Enhanced Bone Fracture Diagnosis in X-rays Using Fine-Tuned DenseNet169 Deep Learning Model

1st *Ali Orangzeb Panhwar
Dr. A. H. S. Bukhari Postgraduate Centre of Information and Communication Technology, University of Sindh, Jamshoro, Pakistan
SZABIST University, Gharo
ali.orangzeb@ghr.szabist.edu.pk
https://orcid.org/https://orcid.org/0009-0004-3885-8894

2nd Shahzad Memon
Department of Computer Science and Digital Technologies Scool of Architecture, Compuing and Engineering University of East London, United Kingdom
s.memon@uel.ac.uk
https://orcid.org/https://orcid.org/0000-0003-3354-5798

3rd Lachhman Das Dhomeja
Dr. A. H. S. Bukhari Postgraduate Centre of Information and Communication Technology, University of Sindh, Jamshoro, Pakistan
lachhman@usindh.edu.pk
https://orcid.org/https://orcid.org/0000-0002-2016-1093

4th Mukesh Prasad
Faculty of Engineering and Information Technology, University of Technology Sydney, Australia
Mukesh.Prasad@uts.edu.au
https://orcid.org/https://orcid.org/0000-0002-7745-9667

5th Asghar Ali Chandio
Department of Artificial Intelligence
Quaid-e-Awam University of Engineering, Science and Technology (QUEST) Nawabshah
Sindh, Pakistan

***Abstract*— The classification of bone fractures from radiographs is an important yet challenging task in clinical diagnosis. Diagnosing fractures through X-rays remains difficult for orthopedic specialists due to image quality issues, which can result in errors, misalignments, and potential harm to patients. However, recent advancements in artificial intelligence (AI) and deep learning have revolutionized medical imaging, with state-of-the-art methods now capable of handling 2D and 3D images. This study focuses on deep-learning approaches for the classification and detection of bone fractures in radiograph images and aims to analyze and compare various deep-learning algorithms and techniques used in fracture detection. It also highlights current cutting-edge approaches in this field, providing insights and guidance for future research and practical applications. In this paper, the application of Fine-tuned DenseNet169 for the automated classification of bone fractures in X-ray images is explored. By using deep learning approaches, our method seeks to enhance the accuracy and efficiency of fracture detection. We trained and evaluated the DenseNet169 model on the MURA Stanford dataset and achieved 83% accuracy in distinguishing fractured and non-fractured elbow bones. The model's performance highlights the potential of DenseNet169 to assist radiologists in clinical settings, promoting better patient outcomes through prompt and reliable fracture diagnosis.**

Keywords—Convolutional neural network (CNN), Deep Learning, Bone Fracture

I. INTRODUCTION

The human skeleton, consisting of 206 bones, serves as the framework of the body and supports its movements and functions while also protecting internal organs. Historically, the treatment of bone fractures involved the use of a wooden frame machine and Steinmann pin, which was simple and less risky compared to surgery. However, there was no effective way to detect fractures until the discovery of X-rays by William Rontgen in 1895. X-rays are a type of photography that uses a cathode ray tube and were initially captured using heavy and expensive glass plates. The study focuses on the advancements in deep learning algorithms and approaches used to detect bone fractures in X-ray images and their strengths and limitations. It also highlights the current approaches within this scope, providing Perspectives for upcoming research and applications. One of the main limitations of using X-rays is the quality of the image which is poor, and the radiation exposure is high. To enhance X-ray technology, new artificial intelligence (AI) based approaches were introduced for X-ray diagnosis. A study [3] suggests that over 1.7 billion individuals could be affected by musculoskeletal disorders, which may lead to significant, chronic pain and fractures. It can take a long time to recover from bone fractures, which are common injuries that often result from accidents. With the rapid development of medical technology, the approaches and procedures for treating accident patients have changed. In addition to diagnosis, hospitals often use X-rays to diagnose fractures. However, clinical hospitals can face challenges due to lacking radiologists or orthopedists. X-ray is the most common and important type of conventional radiography for diagnosing bone fractures. Still, computed tomography (CT) [5], [6] and magnetic resonance imaging (MRI) [7] are also significant and widely used in the treatment of traumatic brain injuries and other fractures. CT scans create 3D images but have the disadvantage of a high radiation dose and high cost, while MRI creates 3D images with low sensitivity and high-dose radiation. However, radiograph images contain low-dose radiations[8]. The authors [12] developed an ensemble model developed for detecting fractures in X-ray images. This model integrates several distinct models to enhance prediction accuracy and compared [13] models VGG16 and Densenet169.

Our main contribution for bone fracture detection using deep learning:

(i) We used a modified DenseNet169 model to achieve high accuracy of bone fracture classification. This result emphasizes the model's strong ability to learn and generalize between different data sets, demonstrating its robustness and adaptability to handle different medical image data.

(ii) The MURA dataset was imbalance and it was very complex and challenging dataset.

(iii) This study employs various regularization techniques, including L1 and L2 regularization, dropout, early stopping, and data augmentation, to mitigate overfitting and enhance model performance. These methods contribute to developing a robust and reliable model for bone fracture classification.

(iv) Our contribution will help out the radiologists to overcome the problem of detection and classification of bone fracture and non-fracture. The Performance of DenseNet169 shows the great improvement Xray bone dataset.

Rajpurkar et al. [11] evaluated a DenseNet-169 model using the Stanford MURA dataset. The study found that the model outperformed radiologists, particularly in a detecting abnormality in finger and hand images. However, a notable decline in performance was observed when the model was applied to examinations of the elbow, forearm, humerus, and shoulder. This disparity across various anatomical regions within the upper extremity underscores the potential challenges and complexities associated with specific types of musculoskeletal abnormalities. In this research, we have trained DenseNet 169 on MURA focused on the local features of bone X-ray images. However, the model and achieved 83% accuracy on elbow (fracture and non-fracture) MURA datasets. It shows the cutting-edge performance as compared to [11] because Rajpurkar et al. worked on MURA datasets and the accuracy of the denseNet 169 models was 71%.

The proposed paper is organized into four sections, as outlined below:

Section II presents the proposed work along with a detailed research methodology regarding the datasets. In this section, we provide an overview of the MURA bone fracture datasets. Section III outlines the results obtained from the dataset. Finally, Section IV describe the conclusion in detail and potential future directions for the proposed structure.

II. PROPOSED WORK

The author [4] reviewed the DNN model for the classification of fracture and non-fracture bones. They used data augmentation to solve the radiograph problem of addition on small data and used these cutting-edge to increase the dataset size. Artificial intelligence approaches are now extensively used for bone fractures and non fractures. This kind of deep learning model can help to find the damage [6]. The authors [5] focused on small hand fractures and proposed the Yolo4 model with data augmentation methods, which achieved 81.91%. Most of the studies conducted compared the performance of artificial intelligence with the results of radiologists. A deep learning model was designed in this project study to classify fracture and non-fracture bones. The deep learning model is overlaid on small datasets. The model overfitting is a problem that occur when model learns from the training data and cannot generalize to new, unseen data. This occurs when the model is overly complex for the training data, and the validation data lacks sufficient diversity or coverage. The model can easily overfit small datasets because it has too much information to learn from limited data. To mitigate this, several techniques can be employed, such as optimization, early stopping, and data augmentation. During the classification, it happens when the model is too close to the training data and cannot be identified due to its poor performance on new data. This is especially pertinent for intricate models that possess numerous parameters. Adjustment techniques that incorporate a penalty in the loss function of the model can significantly limit parameter values and enhance generalization to unfamiliar data. Numerous normalization techniques exist, including L1 normalization (Lasso), L2 normalization (Ridge), and regression. These techniques apply penalties in various manners, and the choice of method is influenced by the problem and the attributes of the data. In deep learning, sorting has been explored and implemented to enhance model performance and reliability. Researchers have examined different combinations of preprocessing techniques and their impacts on performance. Furthermore, innovative deep-learning strategies like weighting, early stopping, and data augmentation have been introduced. Data augmentation has gained substantial popularity in computer vision as it enlarges dataset size during the training phase and aids in reducing the burden by discouraging the model from memorizing the training examples. The literature indicates that data augmentation can greatly enhance the performance and reliability of deep learning models through alterations to the original data such as flipping, rotating, scaling, and introducing noise. Data augmentation additionally aids in addressing class imbalance by producing extra samples for underrepresented classes, leading to a more balanced and resilient model. Besides data augmentation, L1 and L2 regularization, dropout, and early stopping are employed to manage the magnitude of the loads and avoid overfitting. In general, it is crucial for researchers creating deep learning applications to comprehend the different processing techniques and their effects on model performance. A thoughtful assessment of the trade-offs between model complexity and performance, and experimentation with different methods, can lead to better solutions for specific tasks. However, needed to explore new and innovative optimization strategies for deep learning. The study [14] explores bone fracture detection in X-ray images using Deep Learning, specifically DenseNet and VGG19 CNN architectures. The models were trained and optimized using a varied X-ray dataset to enhance the accuracy of fracture detection. Assessment through performance metrics (accuracy, precision, and recall) indicated that the CNN models surpassed conventional methods, providing high sensitivity and specificity. The research also highlights the clinical potential of these models, implying they could assist radiologists in delivering quicker, more precise diagnoses, thereby enhancing patient care and alleviating healthcare workloads. In summary, the study aids in the progress of medical image analysis and computer-assisted diagnosis in radiology.

Research Architecture

1. **Convolutional Neural Network (CNN) based Densely Connected Convolutional Networks (DenseNet)**

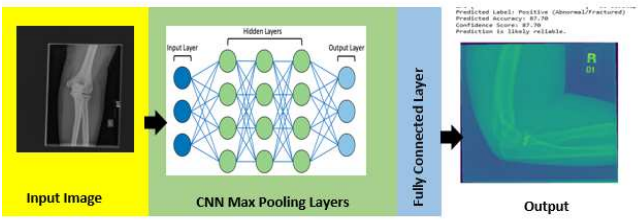

Figure. 1 *Fine Densenet 169proposed Architecture based on CNN*

a) Pre-processing:

Preprocessing is an important part of the process that uses advanced methods beyond conventional methods to improve the quality of input data. These novel pre-processing techniques are aimed at improving the model's capability to extract meaningful features from the images.

b) Data Transformation and Augmentation

The first stage is Preprocessing, where the data is loaded into the model to ensure consistency and enhance performance. Images were resized to a fixed size of 224x224 for the DenseNet169 model. The denseNet model has various variants that require a specific input size for processing images. However, we have taken the standard input size that has compatibility and best performance with the pre-trained model with data augmentation approaches, like rotation and horizontal flipping, which are incorporated using the **Keras Image_DataGenerator.** The data augmentation increases the model's ability to generalize by generating variations of the original images.

Furthermore, we have fine-tuned DeneseNet169 and compared the metric. The Researcher Stanford ML group used the DenseNet169 model on the MURA dataset.

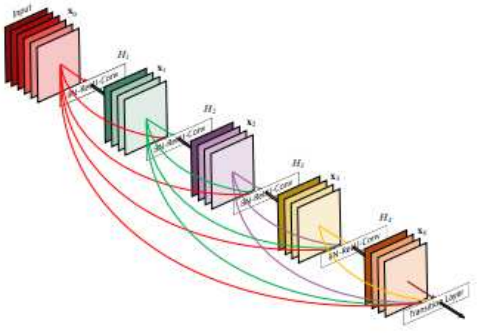

Figure No. 2 *Architecture of Densenet with 5 Layers Dense blocks* [10]

However, we did introduce nuanced modifications in our approach, particularly concerning the loss functions. Due to the significant class imbalance in the MURA dataset, training the model with the standard cross-entropy function risked biasing the model towards the majority class, potentially neglecting the minority class. To mitigate this, we employed a composite loss function that synergistically combined cross-entropy with focal loss. In addition to the loss function, we incorporated training-time augmentations such as flipping, zooming, and random cropping. These augmentations enhanced the model's ability to generalize across varied X-ray presentations. Furthermore, at the network's head, we made refinements to the fully connected (FC) layer and introduced regularization techniques, namely L1 and L2, to prevent overfitting and stabilize training.

Convolution and Pooling Layers: Initial 7x7 convolution with 64 filters, followed by batch normalization and ReLU, then a 3x3 max pooling layer**.**

Dense Block 1: This layer contains 6 bottleneck layers, each with 1x1 and 3x3 convolutions, ensuring dense connectivity with previous layers.

Dense Block 2: This layer consists of 12 bottleneck layers, where outputs of all preceding layers are concatenated for maximum feature reuse.

Dense Block 3: This block contains the 32 bottleneck layers with dense connections for feature propagation.

Dense Block 4: Contains another 32 bottleneck layers, providing further dense connectivity for deep feature extraction.

Transition Layers: This layer consists a 1x1 convolution layer is trailed by 2x2 average pooling between dense blocks, effectively managing dimensionality and controlling network growth.

RESULTS AND DISCUSSION

In this work, we used efficient data processing and reinforcement, which are important for optimizing the performance of machine learning models, especially in the Keras implementation. At the initial step, the MURA dataset is loaded using the create_images_metadata_csv function, which generates CSV files containing image paths and associated tags. Table.1 shows the results

Data Distribution for Training and Validation in Elbow Classification

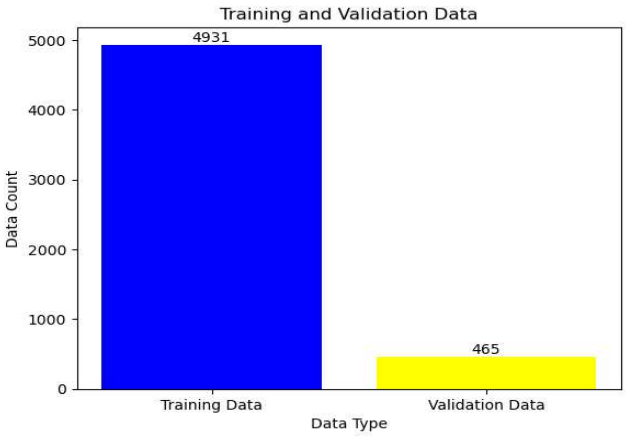

Figure No. 3. *Data Distribution for Training and Validation in Elbow Classification*

The figure no.3 shows that the dataset used in this study comprised a total of 5000 data points, with 465 allocated specifically for validation purposes and the remaining 4931 designated for training. The training data predominantly consists of 4000 images across various categories, providing a robust foundation for model learning. The validation data includes 1000 images, strategically chosen to evaluate the model's accuracy and prevent over fitting. This meticulous data distribution aids in achieving reliable outcomes.

tiveness of the training procedure in attaining steady and de pendable results.

Table 1: MURA Elbow Classification Report

Class	Precision	recall	F1 score	support
0	0.76	0.97	0.85	235
1	0.95	0.70	0.80	239
accuracy			0.83	465
Macro avg	0.86	0.83	0.83	465
Weighted avg	0.86	0.83	0.83	465

The Table no.1 describes bone fracture classification performance of the model is given in Table 1. For class 0, the precision was 0.76, with a recall of 0.97, resulting in an F1-score of 0.85 based on 235 samples. For class 1, the precision reached 0.95, but with a lower recall of 0.70, yielding an F1-score of 0.80 across 239 samples. The overall accuracy of the model was 0.83, calculated over 465 total samples. The macro-average for precision, recall, and F1-score was consistent at 0.86, 0.83, and 0.83, respectively, indicating balanced performance across classes.

Training Accuracy vs Validation Accuracy

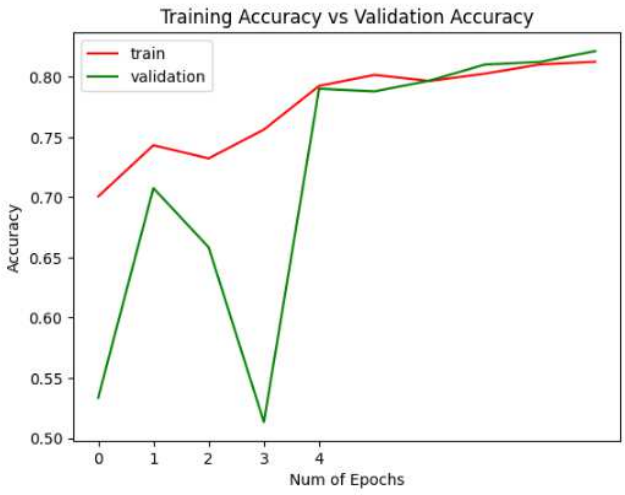

Figure No. 4 *Training Accuracy vs Validation Accuracy*

The figure no.4 shows that the model's performance was evaluated in terms of accuracy across training and validation phases over multiple epochs. The training accuracy started at 0.55 and gradually increased, reaching 0.80 by the final epoch. Similarly, validation accuracy exhibited steady improvement, beginning at 0.50 and peaking at 0.75. This upward trend that shows the model's capability to learn effectively over time, reducing the gap between training and validation accuracy and indicating robust generalization The performance consistency across epochs emphasizes the effec

Data Training Loss vs. Validation Loss

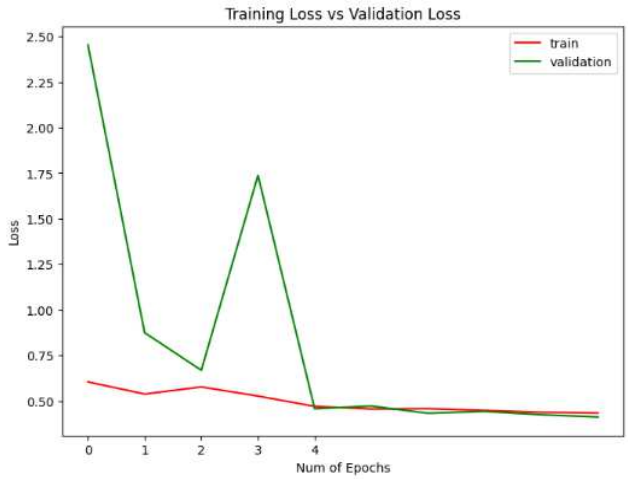

Figure No.5 *Data Training Loss vs. Validation Loss*

The figure no. 5 illustrates that the model's loss metrics were tracked throughout training and validation stages to assess optimization and convergence. At the beginning, the training loss was elevated, commencing at 2. 50, but it consistently dropped as the epochs advanced, attaining a minimum of 0. 50 by the concluding epoch. Likewise, the validation loss exhibited a downward trend, starting at 2. 25 and decreasing to around 0. 75.

Confusion Metric:

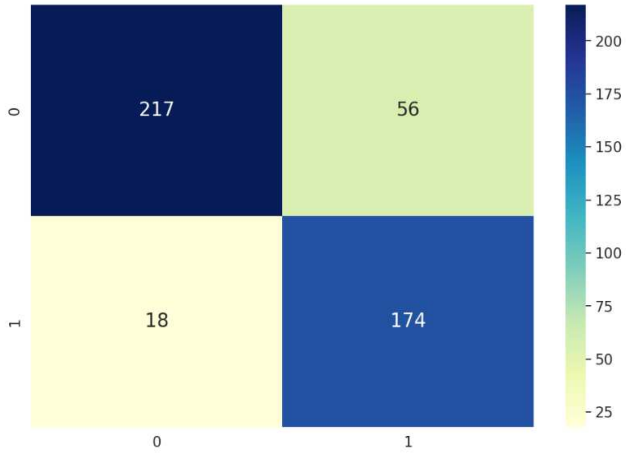

Figure No.6 *Confusion Metric*

The figure no.6 shows the data distribution reveals varying counts across different categories, highlighting imbalances that may influence model performance. Class 0 contains 200 samples, with specific subsets showing counts of 217, 56, and 150. Similarly, class 1 consists of subsets with counts of 175, 125, 100, 75, and 174. These variations suggest the need for preprocessing techniques, like resampling or weighting, to ensure equitable representation of all categories during training. Addressing this imbalance is crucial for achieving consistent and unbiased model predictions.

Receiver operating characteristic (ROC) Curve

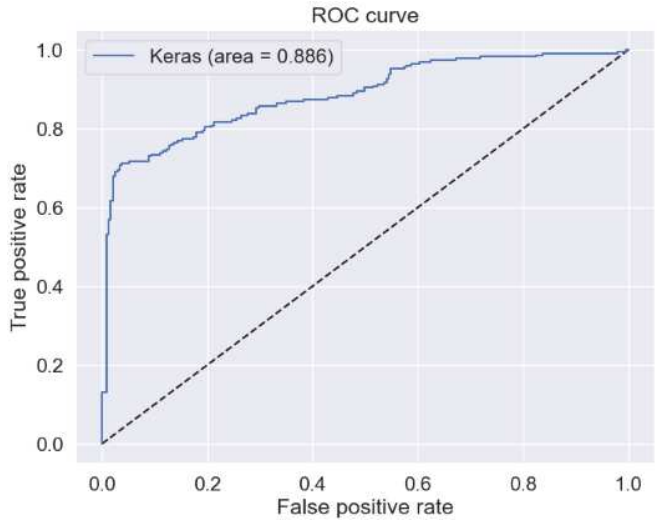

Figure No.7 *Receiver operating characteristic (ROC) Curve*

The figure no. 7 shows that model's performance was further evaluated using the Receiver Operating Characteristic (ROC) curve, which illustrates the trade-off between the TPR and the FPR across different classification thresholds. The curve demonstrated a strong AUC value of 0.886, indicating high discriminative ability. The TPR approached 1.0 at lower FPR values, reflecting the model's capability to correctly classify positive samples while minimizing false positives. This robust ROC performance signifies the model's effectiveness in distinguishing between classes with reliable accuracy.

Prediction in saved model

Figure No. 8 *Prediction in saved model*

In the evaluation of the DenseNet model implemented using Keras-TensorFlow, the following performance metrics were observed. The performance metrics associated with the elbow category achieved a Precision of 0.86, indicating that 86% of instances classified as Elbow were correctly identified. The Recall of 0.83 suggests that the model captured 83% of all actual Elbow instances. The F1 score, a harmonic mean of Precision and Recall, stands at 0.83, indicating a balanced performance. With an ROC value of 0.88, the model exhibits a high discriminative ability in distinguishing between positive and negative instances within the Elbow category. Cohen's Kappa coefficient, calculated at 0.66, indicates substantial agreement between observed and expected classifications

IV. CONCLUSION

In this paper, the implementation of DenseNet169 provides a comprehensive approach to managing the MURA dataset using the DenseNet169 architecture. The combination of efficient data loading, robust processing methods, and a well-defined model architecture contribute to the model's performance observed that it may perform well in the challenging binary classification task. In addition, early detection and performance monitoring ensure that the sample remains wide and avoids overlap, making it suitable for medical applications for fracture detection.

V. FUTURE WORK

During this study, we have analyzed that CNN-based models only focus the local features, However, to capture `both features (Local and global features) from X-ray images, we have to use the ViT models or design the Hybrid approach with CNN to address-es the challenges related to the accuracy, feature breakup at multi-scale, overlapping associated with traditional approaches to deep learning.

ACKNOWLEDGMENT

This paper was authored by Ali Orangzeb, a PhD IT scholar. He wishes to express his sincere appreciation to his supervisor, Dr. Shahzad Memon, at the University of East, London, for his significant contributions and unwavering support throughout the completion of this research paper.

Disclosure of Interests. The authors have no conflicts of interest.

REFERENCES

[1] Sumi, Tahmina Akter, Nanziba Basnin, Md Shahadat Hossain, Karl Andersson, and Md Sazzad Hoassain. "Classifying humerus fracture using x-ray images." In *The Fourth Industrial Revolution and Beyond: Select Proceedings of IC4IR+*, pp. 527-538. Singapore: Springer Nature Singapore, 2023.

[2] Kim, D. H., and T. MacKinnon. "Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks." Clinical radiology 73, no. 5 (2018): 439-445.

[3] CANG ̈OZ, G. B., and G ̈UNEY, S. The Effects of the Traditional Data Augmentation Cutting-edge on Long Bone Fracture. Bilge International Journal of Science and Technology Research, .7(1), 2023, 63-69,2023.

[4] Yadav, D. P., and Rathor, S.. Bone fracture and classification using deep learning approach. In 2020 International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC) ,2020, February, (pp. 282-285). IEEE.

[5] Nguyen, H. P., Hoang, T. P., and Nguyen, H. H. (2021, October). A deep learning-based fracture in arm bone X-ray images. In 2021

international conference on multimedia analysis and pattern recognition (MAPR) (pp. 1-6). IEEE.

[6] Meena, Tanushree, and Sudipta Roy. "Bone fracture detection using deep supervised learning from radiological images: A paradigm shift." *Diagnostics* 12, no. 10 (2022): 2420. [7] Ali, R., Chuah, J. H., Talip, M. S. A., Mokhtar, N., and Shoaib, M. A. . Structural crack using deep convolutional neural networks. Automation in Construction, 133, 103989.

[8 Prijs, Jasper, Zhibin Liao, Minh-Son To, Johan Verjans, Paul C. Jutte, Vincent Stirler, Jakub Olczak et al. "Development and external validation of automated , classification, and localization of ankle fractures: inside the black box of a convolutional neural network (CNN)." European Journal of Trauma and Emergency Surgery 49, no. 2 (2023): 1057-1069.

[10] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In Proceedings of the IEEE.

[11] Rajpurkar, Pranav, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang et al. "Mura: Large dataset for abnormality detection in musculoskeletal radiographs." *arXiv preprint arXiv:1712.06957* (2017).

[12] Tahir, Ayesha, Ayesha Saadia, Khurram Khan, Ammara Gul, Ayman Qahmash, and Raja Naeem Akram. "Enhancing diagnosis: ensemble deep-learning model for fracture detection using X-ray images." *Clinical Radiology* 79, no. 11, e1394-e1402, 2024.

[13] Susmitha, N., and T. Anuradha. "A Review on Techniques and Approaches of Deep Learning in Bone Fracture." *Intelligent Systems Modeling and Simulation 3* 553 ,39. 2024.

[14] Pujitha, B., K. Raga Sravya, N. Krishnasai, and Ch Aparna. "Detection of Bone Fracture Using Deep Learning." In 2024 International Conference on Emerging Systems and Intelligent Computing (ESIC), pp. 703-708. IEEE, 2024.

[9] Cohen, Mathieu, Julien Puntonet, Julien Sanchez, Elliott Kierszbaum, Michel Crema, Philippe Soyer, and Elisabeth Dion. "Artificial intelligence vs. radiologist: accuracy of wrist fracture detection on radiographs." *European radiology* 33, no. 6 (2023): 3974-3983.