the Normal Wrist, and 89% and 91% for the Abnormal Wrist. The Miss Rate for Normal is 5.7% and for Abnormal, it is 11.1%.

As our model is a multi-classifier, we cannot get accuracy and Cohen's Kappa Score for every class separately. The overall Accuracy and Loss of the model are depicted in Fig. 8. So, the overall accuracy of our model is 92.10% and the overall Cohen's Kappa score is 91.5%.

4.2. Comparison of Results

Musculoskeletal abnormalities are very important to detect in the early stages as they can affect the ligaments, tendons, bones, muscles, discs, blood vessels, nerves, and various other parts of the body. Tusher, Hasib, and Hashem published a deep CNN based model for the detection of Bone Abnormality in the MURA dataset. They used only four of the seven upper extremities for their experiment which were Elbow, Finger, Humerus, and Wrist.

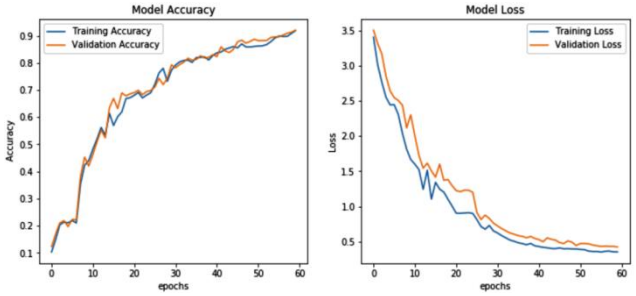

Fig. 8. (a) Accuracy of the Model (b) Loss of the Model

Table 12 Comparison of the Proposed Model and Existing Model

Upper Extremity	Proposed Model		Existing Model	
	Miss Rate	F1 Score	Miss Rate	F1 Score
ELBOW	5%	93%	18%	85%
FINGER	7.60%	92%	8%	83%
HUMERUS	9%	92%	13%	87%
WRIST	8.40%	91.50%	7%	89%

Table 12 shows the comparison of the proposed and existing models. The accuracy they got for these classes was 86.45%, 82.13%, 87.15%, and 87.86%. The overall accuracy of our model is 92.10%. In our model, the precision for Elbow is 91.5%, for the Finger is 93%, for Humerus is 93%, and for the Wrist is 91%. Their paper used VGG-19 and ResNet neural networks for their model. The Miss Rate of the four classes, Elbow, Finger, Humerus, and Wrist, which they studied are 18%, 8%, 13%, and 7%. The Miss Rate calculated in our model was 5% for Elbow, 7.6% for Finger, 9% for Humerus, and 8.4% for Wrist. As we can see, for Elbow, Finger, and Humerus, the Miss Rate of our model is less than the compared model, but for Wrist classes, the Miss Rate of our model is slightly higher than the above model.

F1 score helps us to get a balance between Precision and Recall. In the previous model, the F1 Score was 85% for the Elbow class, 83% for the Finger class, 87% for the Humerus class, and 89% for the Wrist class. Compared to our model, the F1 Score for Elbow, Finger, Humerus, and Wrist is 93%, 92%, 92%, and 91.5%, higher than the previous model. Cohen's Kappa Score in the prior model was 71.1%, whereas, for our model, it is 91.5%, which is much higher than the previous model.

5. Conclusion

Computer-aided detection in medical imaging is vital for reducing radiologists' workload and minimizing error rates in patient reports. Based on Deep CNN, the model enhances prediction accuracy by self-learning from radiological images in the MURA dataset, covering seven upper extremities. Through preprocessing techniques and well-designed architecture, the model achieves an impressive 92.10% accuracy overall, with a remarkable 94% precision precisely for extremities and a Cohen's Kappa score of 91.5%. This research paper advocates the adoption of deep learning algorithms in medical imaging to streamline radiologists' work and enable quicker, more precise therapeutic decisions, benefiting patient care significantly. The success of this model underscores the potential for artificial intelligence to revolutionize medical diagnosis and treatment, ultimately improving patient outcomes and healthcare efficiency.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016, 2017. A systematic analysis for the global burden of disease Study 2016. Lancet, 390(10100), pp.1211-1259. https://doi.org/10.1016/S0140-6736(17)32154-2.

[2] AJR Am J Roentgenol Kim YW, Mansfield LT. Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors. Pages 465–70, 2014. DOI: 10.2214/AJR.13.11493

[3] Berlin, L., 2014. Radiologic errors, past, present and future. Diagnosis, 1(1), pp.79-84. https://doi.org/10.1515/dx-2013-0012

[4] 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. https://doi.org/10.1148/rg.2015150023.

[5] 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.

[6] 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.

[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)" IEEE 2018.

[8] Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., Yang, B., Zhu, K., Laird, D., Ball, R.L. and Langlotz, C., 2017. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957. https://doi.org/10.48550/arXiv.1712.06957

[9] Mennatallah M.Abdel, Sara H.Awwad, Farah H.Ahmed, Amira G.Wasfi, Taraggy M.Ghanim, Ayman M.Nabil, "Survey: Automatic Recognition of Musculoskeletal Disorders from Radiographs" IEEE 2018.

[10] Azian Azamimi Abdullah, Atieqah Yaakob, and Zunaidi Ibrahim, "Pre- diction of Spinal Abnormalities using Machine Learning Techniques" IEEE 2018.

[11] George, J., Subin, T.K. and Rajeev, K., 2008, November. Detection of temporal bone abnormalities using hybrid wavelet Support Vector Machine classification. In TENCON 2008-2008 IEEE Region 10 Conference (pp. 1-6). IEEE. DOI: 10.1109/TENCON.2008.4766549

[12] Tusher Chandra Mondol, Hasib Iqbal and MMA Hashem, "Deep CNN-Based Ensemble CADx Model for Musculoskeletal Abnormality Detection from Radiographs" 5th International Conference on Advances in Electrical Engineering (ICAEE). IEEE 2019.

[13] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25. DOI: 10.1145/3065386

[14] 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.

[15] Dzierżak, R. and Omiotek, Z., 2022. Application of deep convolutional neural networks in the diagnosis of osteoporosis. Sensors, 22(21), p.8189. https://doi.org/10.3390/s22218189

[16] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). IEEE.

[17] Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. and Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging, 35(5), pp.1285-1298.

[18] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep Learning. Nature, 521, 436-444. http://dx.doi.org/10.1038/nature14539

[19] Salamon, J. and Bello, J.P., 2017. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24(3), pp.279-283. https://doi.org/10.3390/s23156972

[20] Zhang, B., Yu, K., Ning, Z., Wang, K., Dong, Y., Liu, X., Liu, S., Wang, J., Zhu, C., Yu, Q. and Duan, Y., 2020. Deep learning of lumbar spine X-ray for osteopenia and osteoporosis screening: A multicenter retrospective cohort study. Bone, 140, p.115561.

[21] Chen, J., Kang, X., Liu, Y. and Wang, Z.J., 2015. Median filtering forensics based on convolutional neural networks. IEEE Signal Processing Letters, 22(11), pp.1849-1853.

[22] Li, Z., Tian, Q., Ngamsombat, C., Cartmell, S., Conklin, J., Filho, A.L.M.G., Lo, W.C., Wang, G., Ying, K., Setsompop, K. and Fan, Q., 2022. High-fidelity fast volumetric brain MRI using synergistic wave-controlled aliasing in parallel imaging and a hybrid denoising generative adversarial network (HDnGAN). Medical Physics, 49(2), pp.1000-1014. DOI: 10.1002/mp.15427

[23] Smith, L.N., 2017, March. Cyclical learning rates for training neural networks. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 464-472). IEEE. https://doi.org/10.48550/arXiv.1506.01186

[24] Huang, G., Liu, S., Van der Maaten, L. and Weinberger, K.Q., 2018. Condensenet: An efficient densenet using learned group convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2752-2761). https://doi.org/10.48550/arXiv.1711.09224

[25] Mall, P.K., Singh, P.K., Srivastav, S., Narayan, V., Paprzycki, M., Jaworska, T. and Ganzha, M., 2023. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. Healthcare Analytics, p.100216. https://doi.org/10.1016/j.health.2023.100216

[26] Ilesanmi, A.E., Ilesanmi, T. and Gbotoso, A.G., 2023. A systematic review of retinal fundus image segmentation and classification methods using convolutional neural networks. Healthcare Analytics, p.100261. https://doi.org/10.1016/j.health.2023.100261

[27] Breit, H.C., Varga-Szemes, A., Schoepf, U.J., Emrich, T., Aldinger, J., Kressig, R.W., Beerli, N., Buser, T.A., Breil, D., Derani, I. and Bridenbaugh, S., 2023. CNN-based evaluation of bone density improves diagnostic performance to detect osteopenia and osteoporosis in patients with non-contrast chest CT examinations. European Journal of Radiology, 161, p.110728.

[28] Eckardt, J.N., Middeke, J.M., Riechert, S., Schmittmann, T., Sulaiman, A.S., Kramer, M., Sockel, K., Kroschinsky, F., Schuler, U., Schetelig, J. and Röllig, C., 2022. Deep learning detects acute myeloid leukemia and predicts NPM1 mutation status from bone marrow smears. Leukemia, 36(1), pp.111-118.

[29] Lee, S.J. and Pickhardt, P.J., 2017. Opportunistic screening for osteoporosis using body CT scans obtained for other indications: the UW experience. Clinical Reviews in Bone and Mineral Metabolism, 15, pp.128-137.

[30] Lorentzon, M., Johansson, H., Harvey, N.C., Liu, E., Vandenput, L., McCloskey, E.V. and Kanis, J.A., 2022. Osteoporosis and fractures in women: the burden of disease. Climacteric, 25(1), pp.4-10.

[31] Zhen, L., Zhang, Y., Yu, K., Kumar, N., Barnawi, A. and Xie, Y., 2021. Early collision detection for massive random access in satellite-based internet of things. IEEE Transactions on Vehicular Technology, 70(5), pp.5184-5189.

[32] Ding, F., Zhu, G., Li, Y., Zhang, X., Atrey, P.K. and Lyu, S., 2021. Anti-forensics for face swapping videos via adversarial training. IEEE Transactions on Multimedia, 24, pp.3429-3441.

[33] Han, S., Oh, J.S. and Lee, J.J., 2022. Diagnostic performance of deep learning models for detecting bone metastasis on whole-body bone scan in prostate cancer. European Journal of Nuclear Medicine and Molecular Imaging, pp.1-11.

[34] Tan, L., Yu, K., Shi, N., Yang, C., Wei, W. and Lu, H., 2021. Towards secure and privacy-preserving data sharing for COVID-19 medical records: A blockchain-empowered approach. IEEE Transactions on Network Science and Engineering, 9(1), pp.271-281.

[35] Grauhan, N.F., Niehues, S.M., Gaudin, R.A., Keller, S., Vahldiek, J.L., Adams, L.C. and Bressem, K.K., 2021. Deep learning for accurately recognizing common causes of shoulder pain on radiographs. Skeletal Radiology, pp.1-8.

[36] Ren, M. and Yi, P.H., 2022. Deep learning detection of subtle fractures using staged algorithms to mimic radiologist search pattern. Skeletal Radiology, pp.1-9.

[37] S. Rani B, G. B, S. G Shivaprasad Yadav, G. Shivakanth and M. B M, "Deep Learning Based Cancer Detection in Bone Marrow using Histopathological Images," 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS), Raichur, India, 2023, pp. 1-8, doi: 10.1109/ICICACS57338.2023.10100116.

[38] Dimlo, UM Fernandes, R. Umanesan, Jonnadula Narasimharao, N. Senthamilarasi, P. S. Ranjit, B. Balaji, I. Thamarai, and Vijay Kumar Dwivedi. "Optimal Configuration Planning of Multi-Energy Systems using Optimization-based Deep Learning Technique." *Electric Power Components and Systems* (2023): 1-16.

[39] A. Sathishkumar, S. Majji, T. Radhika Patnala, S. Rao Karanam, A. Kumar and M. Malyadri, "Experimentation Methodology of Orthogonal Frequency Division Multiplexing Signals Process using Radio over Fiber (RoF) system," 2022 International Virtual Conference on Power Engineering Computing and Control: Developments in Electric Vehicles and Energy Sector for Sustainable Future (PECCON), Chennai, India, 2022, pp. 01-05, doi: 10.1109/PECCON55017.2022.9850996.