JIKO (Jurnal Informatika dan Komputer) Accredited KEMENDIKTI SAINSTEK, No.0173/C3/DT.05.00/2025
Vol. 8, No. 1, April 2025, pp. 21–30 p-ISSN: 2614-8897
DOI: 10.33387/jiko.v8i1.9201 e-ISSN: 2656-1948

CLASSIFICATION OF BONE FRACTURES IN THE WRIST AND HAND USING DENSENET AND XCEPTION

Michelle Swastika Bianglala Nusantara [1], Daniel Martomanggolo Wonohadidjojo [2*]

[1,2]School of Information Technology, Universitas Ciputra Surabaya
Email: [1]mswastika@student.ciputra.ac.id, [*2]daniel.m.w@ciputra.ac.id

(Received: 04 December 2024, Revised: 20 December 2024, Accepted: 29 December 2024)

Abstract

Accurate and timely diagnosis of wrist and hand bone fractures is crucial to ensure effective treatment and avoid complications such as avascular necrosis or non-union. However, manual interpretation of radiographic images is prone to errors due to the fractures' subtle and complex nature. This study applies Convolutional Neural Networks (CNN) to address this challenge, using DenseNet and Xception architectures for automated fracture classification. The models are optimized to enhance diagnostic accuracy and training efficiency by leveraging transfer learning. The research utilizes two publicly available musculoskeletal radiography datasets and employs deep learning techniques within the Keras framework. DenseNet is applied to wrist images due to its dense connectivity, which retains information from earlier layers, while Xception is employed for hand bone images to detect intricate patterns through depthwise separable convolutions. The DenseNet model achieved a test accuracy of 97.5% for wrist fracture classification, and the Xception model achieved 92% accuracy for hand bone fracture classification. These results demonstrate the potential of tailored CNN architectures combined with transfer learning to significantly improve fracture detection, thereby supporting medical professionals in making faster and more accurate clinical decisions.

Keywords: *bone fracture classification, CLAHE, contrast stretching, densenet, unsharp masking, xception*

**Corresponding Author: Daniel Martomanggolo Wonohadidjojo*

1. INTRODUCTION

Hand and wrist bone fractures are a major worldwide health concern because they significantly increase the risk of disability and reduced mobility. The World Health Organization (WHO) estimates that 1.71 billion individuals worldwide suffer from musculoskeletal disorders, of which bone fractures are a leading cause [1]. As complex and vital components of the human body, the hand and wrist are crucial for daily tasks, including eating, writing, and clothing. A person's quality of life may be significantly impacted by injuries to these areas [2]. Consequently, timely and precise fracture detection and classification are essential for efficient treatment and for reducing long-term issues like malunion (inappropriate bone healing) or irreversible functional damage.

Even while doctors are crucial in identifying bone fractures, there is still a significant margin for diagnostic error. According to research by Zhang et al. [3], radiology and orthopedic professionals diagnose fractures with an error margin of 7%–8% and accuracy rates of roughly 92%–93%. These inaccuracies are caused by several factors, including physician weariness, fracture presentation diversity, and the intricacy of X-ray pictures. Patients who receive a misdiagnosis run the risk of experiencing chronic discomfort, decreased functionality, or even permanent impairment as a result of delayed or improper treatment. These difficulties show how urgently sophisticated diagnostic instruments are needed to improve precision and aid in therapeutic judgment.

Recent advancements in Artificial Intelligence (AI), particularly in deep learning using Convolutional Neural Networks (CNNs), have shown promise in improving medical imaging analysis, including fracture classification. CNNs have demonstrated exceptional capabilities in extracting critical features from medical images, enabling accurate classification and anomaly detection [4]. Studies such as those by Solikhun et al. [5] report CNN models achieving

fracture classification accuracies as high as 99%, showcasing the potential of AI to surpass traditional diagnostic methods. Additionally, transfer learning has emerged as an effective strategy for utilizing pre-trained CNN models to improve performance on certain tasks, even with relatively small datasets [6].

Traditional machine learning (ML) techniques were not employed in this study due to the limited size of the datasets. ML models generally require large datasets to achieve reliable performance. With small datasets, these models are prone to overfitting and struggle to generalize effectively. Instead, deep learning approaches using CNNs with transfer learning were chosen, as transfer learning leverages pre-trained models to address data limitations while maintaining high accuracy and generalization.

Many studies have been conducted on the use of CNNs and transfer learning for the classification of bone fractures. The efficiency of transfer learning in musculoskeletal X-ray classification, for example, was investigated by Kandel et al. [7]. They showed that the DenseNet and Xception architectures obtained 81% and 75% accuracy for classifying wrist and hand fractures, respectively. These accuracy levels, however, are not clinically applicable, underscoring the need for additional improvement. Gupta and Sharma [8] addressed challenges such as class imbalance by suggesting techniques like oversampling or undersampling to enhance model reliability and reduce bias. Meanwhile, Meena and Roy[9] emphasized the potential of CNN-based models, including XceptionNet, for fracture detection, though their study did not specifically focus on individual bone regions.

Based on these studies, this research employs DenseNet and Xception architectures to improve fracture classification performance. DenseNet's densely connected layers efficiently retain and reuse information from preceding layers, making it particularly effective for wrist fracture classification [7]. Similarly, Xception's use of depthwise separable convolutions enhances computational efficiency and pattern recognition, making it well-suited for hand fracture classification [7].

This study aims to implement DenseNet and Xception architectures for classifying hand and wrist fractures, leveraging transfer learning to enhance diagnostic accuracy. By addressing limitations in previous research and optimizing dataset preparation, this work seeks to advance the development of accurate and consistent decision-support systems in radiology. These improvements hold the potential to enhance diagnostic precision, reduce error rates, and ultimately improve patient outcomes in clinical settings.

2. RESEARCH METHOD

This study employs a structured methodology comprising several key stages: data collection, data augmentation and preprocessing, model training, and model evaluation. Each stage is essential to ensure the model achieves accurate predictions aligned with the research objectives. The research methodology workflow used in this study can be represented by Figure 1.

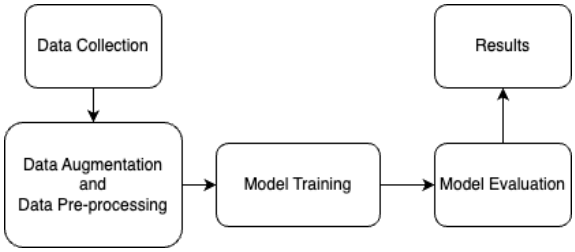

Figure 1. Workflow of the Bone Fracture Classification Model Development

2.1 Data Collection

This study utilizes two public datasets: MURA by Rajpurkar et al. [10] and the Wrist Fracture - X-rays dataset by Malik et al. [11]. The 40,561 musculoskeletal radiography pictures from 14,863 investigations that make up the MURA dataset have been classified as either normal or broken by radiology specialists [12]. Six certified Stanford radiologists provided additional labels for a test set of 207 musculoskeletal studies to evaluate model performance and compare it to radiologist accuracy. This dataset includes images of various parts of the hand, including the wrist and hand bones, making it highly relevant to this research.

Meanwhile, the Wrist Fracture - X-rays dataset by Malik et al. focuses specifically on wrist X-ray images, categorizing them as either "Fracture" or "Normal," with 111 and 82 images, respectively. These images were gathered from the Al-huda Digital X-ray Laboratory in Multan, Pakistan [13], with the primary objective of detecting wrist fractures.

The model for the classification of wrist fractures is trained using the Malik et al. dataset, while the model for the classification of hand bone fractures is trained using the MURA dataset. Figure 2 provides sample images from the MURA dataset, illustrating hand and wrist classes, and Figure 3 showcases examples from the Malik et al. dataset for wrist images.

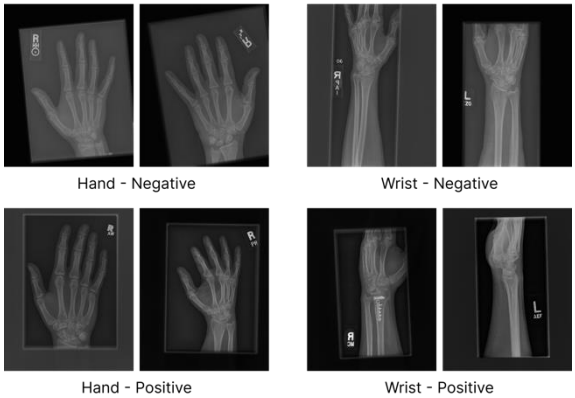

Figure 2. Sample Images from Each Class of Hand and Wrist in the MURA Dataset

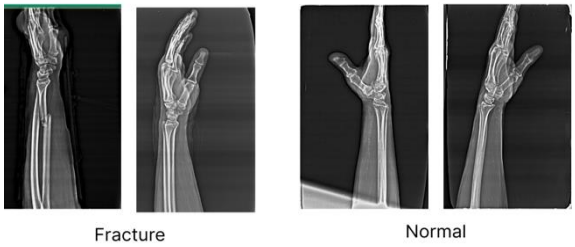

Figure 3. Sample Images from Each Class in the Malik Et Al. Dataset

2.2 Data Augmentation and Data Preprocessing

To prepare the dataset for training, several essential processes are included in the initial data processing. First, the dataset images are sorted. This step filters out inappropriate images, such as those showing objects unrelated to this study, like rings on fingers in hand X-rays or irrelevant background elements. Such images are removed to ensure they do not interfere with model training.

After sorting, the next step is cropping to focus on the relevant areas of the images, specifically the bone regions being classified, while excluding unnecessary parts. Additionally, watermarks on some images are removed to prevent visual interference during classification.

Once the suitable images are prepared, data augmentation is used to expand the amount and variety of the dataset. Various augmentation techniques, including 90° rotations, horizontal and vertical flips, are included in the data augmentation, adding variations in angles and orientations to enrich the dataset. Figure 4 displays examples of each augmentation technique alongside the original hand images. The same techniques are applied to wrist images. This augmentation process helps the model learn from a broader range of variations, improving its generalization ability.

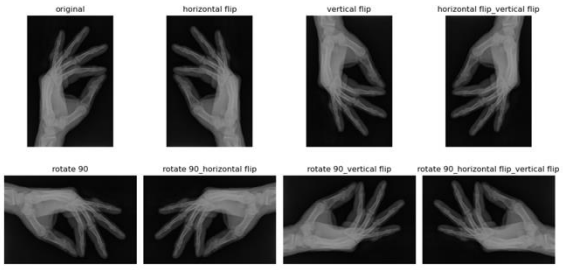

Figure 4. Sample Images of Hand Before and After Each Augmentation

After augmentation, class balancing is performed to ensure an equal number of images for both classes in the dataset. This step prevents the model from being biased toward one class. If one class has fewer images, additional augmentations are applied to balance the numbers. The final wrist dataset contains 200 images each for the negative (no fracture) and positive (fracture) classes, totaling 400 images. For the hand dataset, each class contains 450 images, resulting in a total of 900 images.

The next step is data preprocessing, which differs between the two datasets. The wrist dataset underwent initial preprocessing by the dataset provider, so only CLAHE (Contrast Limited Adaptive Histogram Equalization) is applied in this study to enhance local contrast, making bone details clearer under varying lighting conditions. In contrast, the hand dataset requires more comprehensive preprocessing, including CLAHE, contrast stretching to expand intensity ranges and reveal hidden details, and unsharp masking to improve image sharpness, making bone structures more distinct and easier for the model to identify.

CLAHE modifies Adaptive Histogram Equalization (AHE) by limiting contrast enhancement to prevent artifacts like "halo effects" [14]. In AHE, excessive contrast in narrow histogram areas can amplify noise and cause halos [15]. CLAHE addresses this by setting a "clip limit" for contrast enhancement. The clip limit of a histogram is provided in Equation (1) where α is the clip factor, N is the grayscale value (256), and M is the area size [16].

$$\beta = \frac{M}{N}\left(1 + \frac{\alpha}{100}(s-1)\right) \times 100\% \quad (1)$$

Contrast stretching enhances image contrast by expanding the pixel intensity value range to a desired range, such as the entire pixel range that the image type permits. This technique adjusts pixel intensity distributions to highlight otherwise hard-to-see details, making them more visible and clear [17]. The formula for contrast stretching is presented in Equation (2) [18]:

$$g(x,y) = \frac{f(x,y) - min}{max - min} \times 255 \quad (2)$$

Where $g(x,y)$ is the matrix of the resulting image which range from 0 to 255, $f(x,y)$ is the original intensity value matrix, min and max are the minimum and maximum pixel intensity values of the original image [18].

Unsharp masking enhances image sharpness by subtracting a blurred (lowpass filtered) version of the image from the original. This technique highlights subtle details, making the final image appear sharper and clearer. The process mathematically combines the original image with the blurred version to emphasize clarity and detail. The unsharp masking formula is outlined in Equation (3) [19], where α and β are positive constants with $\alpha \geq \beta$.

$$\hat{f} = \alpha f - \beta f_{l_p} \quad (3)$$

Figures 5 and 6 showcase samples of images before and after preprocessing for wrist and hand objects.

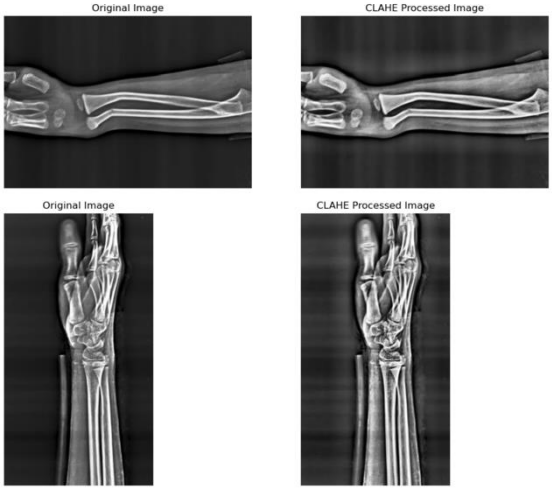

Figure 5. Sample Images of Wrist Before and After Applying CLAHE

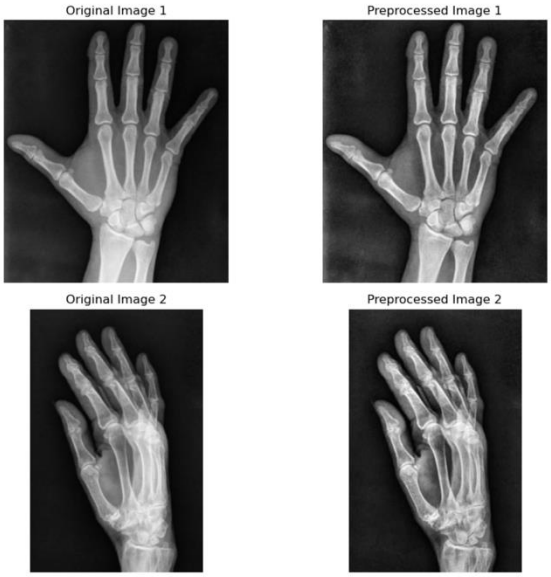

Figure 6. Sample Images of Hand Before and After Applying CLAHE, Unsharp Masking, and Contrast Stretching

After the preprocessing, the datasets are divided into three parts. For the wrist dataset, 70% is allocated for training, 10% goes to validation, and 20% goes to testing. For the hand dataset, 80% goes to training, 10% goes to validation, and 10% goes to testing.

2.3 Model Training

This study employs deep learning as the primary method for medical image analysis, specifically for detecting fractures in hand and wrist bones. The model architecture used in this research is CNN, with two types of architectures: DenseNet and Xception. DenseNet was selected for its superior performance in processing wrist images, while Xception was utilized for hand bone images. Both architectures are well-regarded for their accuracy in feature detection from images.

The Keras framework, which is based on TensorFlow, was utilized to train the deep learning models. Model training was performed on a Graphics Processing Unit (GPU) to accelerate computations.

The hyperparameters for the wrist images are listed in Table 1. Training was conducted with a batch size of 32, L2 regularization set to 0.01, the Nadam optimizer with 100 epochs, and a learning rate of 0.0001. To avoid overfitting, early stopping was used, in which training stops if no increase in validation metrics is observed over several epochs.

Table 1. Hyperparameters for Wrist Model

Hyperparameter	Value
Learning Rate	0.0001
Epoch	100
Batch Size	32
L2 Regularization	0.01
Early Stopping	10 epoch

The hyperparameters for the hand images are provided in Table 2. The training was conducted with a batch size of 32 and included fine-tuning the Xception model to enhance its ability to learn task-specific features. Regularization parameters included L2 regularization set to 0.03 and dropout set to 0.2 to reduce overfitting. The Nadam optimizer with a learning rate of 0.001 was applied, and training lasted for 100 epochs. Additionally, the ReduceLROnPlateau technique was used to adjust the learning rate when no improvement in validation metrics was detected.

Table 2. Hyperparameters for Hand Model

Hyperparameter	Value
Learning Rate	0.001
Epoch	100
Batch Size	32
L2 Regularization	0.03
Dropout	0.2
ReduceLROnPlateau	5 epochs

To make sure the model could generalize to new data, this study used a k-fold cross-validation procedure (k=5) on the training dataset. Cross-validation offers a more accurate evaluation of model performance and lowers the danger of overfitting. The dataset is divided into five subsets (folds), and training and validation are carried out five times. Each fold is used as validation data once, and the rest are utilized for training. This procedure guarantees that every piece of data is used as validation data at least once, resulting in a more representative assessment of performance.

Each iteration involves splitting the data into training and validation sets, training a new model, and evaluating its performance using metrics like recall, accuracy, precision, and F1 score. Early stopping was implemented to prevent overfitting by tracking the validation accuracy and stopping the training after ten consecutive epochs if no progress was seen.

The phases of the transfer learning process used in this study are depicted in Figure 7.

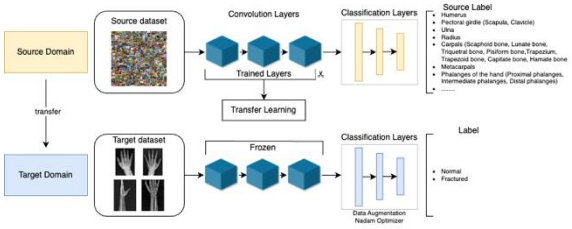

Figure 7. Phases of the Transfer Learning Process for Bone Fracture Classification

2.4 Model Evaluation

A number of metrics, including the confusion matrix, F1 score, recall, accuracy, and precision, were used to evaluate the model's performance. The model's accuracy shows how effectively it can distinguish between fractured and non-fractured images. The precision metric measures the proportion of predicted positive cases that come to pass. Recall gauges how well it can identify each genuine positive sample. F1 score is obtained by calculating the harmonic mean of precision and recall [20]. Equations (4), (5), (6), and (7) represent the formulas for accuracy, precision, recall, and F1 score, respectively [21].

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \times 100\% \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (6)$$

$$F1\ Score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \times 100\% \quad (7)$$

The confusion matrix offers comprehensive metrics regarding the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) [22]. This analysis provides more in-depth understanding of the model's performance trends and identifies potential classification problem areas.

3. RESULT AND DISCUSSION

3.1 Training Results for the Wrist Model

This study involved three experiments focused on wrist images, utilizing different datasets. The first experiment used the MURA dataset from the Large Dataset for Abnormality Detection in Musculoskeletal Radiographs by Rajpurkar et al. [10]. The training and validation accuracy graphs displayed in Figure 8 and the loss graphs in Figure 9 revealed unstable improvements in accuracy and non-converging reductions in loss. These patterns indicated signs of overfitting during training, making it challenging to achieve optimal performance. The final accuracy achieved with this dataset was only 91.7%.

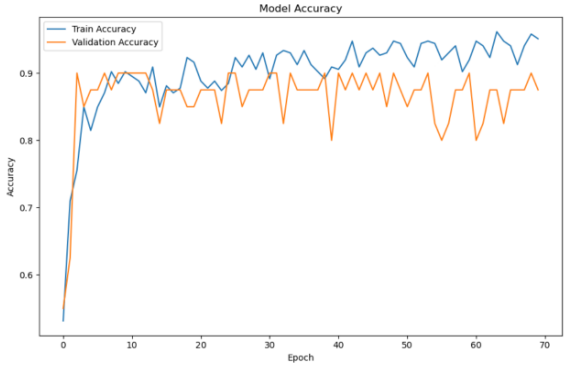

Figure 8. Training and Validation Accuracy Graph of the Wrist Model Using the MURA Dataset

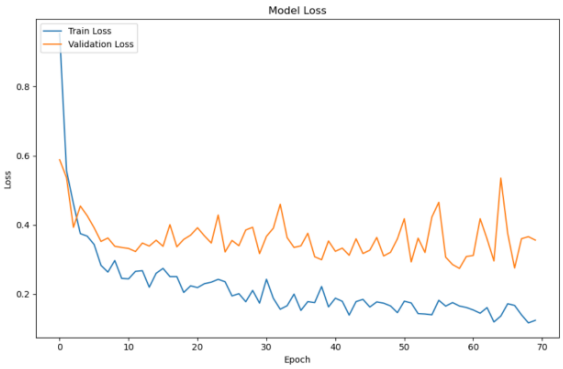

Figure 9. Training and Validation Loss Graph of the Wrist Model Using the MURA Dataset

In the second experiment, the dataset was expanded by combining the Wrist Fracture - X-rays dataset by Malik et al. [11] with the MURA dataset. However, this approach also yielded unsatisfactory results. The combined dataset's performance, as depicted in the training and validation accuracy graphs in Figure 10 and the loss graphs in Figure 11, showed continued difficulties in achieving optimal convergence. The model's test accuracy was only 89.9%, further highlighting the challenges of combining these datasets.

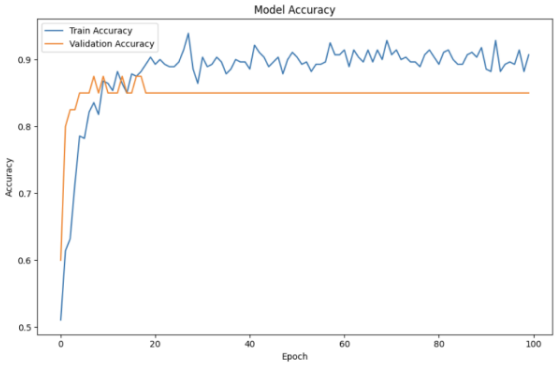

Figure 10. Training and Validation Accuracy Graph of the Wrist Model Using the Combined Dataset

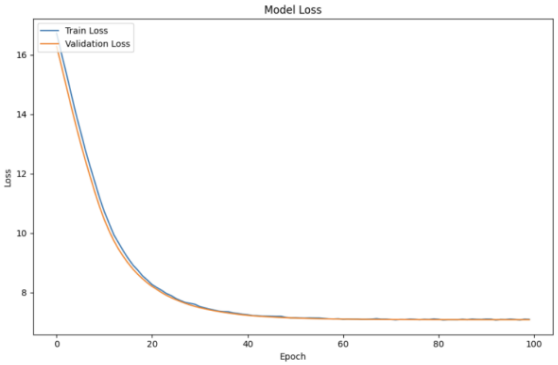

Figure 11. Training and Validation Loss Graph of the Wrist Model Using the Combined Dataset

In the final approach, training was conducted exclusively using the Wrist Fracture - X-rays dataset by Malik et al. [11]. This method produced significantly better results. The model obtained a 97.5% accuracy rate using the DenseNet architecture. The stability of the training process was evident in the training and validation accuracy graphs in Figure 12 and the loss graphs in Figure 13, which demonstrated consistent improvements in accuracy and converging loss reductions. These results confirmed that the model avoided issues of overfitting or underfitting.

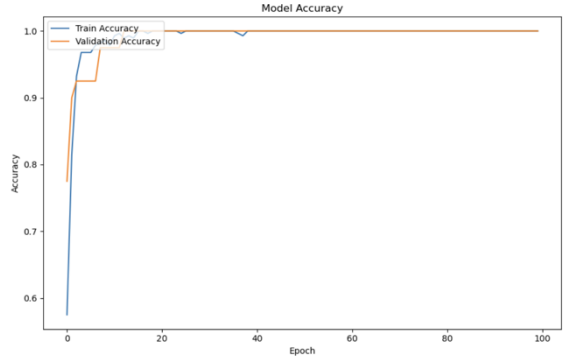

Figure 12. Training and Validation Accuracy Graph of the Wrist Model Using the Dataset by Malik et al.

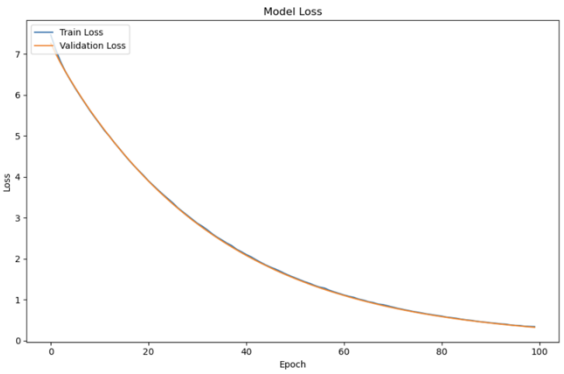

Figure 13. Training and Validation Loss Graph of the Wrist Model Using the Dataset by Malik et al.

A confusion matrix, shown in Figure 14, was implemented to further evaluate the DenseNet model's performance. The model categorized both positive and negative images with very low error rates, as evidenced by the confusion matrix's 40 True Negatives, 0 False Negatives, 2 False Positives, and 38 True Positives. There were no misclassifications of positive images as negative. The model's great performance was further demonstrated by additional metrics, such as accuracy, recall, and F1 score, which showed 100% precision, 95% recall, and 96.44% F1 score.

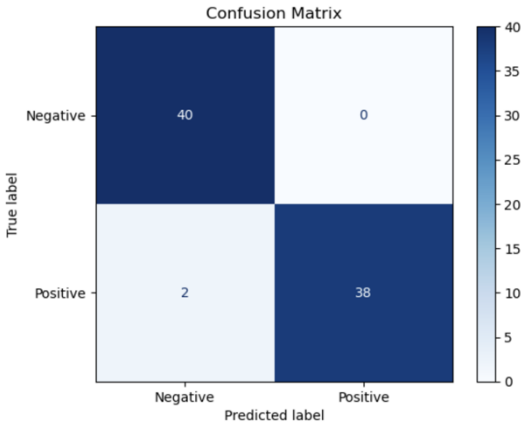

Figure 14. Confusion Matrix for Wrist Model Using the Dataset by Malik et al.

The next step in the evaluation process involved applying cross-validation to assess the model's generalization capabilities. Cross-validation with k=5 was applied, and the average performance across all folds is summarized in Tables 3 and 4. The model obtained an F1 score of 0.93, recall of 0.99, precision of 0.89, and average accuracy of 97.5%. These consistent outcomes indicate that the model performed well across various subsets of the dataset.

Table 3. Comparison of Accuracy Metrics Results Per Fold of Wrist Model's Cross-Validation

Fold	Accuracy	Precision	Recall	F1 Score
1	98.21%	100%	97%	98%
2	92.86%	92%	97%	94%
3	98.21%	89%	100%	94%
4	98.21%	75%	100%	86%
5	100%	90%	100%	95%

Table 4. Comparison of Confusion Matrix Results Per Fold of Wrist Model's Cross-Validation

Fold	True Negative	False Negative	False Positive	True Positive
1	26	0	1	29
2	19	1	3	33
3	28	0	3	25
4	24	0	8	24
5	26	0	3	27

The model in Fold 1 achieved a precision of 1.00, recall of 0.97, and F1 score of 0.98 with only one minor error in positive predictions. Fold 2 showed an F1 score of 0.94, precision of 0.92, and recall of 0.97, with a few errors in the negative categories. With an F1 score of 0.94, recall of 1.00, and precision of 0.89, Fold 3 showed no missed positives but a few minor problems with negative predictions. Recall was 1.00, precision was 0.75, and F1 score was 0.86 because Fold 4 had several mistakes in negative classifications.

With slight misclassifications in negative predictions, Fold 5 showed strong results with a precision of 0.90, recall of 1.00, and F1 score of 0.95.

To further evaluate the overall performance of the model across all folds, the overall confusion matrix results, as shown in Figure 15, indicate a well-balanced performance. The model correctly classified 25 negative cases (True Negatives) and 28 positive cases (True Positives). Misclassifications included 4 positive cases predicted as negative (False Negatives) and 0 negative cases predicted as positive (False Positives). This outcome further validates the model's robustness and its ability to generalize effectively across different subsets of the data.

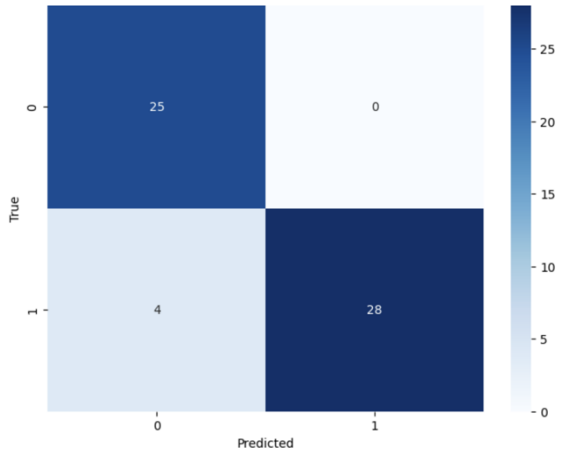

Figure 15. Overall Confusion Matrix Results Across All Folds of Wrist Model Cross-validation

3.2 Training Results for the Hand Model

The classification model for hand fractures, utilizing the Xception architecture, demonstrated strong performance as reflected in its accuracy, loss graphs, and evaluation metrics. The model successfully learns from the training data, as seen by the training and validation accuracy graph in Figure 16, which displays a steady increase in training accuracy until stabilizing after about 20 epochs. Similarly, validation accuracy follows a comparable trend, closely aligning with training accuracy toward the end of the training process. This consistency suggests the model avoids overfitting. Meanwhile, the loss graph in Figure 17 reveals a significant initial drop in both training and validation loss, which levels out as epochs progress. This decline reflects the model's ability to minimize error during training. The model's performance stayed consistent by the end of the training period, despite slight fluctuations in validation loss at the beginning.

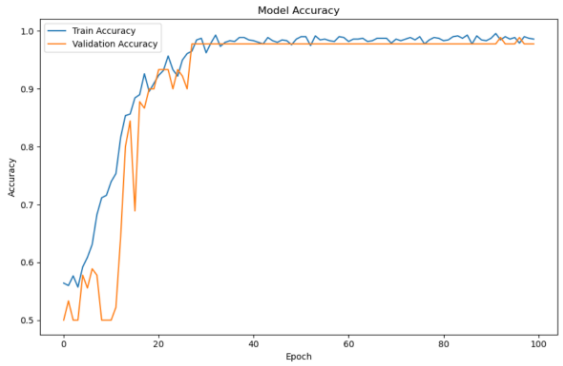

Figure 16. Training and Validation Accuracy Graph of the Hand Model

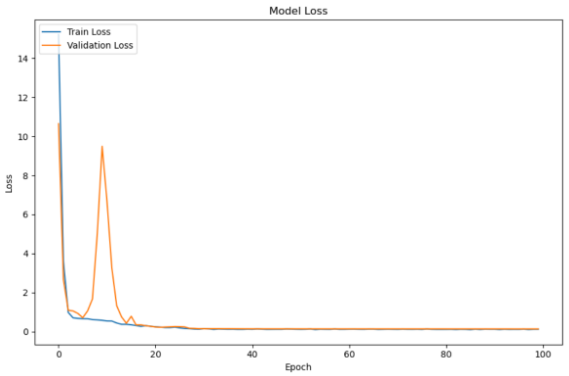

Figure 17. Training and Validation Loss Graph of the Hand Model

The confusion matrix in Figure 18 further underscores the model's high classification accuracy. Most data were correctly classified, with 42 True Negatives and 43 True Positives. The model made seven errors, consisting of 3 False Negatives and 4 False Positives.

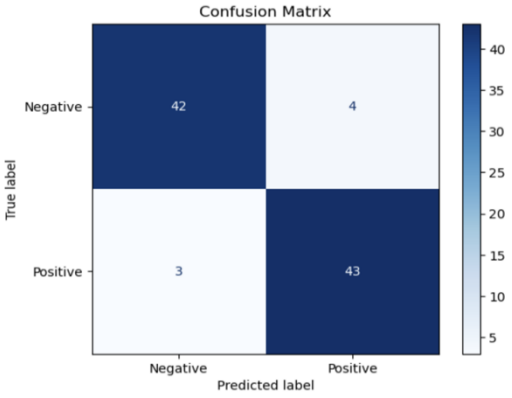

Figure 18. Confusion Matrix Results of the Hand Model

Furthermore, the evaluation metrics support the dependability of the model. A substantial ability to accurately identify positive cases while reducing the misclassification of negative data as positive is indicated by the precision of 91.49%. The model's 93.48% recall indicates it can correctly identify the majority of positive cases. A 92.47% F1 Score emphasizes a balanced connection between precision

and recall. The model performed robustly with few errors, achieving an overall accuracy of 92%. Despite minor fluctuations in validation loss, the results confirm the Xception architecture's effectiveness for hand image classification tasks.

Cross-validation results further confirm the model's consistency, showing an average accuracy of 92.36%, precision of 0.92, recall of 0.93, and F1 score of 0.92. These outcomes show that the model can function effectively across several kinds of data subsets.

The fold-wise cross-validation results, detailed in Tables 5 and 6, reveal varying performances, with trends showing improved accuracy, precision, recall, and F1 scores in the higher folds. These results provide valuable insights into the model's generalization capabilities on unseen data.

Table 5. Comparison of Accuracy Metrics Results Per Fold of Hand Model

Fold	Accuracy	Precision	Recall	F1 Score
1	75%	68%	86%	76%
2	89.58%	95%	84%	89%
3	99.31%	97%	99%	98%
4	97.90%	100%	96%	98%
5	100%	100%	100%	100%

Table 6. Comparison of Confusion Matrix Results Per Fold of Hand Model

Fold	True Negative	False Negative	False Positive	True Positive
1	44	10	29	61
2	66	12	3	63
3	66	1	2	75
4	69	3	0	71
5	80	0	0	63

Across five folds of cross-validation, the model showed steady improvement. Fold 1 had 75% accuracy with high recall (0.86) but low precision (0.68) due to 29 false positives. Accuracy improved to 89.58% in Fold 2, with precision at 0.95 and an F1 score of 0.89. Folds 3 and 4 achieved near-perfect results with accuracies of 99.31% and 97.90%, respectively, and minimal errors. Fold 5 reached 100% accuracy, precision, recall, and F1 score, confirming the model's strong generalization and reliability.

The confusion matrix for the overall model performance across all folds further reinforces its strong predictive capabilities. As seen in Figure 19, the model correctly classified 65 negative cases (True Negatives) and 67 positive cases (True Positives), with only minor misclassifications. Specifically, 5 positive cases were misclassified as negative (False Negatives), while 7 negative cases were incorrectly predicted as positive (False Positives). These results indicate a well-balanced performance, where the majority of predictions are accurate. The model demonstrates high reliability in identifying both positive and negative cases, showcasing a minimal error rate and reinforcing its robust classification performance across all tested folds.

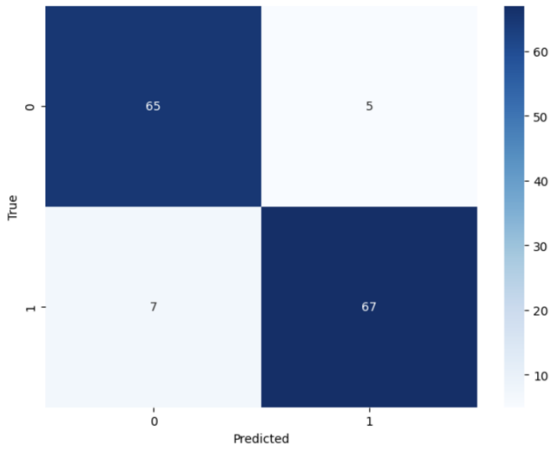

Figure 19. Overall Confusion Matrix Results Across All Folds of Hand Model Cross-validation

3.3 Discussion

This study's findings show significant advances in classifying hand and wrist fractures using deep learning models, particularly DenseNet and Xception. The findings emphasize the importance of model architecture and dataset quality in achieving superior accuracy and reliability for medical imaging tasks.

The DenseNet model achieved remarkable performance in wrist fracture classification with a 97.5% accuracy on the *Wrist Fracture - X-rays* dataset, surpassing the 81% reported by Kandel et al. [7]. Additionally, the evaluation metrics from this study, including the 100% precision, 95% recall, and 97.44% F1 score, demonstrate its ability to minimize false positives and reliably detect fractures. DenseNet's densely connected layers efficiently reuse features, enabling it to learn intricate patterns in X-rays, as reflected in its stable convergenve in training and validation graphs.

Similarly, the Xception model achieved strong results for hand fracture classification, a significant improvement over the 75% accuracy reported by Kandel et al. [7]. Although its precision (91.49%) and F1 score (92.47%) were slightly lower than DenseNet's, its improved performance in higher cross-validation folds demonstrates its potential for tasks requiring detailed pattern recognition. These findings align with Meena and Roy [9], who emphasized the effectiveness of CNN-based models for fracture detection but lacked results for specific bone regions.

The results also address limitations identified in prior studies. Gupta and Sharma [8] noted class imbalance in medical imaging datasets as a challenge, which was mitigated in this study by balancing datasets, resulting in DenseNet's minimal false positives and Xception's low error rates. However, the combined dataset experiment of the wrist model revealed challenges in merging heterogeneous datasets, reducing the accuracy to 89.9% due to inconsistencies in imaging standards and labeling. This finding underscores the need for harmonization techniques, such as domain adaptation, to address such issues.

The DenseNet model's cconfusion matrix revealed one false positive (FP) and two false negatives (FN) out of 82 samples. The minimal FP rate demonstrates DenseNet's high specificity, as it rarely misclassifies normal images as fractures. However, the two FNs indicate that subtle fracture features were missed, emphasizing the need for enhanced feature extraction to further improve sensitivity. These errors, while small, highlight the importance of refining the model to minimize missed diagnoses, which could delay treatment.

The Xception model's confusion matrix revealed one FP and three FNs out of 85 samples. The single FP indicates robust specificity, while the slightly higher FN rate suggests challenges in detecting subtle or overlapping fracture features. This is consistent with Xception's architectural focus on computational efficiency, which may sometimes sacrifice sensitivity to finer details. Targeted augmentation or preprocessing techniques could further enhance its performance in identifying complex patterns.

The results underline the importance of dataset quality and alignment with model architecture. DenseNet and Xception performed better on the curated Wrist Fracture - X-rays dataset compared to the MURA dataset, highlighting the importance of high-quality, curated datasets. The combined dataset experiment demonstrated that merging datasets with different imaging standards and annotation styles can introduce inconsistencies, reducing performance. This underscores the necessity of harmonization and advanced preprocessing techniques to mitigate dataset variability.

This research demonstrates significant progress in achieving clinically applicable accuracies compared to previous studies. The DenseNet and Xception models outperformed the results reported by Kandel et al. [7] and Gupta and Sharma [8], demonstrating superior generalization and robustness across all folds.

In answering the research question, this study confirms that DenseNet and Xception are effective for wrist and hand fracture classification, respectively. DenseNet excelled in wrist fracture classification, achieving clinical-grade accuracy and reliability, while Xception performed well in hand fracture classification with some room for improvement. The results also underscore the pivotal role of dataset quality, balanced data distribution, and architecture alignment in influencing model performance.

4. CONCLUSION

This study explored the application of DenseNet and Xception architectures, enhanced with transfer learning, for classifying hand and wrist fractures. The results demonstrated significant advancements in diagnostic accuracy, addressing key challenges identified in prior research, such as suboptimal model performance and dataset-related issues. This study successfully improved model reliability and generalization by leveraging high-quality datasets and robust preprocessing techniques.

DenseNet proved to be particularly effective for wrist fracture classification, achieving an accuracy of 97.5% and demonstrating strong generalization across cross-validation folds. The model's high precision (100%) and recall (92.5%) highlighted its ability to minimize diagnostic errors, making it a reliable tool for clinical applications. Similarly, Xception showed strong performance in hand fracture classification, achieving an accuracy of 92% and a balanced F1 score of 92.47%. While Xception exhibited variability in early-stage cross-validation folds, its performance improved significantly in later folds, demonstrating its potential for complex fracture classification tasks.

The findings underscore the importance of dataset quality and compatibility with the chosen model architecture. The use of the *Wrist Fracture - X-rays* dataset enabled stable training and better performance compared to the combined dataset approach, which encountered challenges due to inconsistencies in imaging standards. These results emphasize the need for careful dataset preparation, including strategies to address class imbalance and harmonize heterogeneous datasets.

Despite the promising results, the study faced several limitations. First, the variability observed in early-stage cross-validation folds for the Xception model suggests potential sensitivity to data distribution. Second, the combined dataset approach highlighted the challenge of integrating heterogeneous data, which introduced inconsistencies and affected model performance. Additionally, the study focused on limited datasets, which may not fully capture the diversity of fracture cases encountered in clinical settings. These limitations suggest that more study is necessary to overcome these challenges.

Future work should focus on optimizing the Xception model for hand fracture classification. This could involve expanding the dataset. Including multi-institutional and multi-regional datasets with varying imaging modalities, such as CT scans or MRIs, could increase the robustness and generalizability of the models. Addressing dataset heterogeneity through domain adaptation techniques or feature alignment strategies would further enhance performance when merging datasets with different characteristics.

Acknowledgment

We deeply appreciate the support provided by the School of Information Technology of Universitas Ciputra Surabaya.

5. REFERENCE

[1] "Musculoskeletal health." Accessed: Nov. 10, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/musculoskeletal-conditions

[2] F. Costa *et al.*, "Digital rehabilitation for hand and wrist pain: a single-arm prospective longitudinal

cohort study," *PAIN Rep.*, vol. 7, no. 5, p. e1026, Sep. 2022, doi: 10.1097/PR9.0000000000001026.

[3] J. Zhang *et al.*, "Deep learning assisted diagnosis system: improving the diagnostic accuracy of distal radius fractures," *Front. Med.*, vol. 10, p. 1224489, Aug. 2023, doi: 10.3389/fmed.2023.1224489.

[4] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights Imaging*, vol. 9, no. 4, pp. 611–629, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

[5] S. Solikhun, A. P. Windarto, and P. Alkhairi, "Bone fracture classification using convolutional neural network architecture for high-accuracy image classification," *Int. J. Electr. Comput. Eng. IJECE*, vol. 14, no. 6, Art. no. 6, Dec. 2024, doi: 10.11591/ijece.v14i6.pp6466-6477.

[6] P. Kora *et al.*, "Transfer learning techniques for medical image analysis: A review," *Biocybern. Biomed. Eng.*, vol. 42, no. 1, pp. 79–107, Jan. 2022, doi: 10.1016/j.bbe.2021.11.004.

[7] I. Kandel, M. Castelli, and A. Popovič, "Musculoskeletal Images Classification for Detection of Fractures Using Transfer Learning," *J. Imaging*, vol. 6, no. 11, p. 127, Nov. 2020, doi: 10.3390/jimaging6110127.

[8] S. Gupta and D. Sharma, "Bone Fracture Classification using Transfer Learning," Jun. 22, 2024, *arXiv*: arXiv:2406.15958. Accessed: Nov. 10, 2024. [Online]. Available: http://arxiv.org/abs/2406.15958

[9] T. Meena and S. Roy, "Bone Fracture Detection Using Deep Supervised Learning from Radiological Images: A Paradigm Shift," *Diagnostics*, vol. 12, no. 10, p. 2420, Oct. 2022, doi: 10.3390/diagnostics12102420.

[10] P. Rajpurkar *et al.*, "MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs." arXiv, May 22, 2018. Accessed: Nov. 10, 2024. [Online]. Available: http://arxiv.org/abs/1712.06957

[11] H. Malik, "Wrist Fracture - X-rays." Mendeley, Oct. 07, 2020. doi: 10.17632/XBDSNZR8CT.1.

[12] M. Kutbi, "Artificial Intelligence-Based Applications for Bone Fracture Detection Using Medical Images: A Systematic Review," *Diagnostics*, vol. 14, no. 17, p. 1879, Aug. 2024, doi: 10.3390/diagnostics14171879.

[13] T. Anwar and H. Anwar, "LSNet: a novel CNN architecture to identify wrist fracture from a small X-ray dataset," *Int. J. Inf. Technol.*, vol. 15, no. 5, pp. 2469–2477, Jun. 2023, doi: 10.1007/s41870-023-01311-w.

[14] U. Kuran, E. C. Kuran, and M. B. Er, "Parameter Selection of Contrast Limited Adaptive Histogram Equalization Using Multi-Objective Flower Pollination Algorithm," in *Electrical and Computer Engineering*, vol. 436, M. N. Seyman, Ed., in Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 436. , Cham: Springer International Publishing, 2022, pp. 109–123. doi: 10.1007/978-3-031-01984-5_9.

[15] Md. R. Islam and Md. Nahiduzzaman, "Complex features extraction with deep learning model for the detection of COVID19 from CT scan images using ensemble based machine learning approach," *Expert Syst. Appl.*, vol. 195, p. 116554, Jun. 2022, doi: 10.1016/j.eswa.2022.116554.

[16] Nurhidayah, B. Abdul Samad, and B. Abdullah, "Perbandingan Metode Contrast Enhancement pada Citra CT-Scan Kanker Paru-paru," *Gravitasi*, vol. 19, no. 2, pp. 24–28, Dec. 2020, doi: 10.22487/gravitasi.v19i2.15360.

[17] S. I. Sahidan, M. Y. Mashor, A. S. W. Wahab, Z. Salleh, and H. Ja'afar, "Local and Global Contrast Stretching For Color Contrast Enhancement on Ziehl-Neelsen Tissue Section Slide Images," in *4th Kuala Lumpur International Conference on Biomedical Engineering 2008*, vol. 21, N. A. Abu Osman, F. Ibrahim, W. A. B. Wan Abas, H. S. Abdul Rahman, and H.-N. Ting, Eds., in IFMBE Proceedings, vol. 21. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 583–586. doi: 10.1007/978-3-540-69139-6_146.

[18] Computer Engineering, Sriwijaya University, Indralaya, Indonesia and Erwin, "Improving Retinal Image Quality Using the Contrast Stretching, Histogram Equalization, and CLAHE Methods with Median Filters," *Int. J. Image Graph. Signal Process.*, vol. 12, no. 2, pp. 30–41, Apr. 2020, doi: 10.5815/ijigsp.2020.02.04.

[19] C.-F. W. Ron Kikinis and H. Knutsson, "Adaptive Image Filtering," in *Handbook of Medical Imaging*, Elsevier, 2000, pp. 19–31. doi: 10.1016/B978-012077790-7/50005-9.

[20] R. K. Patel and M. Kashyap, "Automated diagnosis of COVID stages from lung CT images using statistical features in 2-dimensional flexible analytic wavelet transform," *Biocybern. Biomed. Eng.*, vol. 42, no. 3, pp. 829–841, Jul. 2022, doi: 10.1016/j.bbe.2022.06.005.

[21] A. Majumder, A. Rajbongshi, Md. M. Rahman, and A. A. Biswas, "Local Freshwater Fish Recognition Using Different CNN Architectures with Transfer Learning," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 11, no. 3, pp. 1078–1083, Jun. 2021, doi: 10.18517/ijaseit.11.3.14134.

[22] F. A. Wicaksana, E. Mulyana, S. Hidayat, and R. Yusuf, "Design and Implementation Submarine Cable Object Detection YOLOv4 based with Graphical User Interface (GUI) for Remotely Operated Vehicle (ROV)," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 9, 2023, doi: 10.14569/IJACSA.2023.01409101.