FEC-IGE: An Efficient Approach to Classify Fracture Based on Convolutional Neural Networks and Integrated Gradients Explanation

Triet Minh Nguyen, Thuan Van Tran, Quy Thanh Lu
Information Technology Department
FPT University
Can Tho, Viet Nam

***Abstract*—In this paper, we propose the FEC-IGE framework includes data preprocessing, data augmentation, transfer learning, and fine-tuning of the pre-trained model of convolutional neural network (CNN) architecture for the problem of bone fracture classification. Bone fractures are a widespread medical issue globally, with a significant prevalence and imposing substantial burdens on individuals and healthcare systems. The impact of bone fractures extends beyond physical injury, often leading to pain, reduced mobility, and decreased quality of life for affected individuals. Moreover, fractures can incur substantial economic costs due to medical expenses, rehabilitation, and lost productivity. In recent years, progress in machine learning methodologies has exhibited potential in tackling issues pertaining to fracture diagnosis and classification. By harnessing the capabilities of deep learning frameworks, scholars aspire to design precise and effective mechanisms for automatically detecting and classifying bone fractures from medical imaging data. In this study, FEC-IGE framework has demonstrated its potential and strength when applied models pre-trained of CNN architecture in the task of classifying X-ray bone fracture images with accuracies of 98.48%, 96.92%, and 97.24% in three experimental scenarios. These outcomes are the consequence of the model's fine-tuning and transfer learning procedures applied to an enhanced dataset including 1129 X-ray pictures classified into ten different kinds of fractures: avulsion fracture, comminuted fracture, fracture dislocation, greenstick fracture, hairline fracture, impacted fracture, longitudinal fracture, oblique fracture, pathological fracture, and spiral fracture. To increase transparency and understanding of the model, Integrated Gradients explanation was also applied in this study. Finally, metrics including precision, recall, F1 score, precision, and confusion matrix were applied to evaluate performance and other in-depth analysis.**

Keywords*—Convolutional neural network; transfer learning; fine-tuning; X-ray image classification; EfficientNet; classification break bone; deep learning; integrated gradients explanation

I. Introduction

The musculoskeletal system, consisting of bones, muscles, and connective tissue, plays an important role in supporting the structure of the body and facilitating movement [1]. The human skeleton is a complex framework, providing protection for vital organs and serving as a fulcrum for muscles and ligaments [2]. Despite its resilience, the skeletal system is susceptible to various disorders and injuries, with fractures being one of the most common musculoskeletal injuries worldwide. Fractures occur when bones are subjected to excessive force or pressure, causing them to break or crack. Fractures can be classified based on severity, location, and whether the bone breaks through the skin (open fracture) or remains in tissue (closed fracture). Common types of fractures include stress fractures [3], hairline fractures, and compound fractures. Each type has distinct symptoms and treatments. Symptoms of a fracture may include localized pain, swelling, bruising, deformity, and impaired mobility, depending on the location and extent of the injury. Early detection and appropriate management of fractures is essential to promote optimal wound healing and prevent long-term complications, highlighting the importance emphasized by medical professionals [4].

A considerable percentage of health impacts are linked to bone fractures, as evidenced by the study [5], which examined 2,625,743 death certificates and found that 2.2% of them had a reference of a bone fracture. The statistics provided are based on data from the Global Burden of Disease Study 2019 (GBD 2019 Fracture Collaborators, 2021) [6]. The prevalence of bone fractures is a significant global health concern, with data indicating a steady increase in incidence over the years. According to the Global Burden of Disease Study 2019, there were approximately 178 million new cases of bone fractures worldwide in 2019, representing a significant increase of 33.4% since 1990. Moreover, an estimated 455 million individuals experienced acute or chronic symptoms associated with bone fractures, reflecting a substantial rise of 70.1% over the same period. The study also revealed that the burden of bone fractures varied across different age groups, with older adults being disproportionately affected. Specifically, individuals aged 95 years and older had the highest age-specific incidence rate of bone fractures, with 15,381.5 cases per 100,000 population. Furthermore, the consequences of bone fractures extend beyond physical discomfort, contributing to years lived with disability (YLD). In 2019, bone fractures resulted in approximately 25.8 million YLD globally, reflecting a 65.3% increase since 1990. These findings underscore the urgent need for comprehensive preventive measures and access to timely screening and treatment interventions to mitigate the overall burden of bone fractures on public health.

Radiography, another name for X-ray imaging [7], is essential for the diagnosis of bone fractures, which are a prevalent musculoskeletal ailment that afflicts people of all ages all over the world. By releasing electromagnetic radiation that enters the body and produces pictures dependent on the density of the tissues it encounters, X-ray scans offer precise representations of bone formations. Because bones absorb more X-rays than soft tissues, they show up in X-ray pictures as white regions, whereas the latter show various degrees of gray. The fact that X-ray imaging is used in clinical settings so often highlights how crucial it is for identifying fractures, determining how serious they are, and directing medical interventions.

Despite its effectiveness, conventional X-ray interpretation relies heavily on the expertise of radiologists and may be prone to errors or delays in diagnosis. As a result, there is a growing interest in leveraging advancements in machine learning techniques, such as transfer learning [8] and fine-tuning [9], to enhance fracture detection and classification accuracy. Transfer learning permits models pre-trained on expansive datasets to be adjusted to unused errands with restricted labeled information, making it well-suited for therapeutic imaging applications where clarified datasets may be rare. Fine-tuning pre-trained models by altering their parameters to superior adjust with the particular characteristics of the target errand, in this manner progressing execution. In their study, Huong Hoang Luong et al. [10] applied transfer learning with fine-tuning in the tasks of classifying abnormal and normal bones in the wrist, humerus, and elbow. By incorporating these machine learning approaches into the interpretation of X-ray images, clinicians can benefit from improved diagnostic accuracy, reduced interpretation time, and enhanced patient care in the diagnosis and management of bone fractures.

Convolutional Neural Networks (CNN) [11] have revolutionized the field of computer vision by enabling high-performance image recognition tasks. These networks are composed of multiple layers, including convolutional layers, pooling layers, and fully connected layers. Some popular CNN architectures that are powerful in the field of medical image analysis, allowing accurate and efficient diagnosis of various types of bone fractures, include AlexNet [12], VGG [13], MobileNet [14], ResNet [15], and EfficientNet [16]. These architectures vary in terms of depth, width, and complexity, with each designed to address specific challenges in image classification, object detection, or segmentation tasks. In particular, EfficientNet stands out as an efficient and high-performance CNN architecture with a relatively smaller model size compared to traditional networks. It introduces a novel compound scaling method that uniformly scales network depth, width, and resolution with a set of fixed scaling coefficients. This approach allows EfficientNet to achieve state-of-the-art performance with significantly fewer parameters compared to other architectures, making it a compelling choice for various computer vision applications.

In this investigation, the Integrated Gradients explanation will be utilized. Proposed by Sundararajan et al. [17], Integrated Gradients are employed to elucidate the predictions generated by our machine learning algorithm. In the context of medical diagnosis, such as classifying bone fracture images from X-ray images, transparency and interpretability are crucial to ensuring the reliability and accuracy of the model's decisions. By integrating Integrated Gradients into bone fracture classification applications, we enhance the interpretability of the model's predictions, thereby facilitating better decision-making and fostering user confidence. As demonstrated by previous studies [18] [19], the use of Integrated Gradients has proven effective in enhancing the transparency and interpretability of machine learning models in various medical imaging tasks.

In this study, we propose the **FEC-IGE** framework which is a combination of the words **F**racture problem, **E**fficient method, **C**lassification and **Integrated Gradients Explanation** for the fracture prediction problem. Additionally, we also implemented five popular CNN models (EfficientNetB3, ResNet50, VGG16, MobileNet và InceptionV3) into the FEC-IGE framework to evaluate the effectiveness of our proposed framework. We introduce three scenarios to assess the efficacy of the 10-class categorization of the dataset. The categorization involving avulsion fracture, comminuted fracture, fracture dislocation, greenstick fracture, and hairline fracture was executed under the initial scenario. The subsequent scenario involves the classification of impacted fracture, longitudinal fracture, oblique fracture, pathological fracture, and spiral fracture. The ultimate scenario entails the classification of all aforementioned 10 classes. The rationale behind the implementation of these three scenarios is to ascertain the effectiveness of the proposed model in classifying varying numbers of classes simultaneously.

The contributions of the research are:

- We propose the FEC-IGE framework including steps of data pre-processing, data augmentation, transfer learning, fine-tuning the pre-trained model CNN architecture, and visual explanation to classify 10 classes of fracture. Based on the augmented dataset, the results obtained are promising compared to other CNN architectures with up to 94.19% accuracy. By applying the techniques in the proposed that framework, we achieve promising results, outperforming other pre-trained models with an accuracy of up to 98.48% - 96.92% - 97.24% in three scenarios. This demonstrated the effectiveness of our proposed FEC-IGE framework in the image classification task.
- Proving that the proposed model (EfficientNetB3) is more effective than the ResNet50, VGG16, MobileNet, and InceptionV3 models in the bone classification problem by deploying all five models in the same situation.
- The empirical findings demonstrate the utility of Integrated Gradients explanation in enhancing comprehension of a machine learning model's decision-making process through the assessment of individual feature impact on model predictions. Integrated Gradients expound on the model at a local level, facilitating insight into the influence of each feature on the model's predictive outcomes.
- Research results will benefit users in early clinical work based on X-ray images. Physicians can benefit from improved diagnostic accuracy, reduced interpretation time, and enhanced patient care in the diagnosis and management of fractures.

Our research report comprises five primary components. Within this section, there is a provision of general information regarding the study and an outline of the methodology devised to address the specific challenge at hand. The references to the relevant research can be found in Section II, with the methodology aligning with the corresponding research segment. Section III delineates all the methodologies utilized in this study. The forthcoming Section IV will delve into the experiments, detailing the procedures followed and the evaluation of the accuracy of the deep learning model. Section V offers a discussion that synthesizes the data and information gathered in support of the objectives of this paper. Lastly, Section VI encapsulates our findings and scrutinizes the key elements pertinent to the research.

II. Related Works

Previously, the classification and diagnosis of bone fractures from X-ray images were mainly performed manually by medical professionals. However, with the rapid advancement of technology, artificial intelligence (AI) has emerged as a valuable tool in supporting crack detection, data collection, and classification. Recent studies, such as that of M. Jarke et al. [20], highlighted the key role of AI paradigms in growing capabilities across many domains. Furthermore, Muhammet Emin Sahin et al. [21] conducted various machine learning techniques using a dataset containing various bone types and finally proposed a computer-aided diagnosis (CAD) system to reduce the burden for doctors by identifying bone fractures with high accuracy.

Fırat Hardalaç et al. [22] investigated the effectiveness of deep learning models in detecting wrist fractures from X-ray images, with a focus on enhancing diagnostic accuracy in emergency care scenarios. Utilizing a comprehensive dataset from Gazi University Hospital, the research evaluates twenty fracture detection approaches employing various deep learning algorithms, including Libra R-CNN, FSAF, Faster R-CNN, Dynamic R-CNN, PAA, RegNet, RetinaNet, and DCN. Additionally, the study develops five ensemble models to fine-tune detection performance, leading to the creation of the innovative 'wrist fracture detection-combo (WFD-C)' model. The WFD-C model achieves the highest detection accuracy, with an average accuracy (AP50) of 86.39%, admitting its potential to significantly improve fracture diagnosis in clinical settings. Overall, this study has provided a lot of information from different methods for bone fracture detection, as well as their reputable data set. Although not as accurate as other studies, it has created a premise for future research. Research by Saurabh Verma et al. discussed in [23] focuses on the application of deep learning, specifically transfer learning, in the detection of open fractures using a limited medical imaging dataset. One of the main challenges addressed in the study is the unavailability of large datasets. To overcome this limitation, the authors used augmented datasets to increase the orientation and number of images. Deep learning-based CNN were used to overcome the limitations of limited training data availability. The study aimed to address the problem of open fracture detection using a limited number of images by applying the Speeded Up Robust Features (SURF) extraction tool to preprocessed radiographic images. The results of the SURF extractor are then fed into pre-trained models using transfer learning techniques. The proposed system achieves a high accuracy of 98.8% in detecting cracks from a given X-ray image. Comparative analysis shows that transfer learning provides comparable or even superior results compared to training models from scratch. However, the study also acknowledges the potential limitations of transfer learning, such as the vulnerability of overfitting with less training data, and the impact of poor preprocessing, which might lead to poor classification of data.

In 2023, research [24] by Mohamed A. Kassem et al. introduces an accurate computer-aided diagnosis system based on deep learning for pelvic fracture detection. In this study, they built an XAI (Explainable AI) framework for pelvic fracture classification. They used a dataset containing 876 X-ray images (472 pelvic fracture images and 404 normal images) to train the model. In this study, feature extraction was performed using GoogleNet, ResNet50, and AlexNet networks and Grad-CAM to validate that appropriate input pelvic segments are being activated during classification according to the relevant label. The results obtained were 98.5% for accuracy, sensitivity, specificity, and precision. Although the research achieved high accuracy and efficiency, the results were generated based on a rather modest data set, easily leading to overfitting and only classifying fracture images and normal images. Jichong Ying et al. [25] have trained several deep learning architectures, notably Adapted ResNet50 with SENet capabilities, to detect ankle fractures in a curated radiological picture dataset. Furthermore, Grad-CAM visuals are utilized to interpret model decisions. ResNet50 was tweaked with a higher SENet capacity than previous models, attaining 93% accuracy. Grad-CAM representations give extensive information about the radiograph regions that are critical to the model's decision-making. Their study observed that the Adapted ResNet50 model upgraded with SENet capabilities performed pretty well in identifying ankle fractures; nevertheless, we discovered that accuracy might still be improved because this is just a matter of defining a kind of ankle fracture.

A novel transfer learning strategy is presented by Zaenab Alammar et al. [26] in an effort to get beyond the restrictions of transfer learning that are present in the ImageNet dataset, which is located in a different domain. They suggested a transfer learning technique that entails fine-tuning a limited collection of annotated medical pictures to take use of previously learned training information, after which deep learning models are trained on many medical radiology images pertaining to the wrist and humerus from the musculoskeletal radiology (MURA) dataset. Their transfer learning approach produced impressive outcomes for models that were trained. The accuracy was 87.85%, the F1 score was 87.63%, and the Cohen's Kappa coefficient was 75.69% for the humerus classification. Similarly, the accuracy was 85.58%, the F1 score was 82.70%, and the Cohen's Kappa coefficient was 70.46% for wrist categorization. Visualization techniques, including gradient-based layer activation heat maps (Grad-CAM) and locally interpretable model-independent interpretation (LIME), have provided additional evidence supporting the superior accuracy of models trained with their Transfer Learning method compared to ImageNet Transfer Learning. Bhan et al. [27] employed feature fusion of deep learning techniques in a related work to determine if the MURA dataset had fractures or not; the five pre-trained models were MobileNetV2, ResNet-50, ResNeXt-50, DenseNet-169, and VGG16, which were then fused in this work. The feature-fusion strategy yielded 87.85% accuracy and a Cohen's Kappa of 75.72% for the humerus, while the shoulder attained 83.13% accuracy and a Cohen's Kappa of 66.25%. Although the accuracy is not high, the research has shown the performance of five separate model types.

In the field of diagnosis using machine learning, research by Huong Hoang Luong et al. contributed two important studies. In the study [28] Huong Hoang Luong et al. proposed a method using k-means clustering algorithm to classify MRI images of the brain into three different types of views (horizontal, facial, and cupping) and combine a The Residual Network (ResNet) was modified to diagnose three types of brain tumors: glioma and meningioma, pituitary adenoma, and identify tumor-free MRI images. The method was evaluated on datasets from Nanfang Hospital and Tianjin University of Medicine and Pharmacy Hospital, China, with MRI images. Their results achieved a brain tumor classification accuracy of 96%, the highest among the previously considered networks. In addition, they presented a model for classifying and detecting benign, malignant, and normal breast cancer that makes use of transfer learning and fine-tuning [29]. To detect breast cancer and improve prediction accuracy, they used transfer learning from a pre-trained MobileNet model to train the suggested model. 780 ultrasound pictures make up the dataset, which is divided into three categories: normal breast (133 photos), malignant breast cancer (210 images), and benign breast cancer (437 images). Applying the MobileNet model's transfer learning and fine-tuning procedures yields good results, according to experimental data, with accuracy values of 96.51%, 94.12%, and 90.60% for each of the three situations. Besides, we also found a lot of their research on classification and diagnosis problems with different models, including UNET [30], MobileNet [31], and ViT [17]. These studies contribute greatly to strengthening the direction of our research.

In the realm of bone fracture diagnosis through machine learning, a recent investigation conducted by Hoai Phuong Nguyen et al. introduced a novel approach rooted in deep learning for the identification of fractures within X-ray images of the humerus [32]. The study entailed the utilization of a composite algorithm comprising YOLACT++ for image segmentation and Contrast Limited Adaptive Histogram Equalization for enhancing image contrast during X-ray image preprocessing. Subsequently, the YOLOv4 model underwent training on a limited dataset employing four distinct data augmentation methods to detect and pinpoint fractures in X-ray images, culminating in an optimal performance of 81.91% with their devised technique. Furthermore, empirical findings validate the superiority of their approach over the Faster-RCNN solution when applied to constrained datasets. The research also underscores the necessity for further enhancements in the model to attain superior accuracy levels compared to commonly used models.

The present literature review is centered on the significant challenge presented by the dearth of annotated data within the medical sector, impeding the realization of the full potential and efficacy of machine learning. Previous research efforts primarily focused on the identification of singular or a small number of fracture categories, posing a challenge in disease diagnosis given the extensive array of fracture types requiring identification. It is this particular challenge that served as the impetus behind the primary aim of this study: to explore methodologies for enhancing performance levels under constrained data conditions in the realm of medical machine learning, while achieving precise identification of an expanding range of fracture variations.

III. Methodology

A. *The Methodology for Research Implementation*

Overall, the framework FEC-IGE, which comprises 11 processes, was employed in this study to create the results; the primary processes are depicted in Fig. 1. The steps are described in more detail below:

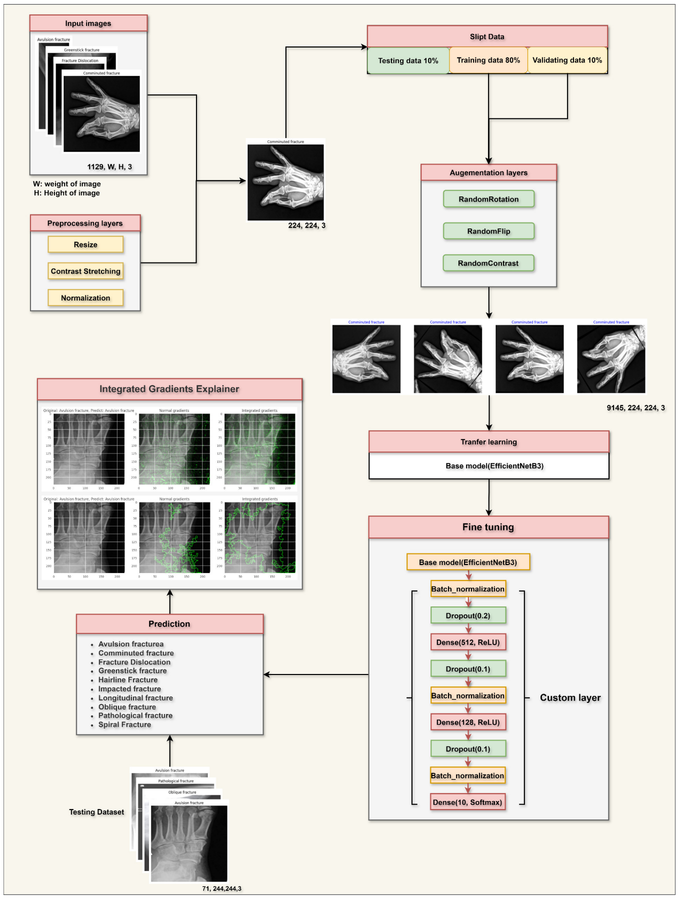

Fig. 1. The proposed FEC-IGE architectural framework.

1) *Data collection:* The selection of a suitable dataset is critical in the field of machine learning since it has a direct influence on model performance and generalization. In the context of bone fracture detection, selecting the appropriate dataset allows researchers to create models that can accurately identify fractures, allowing for faster diagnosis and treatment planning.

2) *Pre-processing Data:* Uses preprocessing methods, such as scaling input to 224x224 and changing brightness and contrast, to improve the quality and visibility of fracture images, making them more acceptable for future classification tasks.

3) *Divide the dataset into three categories: Training, validation, and testing:* TA test set of 10% of the data is used to evaluate the final model's performance on previously unknown data. Meanwhile, the validation set uses 10% of the data to evaluate training

progress and fine-tune the model to avoid overfitting. The remaining 80% constitutes the training set, encompassing all data utilized in training the model. Stratified splitting ensures that each subset maintains a balanced representation of classes, facilitating effective model training, validation, and evaluation.

4) *Data Augmentation:* To enhance the dataset, introduce diversity, establish reliability, and avoid overfitting, a range of data augmentation techniques are employed. Among the augmentation methods employed are RandomFlip, RandomContrast, and RandomRotation from the Keras library's Image augmentation layers. These approaches effectively expand the dataset without requiring additional data collection efforts.

5) *Building the model:* To conduct experiments, we adapted the EfficientNetB3 model architecture, leveraging its efficient and powerful convolutional neural network (CNN) architecture. We retained the core processing layers of the EfficientNetB3 model while making the necessary adjustments to optimize its performance for our specific task. This tailored approach allowed us to achieve exceptional results during training and testing using Keras's model library.

6) *Applying Transfer Learning:* Transfer learning enables the application of previously learned models for comparable tasks, such general picture categorization. These models have learnt fundamental characteristics from massive data, so we will save time and effort over training a model from scratch. Using a pre-trained model decreases the amount of technical time and resources required to deploy the model across several health systems.

7) *Retrain the model using Fine-Tuning:* Fine-tuning is the process of adjusting the weights of a previously trained model to suit your specific task. However, to actually apply these changes and improve model performance, model re-training is necessary. After fine-tuning, the model was adjusted to optimize for the specific task. Re-training the model allows it to learn more from new data, helping to improve generalization and prediction performance on new data.

8) *Validate and collect metrics to evaluate the model:* By measuring metrics like accuracy, precision, recall, and F1-score, we can evaluate how our model performs on new data that was not used during training. This process helps identify the model's loss on the test data set, while also providing an overview of how the model performs across different scenarios. The assessment results may be utilized to alter model hyperparameters such as batch size, neural network design, learning rate, and epochs. Based on the evaluation results, we can propose improvements or adjustments to the model to improve its performance and ensure its generality with new data.

9) *Visual explanation by Integrated Gradients:* Integrated Gradients provide clear explanations for the predictions generated by machine learning models by evaluating the influence of each individual input feature on the final prediction result. By clarifying the role of every input feature in the ultimate prediction, Integrated Gradients help improve the interpretability of the machine learning model, which is particularly advantageous in industries like healthcare, finance, and law where understanding how the model works is crucial.

10) *Compare to other sophisticated methods:* Comparison with other modern methods helps in analyzing the model's effectiveness and identifying the efficacy and uniqueness of the proposed strategy when compared to previously examined and acknowledged ways. This helps you to determine which components of your plan are more effective than others and which need to be modified.

11) *Showing the result:* The results and figures after comparison will be displayed in the form of confusion matrices, line graphs, and tables. The results demonstrate how the model performs in practice and how effective it is in diagnosing bone fractures.

B. Pre-processing Image

Pre-processing is an important step in preparing image data for machine learning tasks since it improves picture quality, consistency, and informativeness, ultimately enhancing model performance. In our study, we used a number of data pre-processing processes, as shown in Fig. 2.

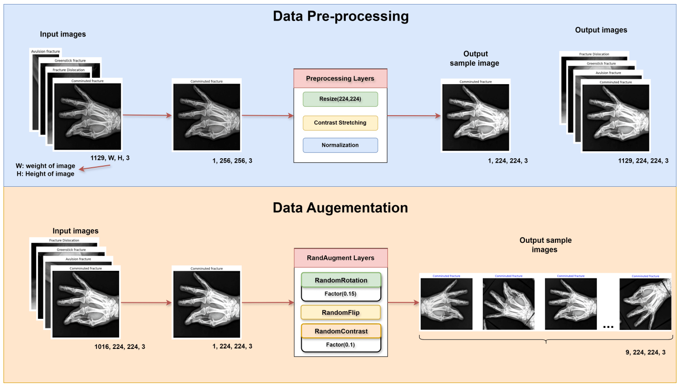

Fig. 2. Detailed proposed framework for image preprocessing.

1) *Resize image:* Achieving consistent input size is an important feature of picture preprocessing. To do this, we scaled all photos to a uniform size of 224 pixels (width) and 224 pixels (height), as specified by Eq. (1).

$$I_{ReSize}(new_{width}, new_{height}) = I_{ReSize}(224, 224) \quad (1)$$

2) *Contrast Stretching:* Contrast stretching is a method employed to enhance the contrast of an image by broadening the range of intensity values. This process involves the redistribution of pixel values to make use of the entire spectrum of intensities. This transformation is illustrated in Eq. (2), where the pixel values are redistributed to utilize the full range of intensities. I_{in} denotes the input intensity value of a pixel, while I_{out} signifies the corresponding output intensity value. The mathematical expression for contrast stretching can be formulated as:

$$I_{out} = \frac{I_{in} - I_{min}}{I_{max} - I_{min}} * 255 \tag{2}$$

where, I_{in} are the minimum and maximum intensity values in the input image, respectively.

3) *Data Augmentation:* Upon completing the initial image preprocessing procedures for data normalization, data augmentation techniques are implemented to enlarge both the training and validation datasets. This approach guarantees model interpretability by preserving the consistency of the test set data, thus preventing overfitting. First, we extract 903 images from the training sets and 113 images from validation sets to increase the number of images. The popular augmentation methods of the RandAugmentation Class in the Keras library were used. Those geometric transformations include rotation, flipping, and contrast adjustment. Finally, we found that the number of training and validation photos grew from 903 to 8128 images. Expanding the data set exposes the model to additional variables and scenarios, resulting in improved generalization and performance in real-world applications. In summary, data augmentation is a key strategy that increases the performance and generalization capacity of machine learning models, especially when vast and varied datasets are not available.

C. Transfer Learning and Fine-Tuning of EfficientNetB3

Transfer learning is a method in machine learning and deep learning in which we train a model on a large data set before reusing (transferring) it to solve a similar or related problem. Instead of starting from scratch on a small dataset, transfer learning allows us to use the knowledge and experience learned from previous training on large datasets to improve the model's performance on the new dataset [33]. During training, we reuse previously trained model parameters. As a result, transfer learning will use the model's current layers instead of retraining from scratch, thereby improving the model's accuracy.

Fine-tuning is the next step after applying transfer learning, the results will improve if we continue to perform fine-tuning. Fine-tuning changes and updates some parts of the pre-trained model (like the final layers) to fit the new dataset. By using information from pre-training and fine-tuning the model's representations to better match the target domain, fine-tuning allows the model to further tune its parameters to match the target, specifically fracture identification. Through the process of unlocking and purposefully training these layers, the model can improve its performance on the given task and the learned features. The final model is capable of learning unique patterns and sensitivities for the specific task, thereby improving the model's accuracy and elasticity in detecting bone fractures.

To maintain their capacity to extract low-level characteristics acquired during pre-training, the model's first layers are usually frozen during this procedure. This freezing method concentrates adaptation on the latter layers that are in charge of task-specific learning, which maximises training efficiency. In order to maximise model performance and avoid overfitting, fine-tuning also entails modifying hyperparameters, include the number of training epochs, hidden layer configurations, learning rate, and batch size. 50-100 epochs, 8–32 batch sizes, and hidden layer configurations like [256, 256, 128] or [512, 128] are examples of common hyperparameter search ranges. It is common practice to investigate the learning rate between 1e-3, 1e-4, and 1e-5.

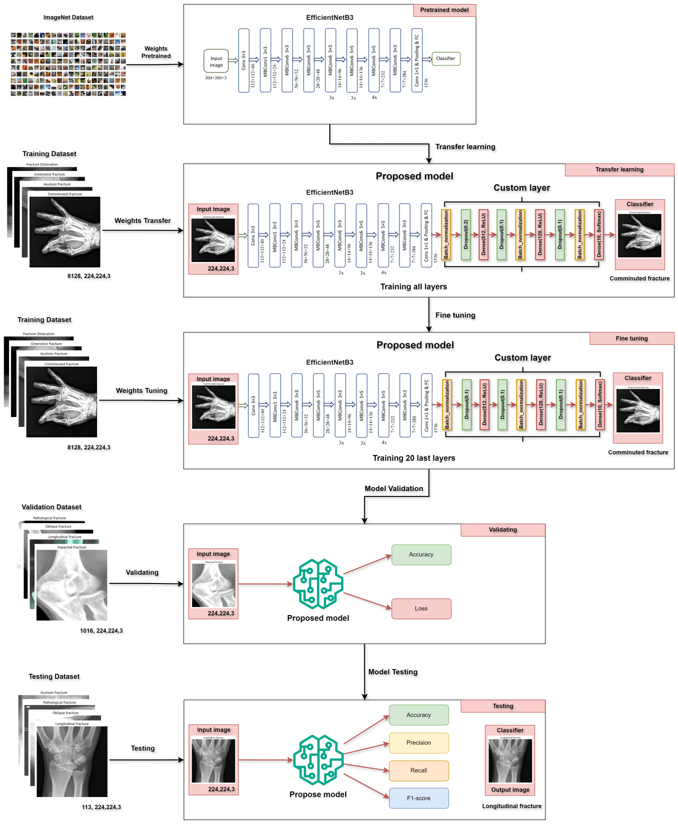

Fig. 3. Detailed proposed framework for transfer learning and fine tuning processes.

We fine-tuned the model using a hyperparameter search to achieve the best results while avoiding overfitting. This search looked at various combinations of training epochs, batch sizes, hidden layer configurations, and learning rates. Based on the findings, the following hyperparameters were chosen to strike a reasonable compromise between training efficiency and performance.

Furthermore, we added BatchNormalization and Dropout layers to the suggested design in order to enhance the model's capacity for generalization and lessen overfitting. Our proposed architecture is presented in Fig. 3.

D. Visual Explanation Using Integrated Gradients

The use of explanations is critical for gaining a better understanding of how a model makes decisions and forecast outcomes. This helps to improve the model's transparency and trustworthiness, particularly in industries such as healthcare, where a detailed description of the decision-making process is extremely useful for diagnosing and treating disorders.

Integrated Gradients is a way for clarifying the predictions of machine learning models, which helps comprehend how the model makes decisions depending on inputs. This approach assesses the relevance of each input characteristic by integrating along the path from a reference point to the individual data point under consideration. During this process, each feature progressively transitions from its reference value to its current value, allowing us to quantify the influence of each feature on the model's final prediction.

Suppose $IG(x)$ represents the Integrated Gradients for input (x), $f(z)$ is the model's output as a function of input (z), and $(\frac{\partial f(z)}{\partial z})$ is the gradient of the model's output concerning the input. The integral is computed from 0 to x, where x signifies the input to the model. The Integrated Gradients method is defined by Eq. (3).

$$IG(x) = \int_0^x \frac{\partial f(z)}{\partial z} dz \tag{3}$$

Integrated Gradients provide several advantages over other interpretation methods. Firstly, it offers computational efficiency and simplicity, allowing for accurate evaluation of individual feature importance. Second, because this technique does not need extensive understanding of the model's structure or properties, it is adaptable and suitable to a wide range of machine learning models. Finally, Integrated Gradients allow both quantitative and qualitative interpretation, providing a thorough knowledge of the model's decision-making process.

The usage of Integrated Gradients has been prevalent in different machine learning models, including deep neural networks, to increase transparency and interpretability. This approach is applicable for both classification and regression models. In scenarios involving non-scalar outputs, such as classification models or multi-target regression, gradients are produced for a single aspect of the output, which is often related with the model's actual or anticipated classes.

In conclusion, the use of Integrated Gradients for visually explanation is a viable technique for improving the transparency, accountability, and dependability of machine learning models, thereby increasing their value and credibility in real-world applications. In future projects, experts and medical practitioners can get significant insights by studying the influence of each feature map on the final choice, as shown in Fig. 4.

IV. Experiments

A. Dataset and Performance Metrics

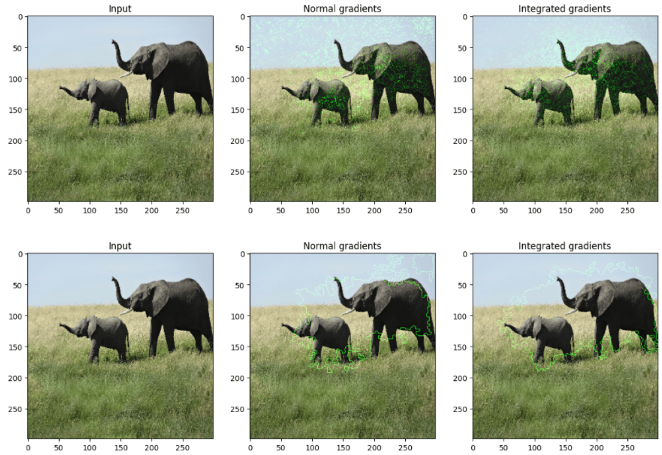

Fig. 4. The demo makes use of the keras library's integrated gradients.

This dataset was initially generated by Jason Zhang and Caden Li as part of Intel's RF100 program to develop a new object identification benchmark for model generalization [34]. The data set contains 1129 photos separated into ten classes; therefore, it is vital to provide various representations while lowering the danger of overfitting and improving the model's generalizability. After enhancing the training dataset and validation dataset, we obtain a new dataset with 8128 images, as shown in Fig. 5. Evaluating a machine learning model's performance is a critical step in both the research and deployment processes. In machine learning, various measures are used to evaluate a model's performance, including precision, recall, accuracy, and the F1-score.

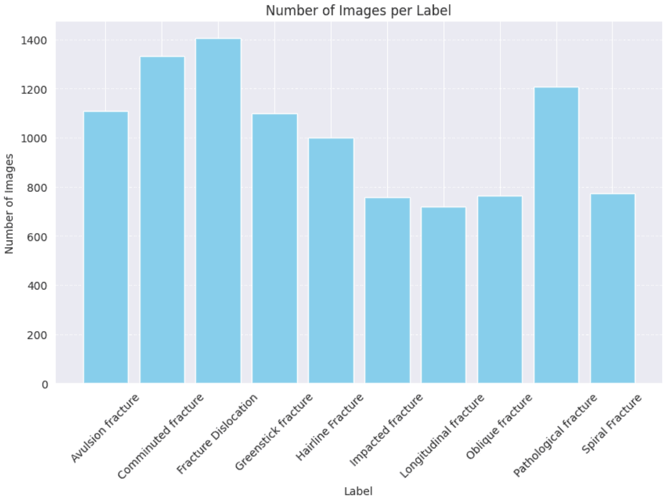

Fig. 5. Dataset characteristics after augmentation.

Accuracy is the ratio between the number of correct predictions and the total number of data samples in the test set. The mathematical formula for accuracy is given in Eq. (4):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Precision measures the ratio between the number of correct positive predictions (True Positive) and the total number of positive predictions (True Positive + False Positive). Precision provides information about the accuracy of positive predictions. Precision's mathematical formula is given in Eq. (5):

$$\text{Precision} \;=\frac{TP}{TP+FP} \tag{5}$$

Recall (also known as Sensitivity) measures the ratio between the number of true positive predictions and the total number of truly positive samples in the data set. Recall provides information about the model's ability to find all positive cases. The mathematical formula of Recall is given in Eq. (6):

$$\text{Recall} \;=\frac{TP}{TP+FN} \tag{6}$$

F1-score is a combined measure of Precision and Recall, often used when both values need to be considered. F1-score is the harmonic average of Precision and Recall and is calculated by the Eq. (7):

$$\text{F1} \;=2*\frac{Precision*Recall}{Precision+Recall} \tag{7}$$

These metrics provide a comprehensive view of the performance of a machine learning model, allowing users to accurately evaluate its predictive ability and find important cases.

B. Scenario 1: X-ray image classification results of the first 5 classes: Avulsion fracture, Comminuted fracture, Fracture Dislocation, Greenstick fracture, Hairline Fracture

In this situation, we use transfer learning and fine-tuning with and without data augmentation to categorize five fracture classifications using five distinct machine learning models. The transfer learning results in Table I demonstrate the model's efficacy on the augmented data set. The suggested model's accuracy has increased from 63.63% to 95.45%. In addition to the suggested model, the ResNet50 model achieves 96.63% accuracy, indicating great efficiency. Table II shows the fine-tuning results before and after increasing the data set as 62.12%-98.48%.

TABLE I. The Results of Categorizing X-Ray Images Into Five First Classes in Transfer Learning

Transfer learning Without Augmentation				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
ResNet50	63,63%	66,85%	63,63%	63,56%
VGG16	51,51%	51,41%	51,51%	50,92%
MobileNet	33,33%	35,43%	33,33%	32,57%
InceptionV3	42,42%	43,98%	42,42%	42,96%
Our Proposed	**63,63%**	**65,53%**	**63,63%**	**63,72%**
Transfer learning With Augmentation				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
ResNet50	96,63%	96,63%	96,63%	96,63%
VGG16	89,22%	89,47%	89,22%	89,21%
MobileNet	69,86%	69,99%	69,86%	69,66%
InceptionV3	60,60%	60,82%	60,60%	60,58%
Our Proposed	**95,45%**	**95,64%**	**95,45%**	**95,46%**

Fig. 6 and Fig. 7 provide a graph of the training process's accuracy and loss. During the training process, the two curves gradually grow and eventually stabilize. This demonstrates that the model strikes a balance between learning from training data and generalizing to new data. Overall, the curves for training and loss accuracy curves are smooth, with no significant variation between them, indicating that the model is appropriate and has strong generalization ability.

TABLE II. The Results of Categorizing X-Ray Images Into Five First Classes in Fine-Tuning

Fine-Tuning Without Augmentation				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
ResNet50	60,60%	62,17%	60,60%	60,30%
VGG16	39,39%	39,75%	39,39%	39,22%
MobileNet	19,69%	14,77%	19,69%	8,41%
InceptionV3	33,33%	35,94%	33,33%	33,66%
Our Proposed	**62,12%**	**63,54%**	**62,12%**	**62,22%**
Fine-Tuning With Augmentation				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
ResNet50	98,48%	98,51%	98,48%	98,48%
VGG16	58,41%	66,57%	58,41%	56,38%
MobileNet	38,55%	64,89%	38,55%	30,04%
InceptionV3	64,14%	64,51%	64,14%	64,13%
Our Proposed	**98,48%**	**98,49%**	**98,48%**	**98,48%**

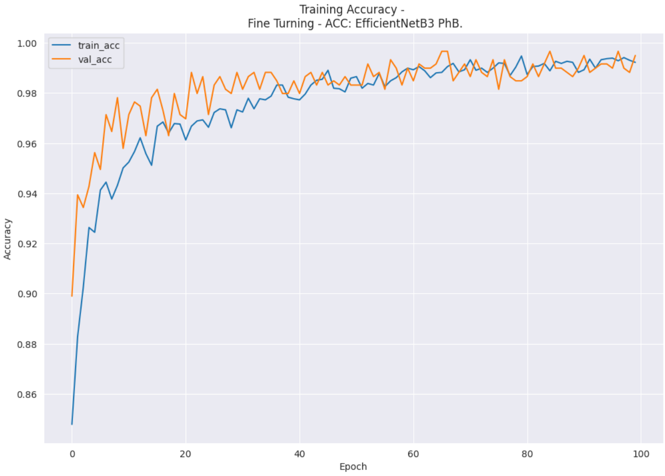

Fig. 6. Accuracy of training and validation while fine-tuning our model (5 First Classes).

The confusion matrix pictures of five different kinds of fractures—avulsion, comminuted, fracture dislocation, greenstick, and hairline—are shown in Fig. 8. The outcome of the Integrated Gradients explanation is Fig. 9, which illustrates how each feature helps to push the model output from the baseline value—the average model output across the training dataset we passed—to the model output. The training process is transparent, as seen by the two images above, and overfitting is not an issue.

C. Scenario 2: X-ray image classification results of the last 5 classes: Impacted fracture, Longitudinal fracture, Oblique fracture, Pathological fracture, Spiral fracture

In this scenario, we classify the next five types of fractures out of a total of 10 types, including spiral fracture, impacted

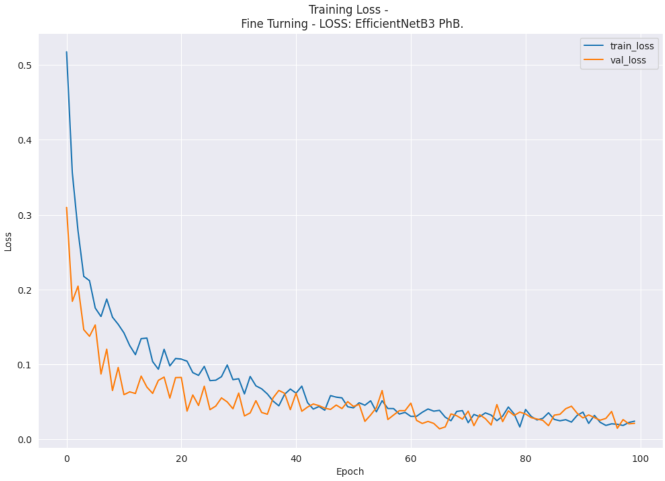

Fig. 7. Accuracy of validation and training loss throughout our model's fine-tuning (5 First Classes).

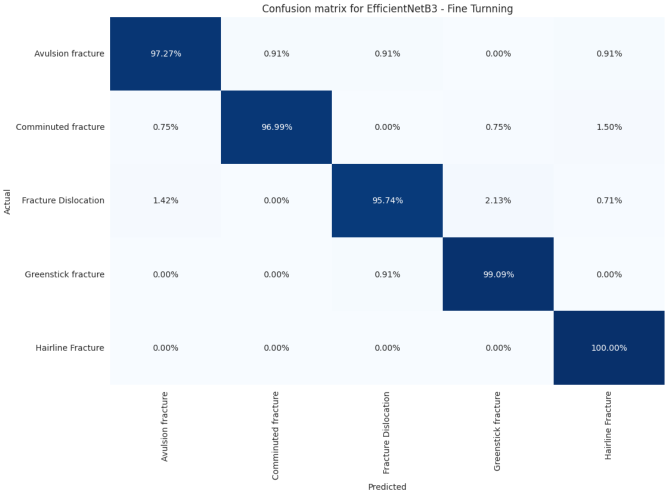

Fig. 8. Confusion matrix during our model's fine tuning (5 First Classes).

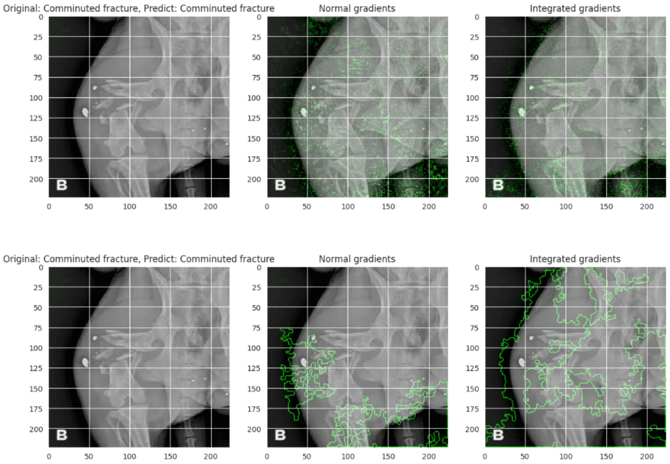
Original: Comminuted fracture, Predict: Comminuted fracture
Normal gradients
Integrated gradients
Original: Comminuted fracture, Predict: Comminuted fracture
Normal gradients
Integrated gradients
B

Fig. 9. Our model's output in scenario 1 with integrated gradients explanation.

fracture, pathological fracture, oblique fracture, and longitudinal fracture. Transfer learning and fine-tuning are carried out by the scenario both with and without data augmentation. 96.21% accuracy was achieved in the transfer learning portion of the suggested model, which is better than 40% when compared to training on the original data set (Table III). Additionally, Table IV illustrates the efficacy of fine-tuning when the achieved accuracy is greater than transfer learning, at 96.92%.

The accuracy and loss of the training process in the second scenario experiment are displayed in Fig. 10 and 11. The two curves grow steadily and don't differ much from one another during the training period. The training and loss accuracy curves are generally smooth and show little variance, suggesting that the model is suitable and capable of high generalization.

TABLE III. The Results of Categorizing X-Ray Images Into Five Last Classes in Transfer Learning

Transfer learning Without Augmentation				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
ResNet50	55,31%	52,38%	55,31%	51,54%
VGG16	27,65%	7,65%	27,65%	11,98%
MobileNet	34,04%	31,42%	34,04%	32,26%
InceptionV3	25,53%	24,48%	25,53%	23,97%
Our Proposed	**53,19%**	**57,40%**	**53,19%**	**52,93%**
Transfer learning with Augmentation				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
ResNet50	93,38%	93,45%	93,38%	93,38%
VGG16	88,41%	88,40%	88,41%	88,35%
MobileNet	71,63%	71,64%	71,63%	71,51%
InceptionV3	62,64%	62,75%	62,64%	62,59%
Our Proposed	**96,21%**	**96,27%**	**96,21%**	**96,21%**

TABLE IV. The Results of Categorizing X-Ray Images Into Five Last Classes in Fine-Tuning

Fine-Tuning Without Augmentation				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
ResNet50	51,06%	50,10%	51,06%	48,34%
VGG16	19,14%	3,66%	19,14%	6,15%
MobileNet	25,53%	19,55%	25,53%	16,02%
InceptionV3	40,42%	42,47%	40,42%	40,03%
Our Proposed	**42,55%**	**36,34%**	**42,55%**	**38,00%**
Fine-Tuning With Augmentation				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
ResNet50	96,69%	96,71%	96,69%	96,69%
VGG16	74,94%	77,04%	74,94%	75,04%
MobileNet	38,55%	43,02%	43,02%	42,78%
InceptionV3	62,17%	62,17%	62,17%	62,09%
Our Proposed	**96,92%**	**96,97%**	**96,92%**	**96,93%**

Fig. 12 presents the confusion matrix images of 5 types of fractures, including spiral fracture, impacted fracture, pathological fracture, oblique fracture, and longitudinal fracture. The matrix shows that, with an accuracy rate of 100%, the model performs best when diagnosing oblique fractures. In addition, compared to the other classes, the longitudinal fracture class has a larger mistake rate. The outcome of the Integrated Gradients explanation for this case is shown in Fig. 13.

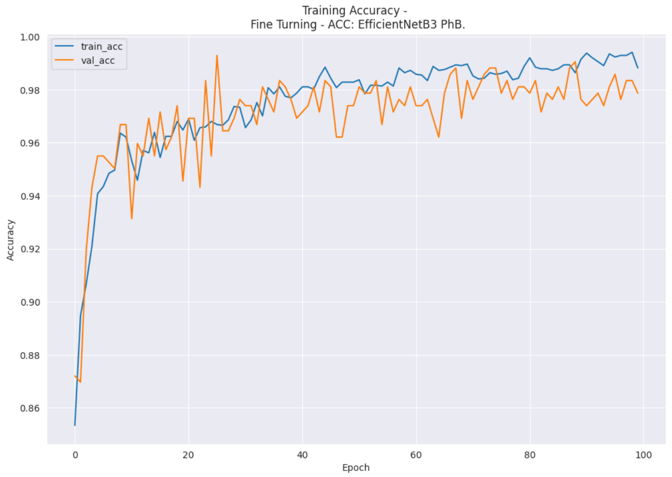

Fig. 10. Accuracy of training and validation in optimizing our model (5 Last Classes).

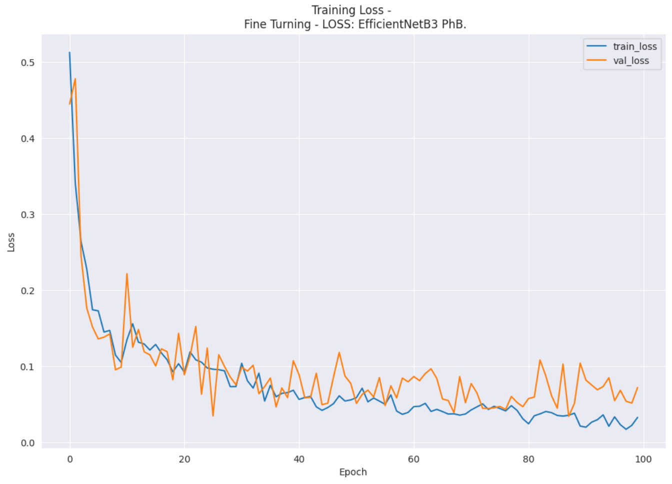

Fig. 11. Accuracy of validation and training loss throughout our model's fine-tuning (5 Last Classes).

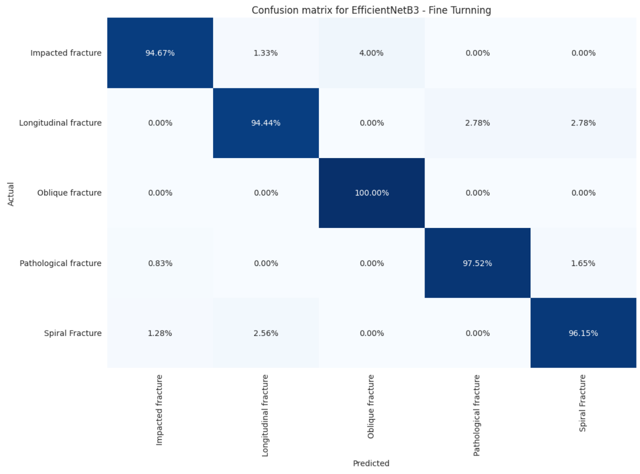

Fig. 12. Confusion matrix during our model's fine tuning (5 Last Classes).

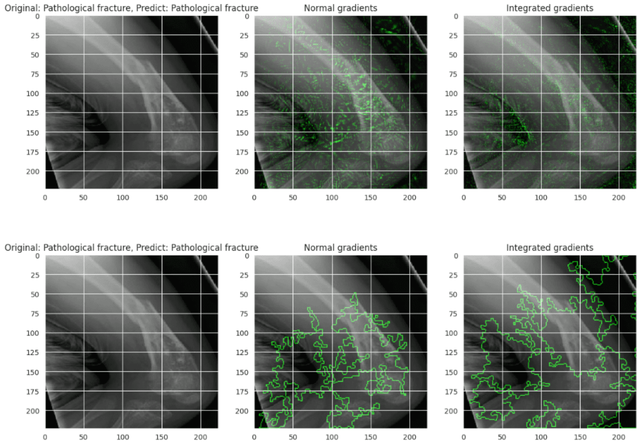
Original: Pathological fracture, Predict: Pathological fracture
Normal gradients
Integrated gradients
Original: Pathological fracture, Predict: Pathological fracture
Normal gradients
Integrated gradients
0
25
50
75
100
125
150
175
200
0
50
100
150
200

Fig. 13. Our model's output in scenario 2 with integrated gradients explanation.

D. Scenario 3: X-ray image classification results of the 10 classes: avulsion fracture, comminuted fracture, fracture dislocation, greenstick fracture, hairline fracture, impacted fracture, longitudinal fracture, oblique fracture, pathological fracture, and spiral fracture

This crucial case demonstrates the suggested model's excellent performance in handling a classification issue with up to ten classes. Table V illustrates that the suggested model attained 94% accuracy following the transfer learning procedure, which is greater than ResNet50's 92.82%. Following the phase of fine-tuning the suggested model using the expanded data set, Table VI presents the final accuracy result, which is 96.85%.

TABLE V. THE RESULTS OF CATEGORIZING X-RAY IMAGES INTO TEN CLASSES IN TRANSFER LEARNING

Transfer learning Without Augmentation				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
ResNet50	43,36%	42,69%	43,36%	42,85%
VGG16	19,46%	5,25%	19,46%	8,23%
MobileNet	30,97%	33,34%	30,97%	30,66%
InceptionV3	38,05%	39,12%	38,05%	37,85%
Our Proposed	**51,32%**	**52,41%**	**51,32%**	**51,10%**
Transfer learning with Augmentation				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
ResNet50	92,82%	92,92%	92,82%	92,81%
VGG16	84,75%	84,91%	84,75%	84,71%
MobileNet	56,93%	57,10%	56,93%	56,88%
InceptionV3	53,29%	53,64%	53,29%	53,26%
Our Proposed	**94,00%**	**94,05%**	**94,00%**	**93,99%**

Fig. 14 and Fig. 15 illustrate the accuracy and loss of the training process in the experiment of scenario 3. During the training process, the two curves steadily increase and do not deviate significantly from each other, indicating the transparency and reliability of the proposed model.

Fig. 17 presents the confusion matrix images of 10 types of fractures, including avulsion fracture, comminuted fracture, fracture dislocation, greenstick fracture, hairline fracture, impacted fracture, longitudinal fracture, Oblique fracture, pathological fracture, and spiral fracture. The matrix shows that,

TABLE VI. The Results of Categorizing X-Ray Images Into Ten Classes in Fine-Tuning

Fine-Tuning Without Augmentation				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
ResNet50	49,55%	48,03%	49,55%	48,42%
VGG16	1,54%	12,38%	2,75%	2,75%
MobileNet	30,08%	50,86%	30,08%	24,59%
InceptionV3	37,16%	38,05%	37,16%	36,75%
Our Proposed	**51,32%**	**50,36%**	**51,32%**	**50,09%**
Fine-Tuning With Augmentation				
Model	**Accuracy**	**Precision**	**Recall**	**F1**
ResNet50	94,19%	94,27%	94,19%	94,15%
VGG16	56,93%	59,22%	56,93%	56,53%
MobileNet	39,23%	64,31%	39,23%	34,47%
InceptionV3	51,72%	51,66%	51,72%	51,51%
Our Proposed	**97,24%**	**96,92%**	**97,24%**	**96,86%**

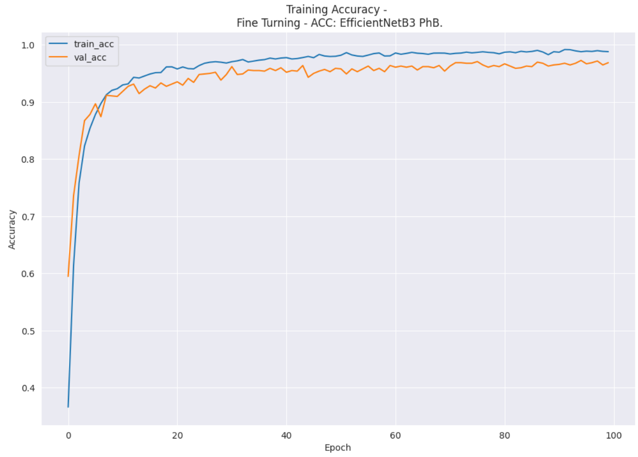

Fig. 14. Accuracy of training and validation in optimizing our model (Full 10 Classes).

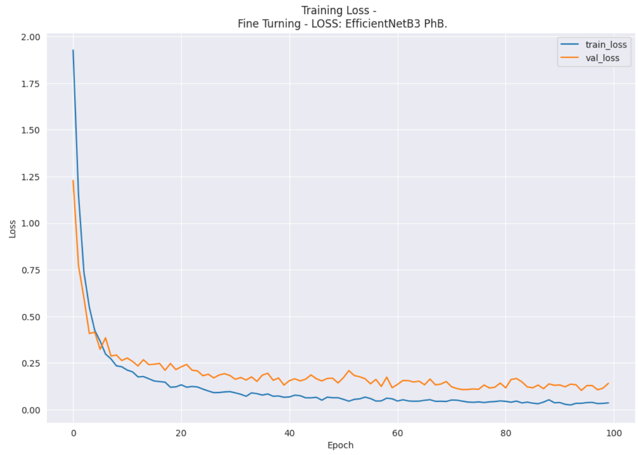

Fig. 15. Accuracy of validation and training loss throughout our model's fine-tuning (Full 10 Classes).

with a 98% accuracy rate, the model performs best when it comes to diagnosing fracture dislocation. Spiral fault layers, at around 10%, have the highest failure rate at the same time. The outcome of the Integrated Gradients explanation for categorizing ten different types of fractures is shown in Fig. 16.

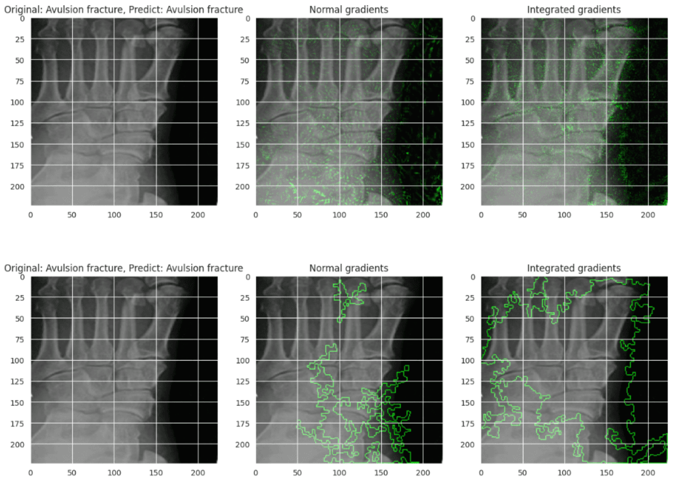
Original: Avulsion fracture, Predict: Avulsion fracture
Normal gradients
Integrated gradients
Original: Avulsion fracture, Predict: Avulsion fracture
Normal gradients
Integrated gradients

Fig. 16. Our model's output in scenario 3 with integrated gradients explanation.

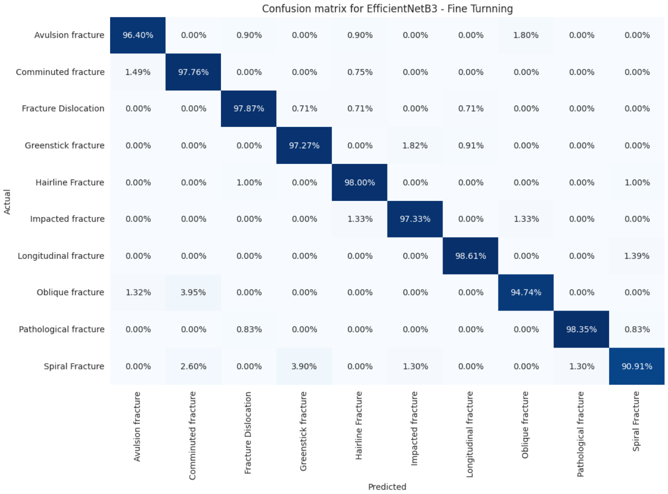

Fig. 17. Confusion matrix during our model's fine tuning (Full 10 Classes).

E. Comparison with others State-of-the-art Methods

This section totally compares our proposed method to several existing state-of-the-art categorization methodologies. Table VII compares categorization methods, their respective accuracy rates, and our proposed strategy.

The models studied include a wide range of approaches and architectures for different machine learning problems. MobileNet, ResNet50, EfficientNetV2, GoogleNet, and YOLO are all convolutional neural network (CNN) models notable for their performance and efficiency in image categorization and feature extraction. Additionally, Vision Transformer (ViT) is a recently suggested architecture that employs a self-attention mechanism to capture long-range relationships in pictures, making it appropriate for tasks like image categorization and object identification. The essential point is that the model we present outperforms similar and recent studies on the topic of

TABLE VII. Comparison with Others State-Of-The-Art Methods

Ref.	Architecture	ACC
Huong Hoang Luong et al. [10]	MobileNet	84%
Hadeer El-Saadawy et al. [14]	MobileNet	73,42%
Lee-Ren Yeh et al. [15]	ResNet	92%
Fırat Hardalac et al. [22]	WFD-C	86,39%
Saurabh Verma et al. [23]	CNN	98,8%
Mohamed A. Kassem et al. [24]	GoogleNet	98,5%
Hoai Phuong Nguyen et al. [32]	YOLOv4	81,91%
Bhan et al. [31]	CNN	87,85%
Jichong Ying et al. [25]	ResNet50	93%
Xuebin Xu et al. [35]	EfficientNetV2	78,12%
Hang Min et al. [36]	YOLOv5	81%
Leonardo Tanzi et al. [37]	ViT	97%
Our Proposed Model		**97,24%**

bone fracture classification.

V. Discussion

Upon the application of the FEC-IGE framework, not only is the power of deep learning harnessed, but also advanced techniques such as data preprocessing, augmentation, transfer learning, and fine-tuning of the EfficientNetB3 pre-trained model [16] are integrated. A comprehensive series of experiments has been carried out to assess the efficacy of this proposed methodology.

Aside from the exceptional performance demonstrated by the FEC-IGE framework, surpassing previous studies on pre-trained models in skin disease classification, several aspects deserve consideration. Initially, although data augmentation methods have played a crucial role in addressing data imbalance and enhancing model performance, the most straightforward approach to enhancing model efficacy remains the enlargement of the original dataset. This is particularly pertinent given the current constraints in obtaining high-quality, annotated datasets for skin diseases. Secondly, the incorporation of pre-trained model weights into the revamped model has notably enhanced both the training efficiency and model performance. This strategy has been investigated in recent research, showcasing its effectiveness in boosting model performance. Nevertheless, the issue of how pre-trained models effectively bridge the gap between medical and natural images remains a subject requiring further exploration.

The restricted availability of training data presents a hurdle in fully exploiting the discriminative capabilities of the FEC-IGE framework. Consequently, while the proposed EfficientNetB3 model yielded satisfactory outcomes in five models utilizing the FEC-IGE framework, instances persist where its performance falls short (MobileNet [14], InceptionV3 [38]). Despite the enhancement in performance across all models post-framework implementation, there are certain models that do not attain high accuracy levels. This underscores specific challenges that have not been adequately tackled within the existing framework.

In conclusion, the FEC-IGE framework makes notable contributions to skin disease classification through its superior performance, versatility, and the incorporation of Integrated Gradients for visual explication. Nonetheless, there is room for improvement, particularly in elevating model accuracy and deploying the model on mobile or web-based platforms for fracture classification. This area represents a promising avenue for future investigation, aimed at rendering fracture classification more accessible and precise for healthcare practitioners and patients alike.

VI. Conclusion

In the realm of fracture classification, our proposed FEC-IGE framework stands out for its innovative approach and superior performance compared to other state-of-the-art methods. The FEC-IGE framework, which encompasses data preprocessing, data augmentation, transfer learning, and fine-tuning of the EfficientNetB3 pre-trained model, has demonstrated remarkable effectiveness in classifying ten distinct classes of fracture.

Our framework's performance is particularly noteworthy when applied to other pre-trained models such as ResNet50, VGG16, MobileNet, InceptionV3, and EfficientNetB3. In three different cases, our FEC-IGE framework achieved an accuracy of 98.48% - 96.92% - 97.24%, respectively, significantly outperforming these models. This superior performance is attributed to the meticulous steps of data preprocessing and augmentation, which enhance the model's ability to generalize from the training data to unseen fracture images. Additionally, the fine-tuning of the EfficientNetB3 pre-trained model tailored to our specific task has allowed our framework to adapt and optimize its performance for fracture classification.

Furthermore, the trying to apply the FEC-IGE framework to five well-known CNN architectures (ResNet50, VGG16, MobileNet, InceptionV3, and EfficientNetB3) resulted in a substantial performance improvement across all models. This demonstrates the versatility and robustness of our framework, capable of enhancing the performance of a wide range of CNN architectures in the classification of fractures.

The high accuracy rate of the FEC-IGE framework after applying it to the EfficientB3 model of 97.24% in fracture classification is a testament to its effectiveness. This level of accuracy not only enables precise recognition of distinct skin conditions but also supports the development of precise treatment strategies. The validation process has further highlighted the importance of data augmentation and fine-tuning in improving the system's efficacy.

Another significant contribution of our work is the integration of Integrated Gradients for visual explanation. This method has proven to be beneficial in enhancing the understanding of the decision-making process of the model. By providing lucid and comprehensible explanations, Integrated Gradients contribute to the reliability and credibility of the model's predictions. This approach is particularly valuable in domains such as medicine and security, where transparency and understanding of the model's decision-making process are paramount.

In conclusion, the FEC-IGE framework's contributions to fracture classification through superior performance, versatility, and the integration of Integrated Gradients for visual explanation, set it apart from other state-of-the-art methods. These advancements not only demonstrate the effectiveness of our proposed framework but also pave the way for future research in the application of machine learning in healthcare.

Availability of Data, Code, and Material

Data for this study are published on repository link at [1] and code is at [2]

Acknowledgment

We would like to express my sincere gratitude to Huong Hoang Luong, Duy Khanh Nguyen, and Bang Huu Do Dang for their invaluable support and assistance throughout the course of this research. Their expertise, guidance, and encouragement have been instrumental in the successful completion of this study. We are truly grateful for their dedication and commitment, which have greatly contributed to the quality and depth of this research endeavor.

References

[1] M. Nordin and V. H. Frankel, *Basic biomechanics of the musculoskeletal system*. Lippincott Williams & Wilkins, 2001.

[2] R. B. Martin, D. B. Burr, N. A. Sharkey, D. P. Fyhrie *et al.*, *Skeletal tissue mechanics*. Springer, 1998, vol. 190.

[3] A. D. Perron, W. J. Brady, and T. A. Keats, "Principles of stress fracture management: the whys and hows of an increasingly common injury," *Postgraduate medicine*, vol. 110, no. 3, pp. 115–124, 2001.

[4] S. D. Kingma and A. I. Jonckheere, "Mps i: Early diagnosis, bone disease and treatment, where are we now?" *Journal of Inherited Metabolic Disease*, vol. 44, no. 6, pp. 1289–1310, 2021.

[5] N. Ziadé, E. Jougla, and J. Coste, "Using vital statistics to estimate the population-level impact of osteoporotic fractures on mortality based on death certificates, with an application to france (2000-2004)," *BMC public health*, vol. 9, pp. 1–14, 2009.

[6] G. . F. Collaborators, "Global, regional, and national burden of bone fractures in 204 countries and territories, 1990-2019: a systematic analysis from the global burden of disease study 2019," *The Lancet. Healthy longevity*, vol. 2, no. 9, pp. e580–e592, 2021.

[7] B. E. Warren, *X-ray Diffraction*. Courier Corporation, 1990.

[8] S. Goswami, C. Anitescu, S. Chakraborty, and T. Rabczuk, "Transfer learning enhanced physics informed neural network for phase-field modeling of fracture," *Theoretical and Applied Fracture Mechanics*, vol. 106, p. 102447, 2020.

[9] U. B. Abubakar, M. M. Boukar, and S. Adeshina, "Evaluation of parameter fine-tuning with transfer learning for osteoporosis classification in knee radiograph," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 8, 2022.

[10] H. H. Luong, L. T. T. Le, H. T. Nguyen, V. Q. Hua, K. V. Nguyen, T. N. P. Bach, T. N. A. Nguyen, and H. T. Q. Nguyen, "Transfer learning with fine-tuning on mobilenet and grad-cam for bones abnormalities diagnosis," *Complex, Intelligent and Software Intensive Systems*, pp. 171–179, 2022.

[11] K. O'shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[12] B. Y. Panchal, B. Talati, S. Shah, and S. PATEL, "Bone fracture classification using modified alexnet," *Stochastic Modeling & Applications*, vol. 26, no. 3, 2022.

[13] A. M. Barhoom, M. R. J. Al-Hiealy, and S. S. Abu-Naser, "Bone abnormalities detection and classification using deep learning-vgg16 algorithm," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 20, pp. 6173–6184, 2022.

[14] H. El-Saadawy, M. Tantawi, H. A. Shedeed, and M. F. Tolba, "A two-stage method for bone x-rays abnormality detection using mobilenet network," in *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*. Springer, 2020, pp. 372–380.

[15] L.-R. Yeh, Y. Zhang, J.-H. Chen, Y.-L. Liu, A.-C. Wang, J.-Y. Yang, W.-C. Yeh, C.-S. Cheng, L.-K. Chen, and M.-Y. Su, "A deep learning-based method for the diagnosis of vertebral fractures on spine mri: retrospective training and validation of resnet," *European Spine Journal*, vol. 31, no. 8, pp. 2022–2030, 2022.

[16] A. S. Bayangkari Karno, W. Hastomo, T. Surawan, S. R. Lamandasa, S. Usuli, H. R. Kapuy, and A. Digdoyo, "Classification of cervical spine fractures using 8 variants efficientnet with transfer learning." *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 13, no. 6, 2023.

[17] H. T. Nguyen, T. D. Tran, T. T. Nguyen, N. M. Pham, P. H. N. Ly, and H. H. Luong, "Strawberry disease identification with vision transformer-based models," *Multimedia Tools and Applications*, Feb. 2024.

[18] M. Schwegler, C. Müller, and A. Reiterer, "Integrated gradients for feature assessment in point cloud-based data sets," *Algorithms*, vol. 16, no. 7, p. 316, 2023.

[19] M. Bontonou, A. Haget, M. Boulougouri, J.-M. Arbona, B. Audit, and P. Borgnat, "Studying limits of explainability by integrated gradients for gene expression models," *arXiv preprint arXiv:2303.11336*, 2023.

[20] M. Jarke and F. J. Radermacher, "The ai potential of model management and its central role in decision support," *Decision Support Systems*, vol. 4, no. 4, pp. 387–404, 1988.

[21] M. E. Sahin, "Image processing and machine learning-based bone fracture detection and classification using x-ray images," *International Journal of Imaging Systems and Technology*, vol. 33, no. 3, pp. 853–865, 2023.

[22] F. Hardalaç, F. Uysal, O. Peker, M. Çiçeklidağ, T. Tolunay, N. Tokgöz, U. Kutbay, B. Demirciler, and F. Mert, "Fracture detection in wrist x-ray images using deep learning-based object detection models," *Sensors*, vol. 22, no. 3, p. 1285, 2022.

[23] S. Verma, S. Kulshrestha, C. Rajput, and S. Patel, "Detecting bone fracture using transfer learning," *Advancement of Machine Intelligence in Interactive Medical Image Analysis*, pp. 215–228, 2020.

[24] M. A. Kassem, S. M. Naguib, H. M. Hamza, M. M. Fouda, M. K. Saleh, K. M. Hosny *et al.*, "Explainable transfer learning-based deep learning model for pelvis fracture detection," *International Journal of Intelligent Systems*, vol. 2023, 2023.

[25] J. Ying, H. Wang, J. Liu, T. Yu, and D. Huang. (2023) Harnessing resnet50 and senet for enhanced ankle fracture identification.

[26] Z. Alammar, L. Alzubaidi, J. Zhang, Y. Li, W. Lafta, and Y. Gu, "Deep transfer learning with enhanced feature fusion for detection of abnormalities in x-ray images," *Cancers*, vol. 15, no. 15, p. 4007, 2023.

[27] A. Bhan, S. Singh, S. Vats, and A. Mehra, "Ensemble model based osteoporosis detection in musculoskeletal radiographs," in *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2023, pp. 523–528.

[28] H. T. Nguyen, H. H. Luong, T. H. N. Kien, N. T. L. Phan, T. M. Dang, T. T. Duong, T. D. Nguyen, and T. C. Dinh, "Brain tumors detection on mri images with k-means clustering and residual networks," *Advances in Computational Collective Intelligence*, pp. 317–329, 2022.

[29] H. H. Luong, N. T. L. Phan, T. C. Dinh, T. M. Dang, T. T. Duong, T. D. Nguyen, and H. T. Nguyen, "Fine-tuning mobilenet for breast cancer diagnosis," *Inventive Computation and Information Technologies*, pp. 841–856, 2023.

[30] H. T. Nguyen, H. H. Luong, P. T. Phan, H. H. D. Nguyen, D. Ly, D. M. Phan, and T. T. Do, "Hs-unet-id: An approach for human skin classification integrating between unet and improved dense convolutional network," *International Journal of Imaging Systems and Technology*, vol. 32, no. 6, pp. 1832–1845, Jun. 2022.

[31] L. H. Huong, N. H. Khang, L. N. Quynh, L. H. Thang, D. M. Canh, and H. P. Sang, "A proposed approach for monkeypox classification," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023.

[32] H. P. Nguyen, T. P. Hoang, and H. H. Nguyen, "A deep learning based fracture detection in arm bone x-ray images," in *2021 international conference on multimedia analysis and pattern recognition (MAPR)*. IEEE, 2021, pp. 1–6.

[33] J. A. Wani and N. Sharma, "Comparative analysis of transfer learning models in classification of histopathological whole slide images," in *Proceedings of International Conference on Recent Innovations in Computing: ICRIC 2022, Volume 1*. Springer, 2023, pp. 351–369.

[1] https://www.kaggle.com/datasets/gauravduttakiit/fracturefusion-a-symphony-of-bone-breaks

[2] https://github.com/lhhuong/FEC-IGE

[34] S. Research, "Science research 2022: Bone fracture detection dataset," dec 2022, visited on 2024-03-01. [Online]. Available: https://universe.roboflow.com/science-research/science-research-2022:-bone-fracture-detection

[35] X. Xu, M. Wang, D. Liu, M. Lei, and X. Cheng, "Sternal fracture recognition based on efficientnetv2 fusion spatial and channel features," in *The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery*. Springer, 2022, pp. 191–200.

[36] H. Min, Y. Rabi, A. Wadhawan, P. Bourgeat, J. Dowling, J. White, A. Tchernegovski, B. Formanek, M. Schuetz, G. Mitchell *et al.*, "Automatic classification of distal radius fracture using a two-stage ensemble deep learning framework," *Physical and Engineering Sciences in Medicine*, pp. 1–10, 2023.

[37] L. Tanzi, A. Audisio, G. Cirrincione, A. Aprato, and E. Vezzetti, "Vision transformer for femur fracture classification," *Injury*, vol. 53, no. 7, pp. 2625–2634, 2022.

[38] X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," in *2017 2nd international conference on image, vision and computing (ICIVC)*. IEEE, 2017, pp. 783–787.