Bone Fracture Detection Using Convulutional Neural Network (CNN) K-Fold Validation Method

Laith Taani
Computer Engineering Depatrment
Princess Sumaya University for Technology
Amman, Jordan
laithtaani3@gmail.com

Yousef Jarbou
Computer Engineering Department
Princess Sumaya University for Technology
Amman, Jordan
yousefjarbou@outlook.com

Heba Abdel-Nabi
Computer Engineering Department
Princess Sumaya University for Technology
Amman, Jordan
h.abdelnabi@psut.edu.jo

***Abstract*—This work presents an automated deep learning-based approach for detection and classification of bone fractures from X-ray images. The study incorporates a combination of two publicly available databases that are vastly distinguished and representative of the sample. Further, extensive data preprocessing has been done, including augmentation methods, aimed at assuring the best performance and generalization of the model. A custom convolutional neural network (CNN) architecture developed and trained using properly optimized hyper-parameters was able to discriminate between fractured and unfractured bones. This proposed technique provides solutions with good performance by achieving an accuracy of 99.17% in fracture detection. Hence, this solution is well meant for the healthcare professional by providing a comprehensive automated diagnosis tool for possible reduction of manual interpretations as well as increasing diagnostic accuracy in the clinical environment.**

Keywords—Bone Fracture Detection; Deep Learning; Convolutional Neural Networks; X-ray Imaging; Medical Diagnostics.

I. Introduction

Bone fractures constitute a major health hazard across the world since they are responsible for a substantial percentage of emergency medical cases each year. The injuries can reduce the mobility of the patients and the quality of life, thus calling for timely and accurate diagnosis, which is prerequisite to effective treatment. Traditionally, fracture injuries are diagnosed with the help of X-ray imaging and their interpretation manually by radiologists, which is a laborious process and subject to errors of man [1].

The accuracy of these diagnoses can be significantly improved by deep learning. Its best-known application, CNNs, has transformed the field of medical image analysis by granting it the bulk of fully automated, accurate, and efficient diagnostics. CNNs have made it cumulative for applications to be widely applied in medical imaging involving the detection of tumors, classification of diabetic retinopathy, and diagnosis of lung diseases [2]. Recently, CNNs have been found to show prominent performance with respect to detecting bone fractures because of their ability to extract complex patterns and hierarchical features from X-ray images [3].

This paper discusses how CNNs can apply to detect and classify bone fractures. The proposed model features highly advanced CNN architectures integrated with data augmentation and transfer learning techniques to tackle the common issues of data imbalance and variation in fracture type. This work with its automated diagnostics comes as a part of helping healthcare practitioners reduce possible errors in diagnosing and improve patient outcomes [4] [5].

II. Related Work

In the field of image classification, especially for medical purposes, deep CNNs have been extensively used due to their high performance and ability for automatic feature extraction. A wide variety of approaches and architectural innovations have been explored.

The work in [6] introduced a novel methodology for optimizing CNN architectures in the context of X-ray image classification. The authors leveraged transfer learning combined with residual CNNs to enhance feature extraction and classification accuracy. This approach demonstrated significant improvement in identifying abnormalities in chest X-rays, with an overall accuracy of 92.4%, underscoring the potential of residual networks for medical imaging tasks more, the research in [7] employed EfficientNet architectures for detecting dental diseases in X-ray images. The study optimized feature extraction using advanced CNN layers and inverted residual blocks, achieving high classification accuracy of 94.6% across dental disease categories. This demonstrates the adaptability of lightweight CNN architectures for medical imaging applications.

The work in [8] proposed a hybrid CNN-XGBoost approach for early detection of lung diseases from chest X-ray images. By combining CNN feature extraction with XGBoost classification, the study achieved notable accuracy improvements while maintaining computational efficiency. The methodology emphasized the use of explainable AI to interpret results, making it suitable for clinical applications.

III. EXPERIMENTAL SETUP

The objective of this paper is to produce a bone fracture prediction system using an X-Ray dataset. This dataset represents real data which serves the purpose of this paper and allows the prediction system to generalize to any new data.

A. Data Attribute Description

The quality and training of data are central to the performance of deep learning (DL) models. As a result, we gave two publicly available datasets that focus on bone fractures

Fracture Multi-Region X-ray Data [9] and Bone Fracture Detection Using X-rays [10]. These datasets comprise labeled images with the categories as "fractured" and "unfractured" and thus constituted the training and validation basis. Merging these datasets has broadened the diversity and size of the dataset, thus increasing the generalization capacity of the model.

During the merging process, a critical issue was identified overlapping images with identical names appeared in both training and validation subsets. To address this, the training and validation files were combined, and duplicate images were removed, ensuring that only unique instances were retained. Subsequently, the dataset was re-split into distinct training and validation subsets with a balanced distribution.

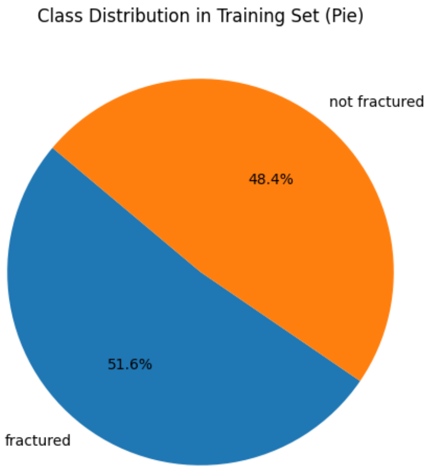

Figure 1 - Data Distribution

B. Data Preprocessing

To further improve the model's generalizability and robustness, we implemented a data augmentation strategy. This involved applying random transformations to simulate real-world variations commonly observed in medical imaging, such as changes in orientation, scale, and position. Additionally, the images were normalized to standardize pixel intensity values, enhancing consistency across the dataset. These preprocessing steps were critical in maximizing the model's capacity to perform reliably in diverse clinical scenarios.

IV. MACHINE LEARNING ALGORITHM

After analyzing the data, the dataset was divided into training and testing in an 80:20 ratio. The split of the data ensures enough for training, with a significant portion reserved for testing on unseen data. Further, to increase the generalizing capability of the model, 5-fold cross-validation of the training data was employed. The approach typically involves splitting given data into subsets of five-folds where four subsets are iteratively used to train and validated against the leftover. This systemic validation ensures better robustness over the performance estimations and restricts overfitting as well [11].

DL is a subset of machine learning that embraces the computational model comprising multiple layers to represent data in an efficiently abstracted way. These methodologies have been very successful for at least image classification, object detection, and natural language processing applications over the last years [12]. DL explicitly exploits intricate structure in large datasets by using the backpropagation algorithm, modifying internal parameters to allow for representation and, hence, learning across multiple layers [13].

Deep convolutional networks have been specifically revolutionary in the handling of images, videos, speech, and audio, while recurrent networks perform exceptionally well on sequential data such as text and time-series analysis. Amongst the various DL architectures, convolutional neural networks (CNNs) have found their application in several tasks involving medical diagnostics by making use of convolutional layers that extract the spatial hierarchies of features from the images [14]. Figure 1 provides a visualization of the CNN model architecture used in this study.

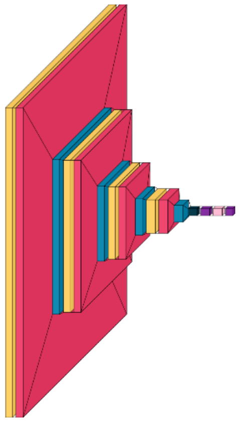

Figure 2 - CNN Model Layers

The general architecture of CNNs consists of convolutional, pooling, and fully connected layers. The convolutional layer extracts the local features by applying the kernel to parts of an image; this operation is repeated across the entire image to capture spatial patterns. The pooling layers, which may be in the form of max-pooling or average-pooling, reduce the spatial dimensions of feature maps while retaining important information and hence enhance translation and rotation invariance. These layers are often interspersed and terminate in fully connected layers that merge features for coherent representation at classification or detection tasks *[15]*. Table 1 shows the specific model architecture used in this study.

Table 1 - Summary of CNN Architecture Used

Layer	Type	Output Shape
Input Layer	-	(224, 224, 3)
Conv2D_1	Convolutional (32 filters, 3x3 kernel)	(222, 222, 32)
BatchNormalization_1	Batch Normalization	(222, 222, 32)

MaxPooling_1	Max Pooling (2x2)	(111, 111, 32)
Conv2D_2	Convolutional (64 filters, 3x3 kernel)	(109, 109, 64)
BatchNormalization_2	Batch Normalization	(109, 109, 64)
MaxPooling_2	Max Pooling (2x2)	(54, 54, 64)
Conv2D_3	Convolutional (128 filters, 3x3 kernel)	(52, 52, 128)
BatchNormalization_3	Batch Normalization	(52, 52, 128)
MaxPooling_3	Max Pooling (2x2)	(26, 26, 128)
Conv2D_4	Convolutional (256 filters, 3x3 kernel)	(24, 24, 256)
BatchNormalization_4	Batch Normalization	(24, 24, 256)
MaxPooling_4	Max Pooling (2x2)	(12, 12, 256)
GlobalAveragePooling	Global Average Pooling	(256)
Dense_1	Fully Connected	(256)
Dropout_1	Dropout (rate=0.5)	(256)
Dense_2 (Output)	Fully Connected	(1)

V. Results and Analysis

Once the architecture of the model is decided, it would be necessary to tune the hyperparameters such as learning rate, batch size, and number of layers. The approach used was RandomizedSearchCV, whereby a different combination of hyperparameters is iterated upon in order to get the near best settings for the model. During training, input data, X-ray images, passes through the network, updating the weights of every neuron with a back-propagation algorithm to minimize error between predicted outputs and true labels. This is repeated iteratively until the model learns adequately to generalize from the training data [16].

To gauge the performance of the trained model, certain metrics about its performance will be necessary in order to ensure its reliability. Some common metrics used for the evaluation of classification performance include accuracy, precision, recall, and an F1 score. Generally, a confusion matrix is used to explain the result of binary classification: true positives (TP - correctly predicted positive values), false positives (FP - wrongly predicted positive values), true negatives (TN - correctly predicted negative values), and false negatives (FN - wrongly predicted negative values). Figure 3 shows the confusion matrix obtained from this model [17].

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (1)$$

The proportion of correctly predicted cases to the total number of predictions.

$$Precision = \frac{TP}{TP+FN} \qquad (2)$$

The proportion of correctly identified fractures among all cases predicted as fractures.

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

Measures the proportion of actual fractures correctly identified.

$$F1\,Score = \frac{2}{\frac{1}{P}+\frac{1}{R}} = \frac{2 \times P \times R}{P+R} = \frac{TP}{TP+\frac{FN+FP}{2}} \qquad (4)$$

Balances precision and recall, providing a harmonic mean of the two.

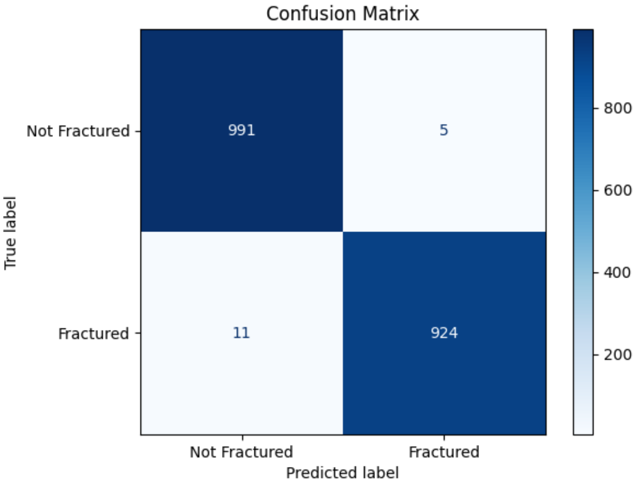

Figure 3 - Confusion Matrix

Larger values for accuracy, precision, recall, and F1 score indicate better model performance. After training and testing, the trained model can now be used in the processing of new data; thus, enabling fast and accurate detection of fractures in bones. It is a good contribution to the clinical diagnosis for doctors and hospitals to reduce misdiagnosis or diagnosis missed due to oversight [18].

The ROC curve was plotted to further analyze the performance of the developed models for the detection of bone fractures, represented in Figure 4. In this ROC curve, the diagnostic capability of the classifier is represented by plotting the true positive rate or sensitivity against the false positive rate or 1-specificity at different threshold levels. The area under the ROC curve was also calculated and is represented in Table 2. The AUC provides a single scalar to quantify model performance; the closer to one, the better the diagnostic ability of the model, which in turn means stronger model performance. This allows an extensive comparison of models involved in the project.

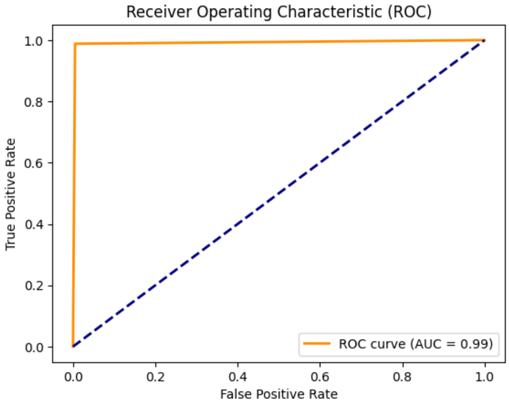

Figure 4 - ROC Curve

Table 2 - Results Comparison

Study	Methodology	Accuracy (%)
Study A [6]	Residual CNN	92.4
Study B [7]	EfficientNet	94.6
Study C [8]	CNN-XGBoost	95.2
Proposed Method	Custom CNN	99.17

The proposed method outperformed existing models in accuracy and generalizability, demonstrating its potential for clinical use. Data preprocessing and augmentation significantly contributed to model accurateness.

VI. Conclusion

This study developed a highly accurate CNN-based approach for automated bone fracture detection. By integrating advanced preprocessing techniques and optimizing the architecture, the model achieved state-of-the-art performance. Future work should focus on expanding the dataset and exploring lightweight models for deployment in resource-constrained settings.

References

[1] X. Chen, Y. Chen, H. Liu, G. Goldmacher, C. Roberts, D. Maria and W. Ou, "PIN92 PEDIATRIC BACTERIAL PNEUMONIA CLASSIFICATION THROUGH CHEST X-RAYS USING TRANSFER LEARNING," *Value in Health,* vol. 22, pp. S209-S210, 2019.

[2] P. Gupta and S. Gupta, "Deep Learning in Medical Image Classification and Object Detection: a Survey," *International Journal of Image Processing and Pattern Recognition,* 2022.

[3] J. Olczak, N. Fahlberg, A. Maki, A. S. Razavian, A. Jilert, A. Stark, O. Sköldenberg and M. Gordon, "Artificial intelligence for analyzing orthopedic trauma radiographs," *Acta Orthopaedica,* vol. 88, no. 6, pp. 581-586, 2017.

[4] S. Parvin and A. Rahman, "A real-time human bone fracture detection and classification from multi-modal images using deep learning technique," *Applied Intelligence,* vol. 54, no. 19, pp. 9269-9285, 2024.

[5] A. M. Raisuddin, E. Vaattovaara, M. Nevalainen, M. Nikki, E. Järvenpää, K. Makkonen, P. Pinola, T. Palsio, A. Niemensivu, O. Tervonen and A. Tiulpin, "Critical evaluation of deep neural networks for wrist fracture detection," *Scientific Reports,* vol. 11, no. 1, p. 6006, 2021.

[6] S. Ahlawat, A. Choudhary, A. Nayyar, S. Singh and B. Yoon, "Improved Handwritten Digit Recognition Using Convolutional Neural Networks (CNN)," *Sensors,* vol. 20, no. 12, p. 3344, 2020.

[7] S. A. Ahmad, M. N. Taib, N. E. A. Khalid and H. Taib, "An Analysis of Image Enhancement Techniques for Dental X-ray Image Interpretation," *International Journal of Machine Learning and Computing,* pp. 292-297, 2012.

[8] M. Ž. N. B. M. A. B. N. G. K. M. Marjanovic and N. S. , "Hybrid CNN and XGBoost Model Tuned by Modified Arithmetic Optimization Algorithm for COVID-19 Early Diagnostics from X-ray Images," *Electronics,* vol. 11, no. 22, pp. 3798-3798, 2022.

[9] M. Rodrigo, "Bone Fracture Multi-Region X-ray Data," Kaggle.com, 2024. [Online]. Available: https://www.kaggle.com/datasets/bmadushanirodrigo/fracture-multi-region-x-ray-data. [Accessed 2024].

[10] Anon, "bone fracture detection using x-rays," Kaggle.com, 2024. [Online]. Available: https://www.kaggle.com/datasets/vuppalaadithyasairam/bone-fracture-detection-using-xrays. [Accessed 2024].

[11] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto and S. Ridella, "The 'K' in K-fold Cross Validation," 2012. [Online]. Available: https://www.esann.org/sites/default/files/proceedings/legacy/es2012-62.pdf.

[12] Y. LeCun, Y. Bengio and G. Hinton, "Deep Learning," *Nature,* vol. 521, no. 7553, pp. 436-444, 2015.

[13] T. Shanthi and R. Sabeenian, "Modified Alexnet architecture for classification of diabetic retinopathy images," *Computers & Electrical Engineering,* vol. 76, pp. 56-64, 2019.

[14] B. Kayalibay, G. Jensen and P. van der Smagt, "CNN-based Segmentation of Medical Imaging Data," 2017.

[15] G. Huang, Z. Liu, L. Van Der Maaten and K. Weinberger, "Densely Connected Convolutional Networks," 2017. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/papers/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.pdf.

[16] D. Bhatt, C. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya, K. Modi and H. Ghayvat, "CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope," *Electronics,* vol. 10, no. 20, p. 2470, 2021.

[17] N. Alamsyah, B. Budiman, T. P. Yoga and R. Y. R. Alamsyah, "XGBOOST HYPERPARAMETER OPTIMIZATION USING RANDOMIZEDSEARCHCV FOR ACCURATE FOREST FIRE DROUGHT CONDITION PREDICTION," *Jurnal Pilar Nusa Mandiri,* vol. 20, no. 2, pp. 103-110, 2024.

[18] C. Goutte and E. Gaussier, "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation," *Lecture Notes in Computer Science,* pp. 345-359, 2005.

[19] J. Lu, L. Tan and H. Jiang, "Review on Convolutional Neural Network (CNN) Applied to Plant Leaf Disease Classification," *Agriculture,* vol. 11, no. 8, p. 707, 2021.

[20] X. Li, S. Wei, S. Niu, X. Ma, H. Li, M. Jing and Y. Zhao, "Network pharmacology prediction and molecular docking-based strategy to explore the potential mechanism of Huanglian Jiedu Decoction against sepsis," *Computers in Biology and Medicine,* vol. 144, p. 105389, 2022.