Article

Context-Aware Level-Wise Feature Fusion Network with Anomaly Focus for Precise Classification of Incomplete Atypical Femoral Fractures in X-Ray Images

Joonho Chang [1], **Junwon Lee** [1], **Doyoung Kwon** [1], **Jin-Han Lee** [2], **Minho Lee** [1], **Sungmoon Jeong** [3], **Joon-Woo Kim** [2], **Heechul Jung** [1,*] and **Chang-Wug Oh** [2,*]

[1] Department of Artificial Intelligence, Kyungpook National University, Daegu 41566, Republic of Korea; joonnoz@knu.ac.kr (J.C.); zini9405@knu.ac.kr (J.L.); entropy0437@knu.ac.kr (D.K.); mholee@knu.ac.kr (M.L.)
[2] Department of Orthopedic Surgery, School of Medicine, Kyungpook National University Hospital, Daegu 41566, Republic of Korea; jhlee.knuh@gmail.com (J.-H.L.); joonwoo@knu.ac.kr (J.-W.K.)
[3] Department of Medical Informatics, School of Medicine, Kyungpook National University, Daegu 41566, Republic of Korea; jeongsm@knu.ac.kr
* Correspondence: heechul@knu.ac.kr (H.J.); cwoh@knu.ac.kr (C.-W.O.)

Abstract: Incomplete Atypical Femoral Fracture (IAFF) is a precursor to Atypical Femoral Fracture (AFF). If untreated, it progresses to a complete fracture, increasing mortality risk. However, due to their small and ambiguous features, IAFFs are often misdiagnosed even by specialists. In this paper, we propose a novel approach for accurately classifying IAFFs in X-ray images across various radiographic views. We design a Dual Context-aware Complementary Extractor (DCCE) to capture both the overall femur characteristics and IAFF details with the surrounding context, minimizing information loss. We also develop a Level-wise Perspective-preserving Fusion Network (LPFN) that preserves the perspective of features while integrating them at different levels to enhance model representation and sensitivity by learning complex correlations and features that are difficult to obtain independently. Additionally, we incorporate the Spatial Anomaly Focus Enhancer (SAFE) to emphasize anomalous regions, preventing the model bias toward normal regions, and reducing False Negatives and missed IAFFs. Experimental results show significant improvements across all evaluation metrics, demonstrating high reliability in terms of accuracy (0.931), F1-score (0.9456), and AUROC (0.9692), proving the model's potential for application in real medical settings.

Keywords: Incomplete Atypical Femoral Fracture; Atypical Femoral Fracture; X-ray; feature fusion; anomaly focus; tiny lesion; image classification; deep learning

MSC: 92C55

1. Introduction

Atypical Femoral Fracture (AFF) is a dangerous fracture type that occurs in the subtrochanteric or diaphyseal regions of the femur [1]. It can develop with slight or without any injury and is characterized by radiographic findings such as a simple transverse or short oblique fracture [2]. Several factors have been identified as contributing to the occurrence of AFF, including excessive femoral curvature [3], vitamin D deficiency [4], and the use of proton pump inhibitors (PPIs) and corticosteroids [5]. Notably, a strong correlation has been observed between AFF and the prolonged use of bisphosphonates (BPs) and denosumab for osteoporosis treatment [6–8], so AFF predominantly occurs in the elderly population. If an AFF occurs, surgical intervention becomes more challenging and can lead to functional impairment as well as various complications such as nonunion and femoral head necrosis [9], leading to an increased mortality risk following AFF [10,11]. For these reasons, with the increasing aging population, the incidence of AFF is also expected to rise,

emphasizing the importance of early diagnosis and intervention before AFF occurrence to mitigate these risks.

In the early stages preceding the occurrence of AFF, cortical buckling will develop in the lateral cortex of the femur due to repeated cycles of microfracture and healing [12]. This condition is termed Incomplete Atypical Femoral Fracture (IAFF). IAFF exhibits various characteristics and is classified based on its location as shown in Figure 1: Diaphyseal IAFF (D-IAFF), which occurs in the femoral shaft, and Subtrochanteric IAFF (S-IAFF), which occurs in the subtrochanteric region [13]. Although IAFF is a crucial precursor to AFF, it is often asymptomatic or presents with vague features, making detection difficult and often resulting in delayed diagnosis. As a precursor to AFF, the progression process from IAFF to AFF is illustrated in Figure 2. IAFF is typically diagnosed through bone scans [14] or Magnetic Resonance Imaging (MRI) [15]. However, these diagnostic methods have notable drawbacks, including high costs and time consumption. Furthermore, there remains the risk of misdiagnosis [14], which can lead to either unnecessary or delayed interventions, ultimately culminating in a complete fracture.

(a) Diaphyseal IAFF (D-IAFF)　　　　　　　　(b) Subtrochanteric IAFF (S-IAFF)
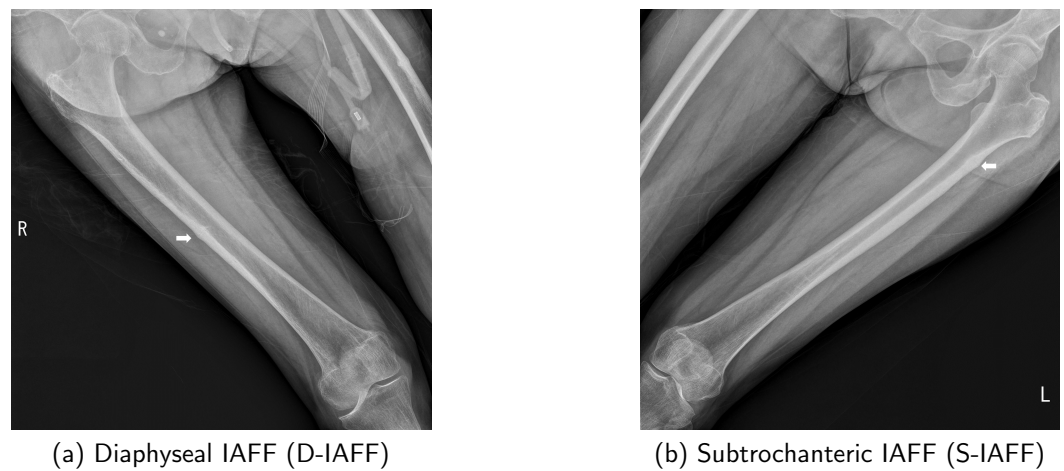
Figure 1. Types of Incomplete Atypical Femoral Fracture (IAFF) categorized by their anatomical location of occurrence. The arrow indicates the location of the IAFF.

(a) 8 months　　　　(b) 3 months　　　　(c) 1 week　　　　(d) AFF
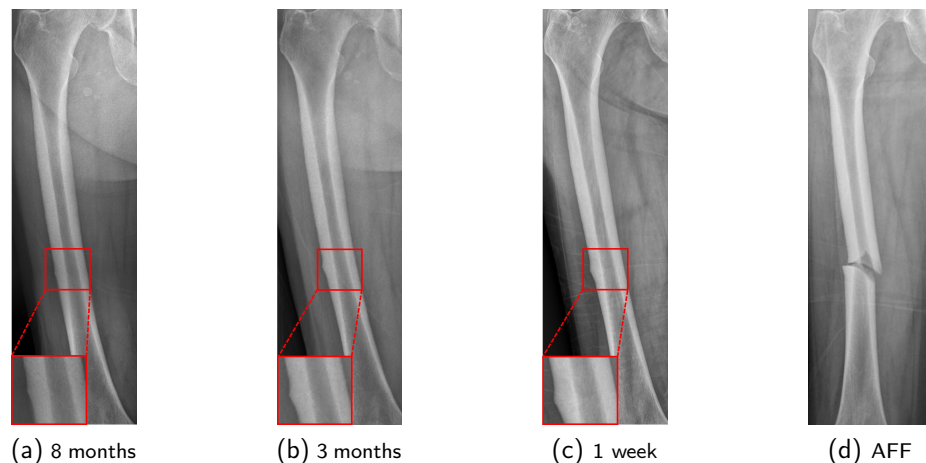
Figure 2. Progression process from IAFF to AFF. This image is an AP radiographic view of a case missed at the hospital due to limited information on IAFF: (**a**) 8 months prior to AFF occurrence, (**b**) 3 months prior to AFF occurrence, (**c**) 1 week prior to AFF occurrence, (**d**) AFF occurrence.

To mitigate these problems, the development of a diagnostic support system utilizing X-rays is essential, but there are several obstacles to be overcome: (1) An IAFF is often extremely small in size, (2) lacks distinct characteristics, and (3) is similar to normal anatomical deformations, making it easy to overlook. Additionally, (4) the location and

features of an IAFF vary depending on the type, and (5) its appearance may differ slightly depending on the radiographic view, even for the same patient (Figure 3). Due to these factors, even experienced orthopedic specialists may miss an IAFF if they do not examine it meticulously.
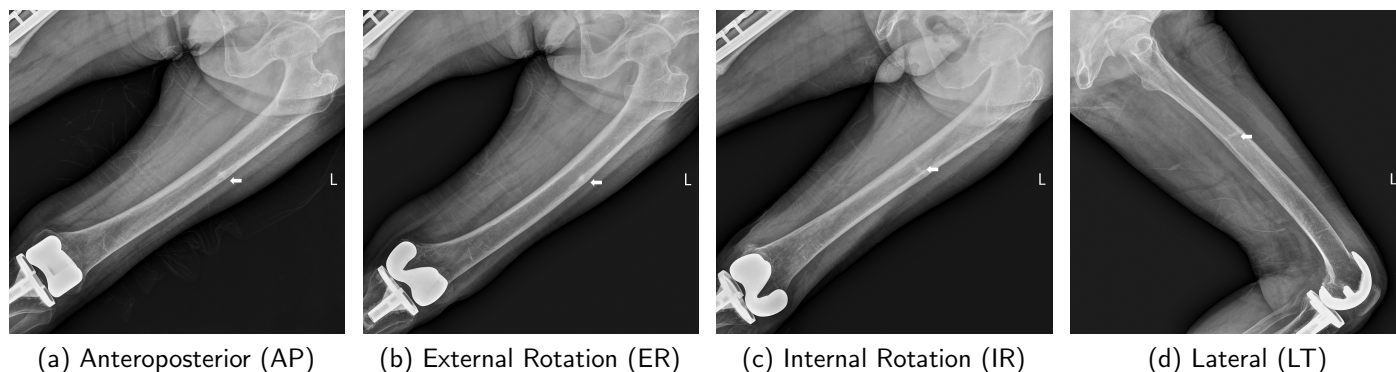
Figure 3. Examples of various X-ray radiographic views used for the detection of IAFF. The arrow indicates the location of the IAFF.

To address these challenges, we propose a universal model, Context-aware Level-wise Feature Fusion Network with Anomaly Focus (CFNet), inspired by the diagnostic methods employed by orthopedic specialists for IAFF diagnosis. This model is designed to be seamlessly integrated into various classification frameworks. When diagnosing IAFF, specialists typically begin by assessing the overall shape, curvature, and suspicious regions on a femur X-ray scan. They then carefully examine areas where IAFFs frequently occur and make the final diagnosis by comparing these regions with other potential conditions or deformities across all regions. Based on this diagnostic approach, CFNet extracts features at multiple levels from both a single entire X-ray image and high-resolution images segmented into four sections, utilizing Dual Context-aware Complementary Extractor (DCCE) blocks within each input branch of DCCE. This design allows the model to capture the overall femoral features from the entire X-ray image while simultaneously identifying tiny and ambiguous IAFF characteristics from the high-resolution sliced images, thus providing a mutual complement. The features extracted from each DCCE block are fused by the Level-wise Perspective-preserving Fusion Network (LPFN) to minimize information loss and ensure that features at the same level are integrated without interference. LPFN enhances the model's representational capacity by learning features and correlations that are difficult to capture independently, thereby improving prediction accuracy. Moreover, we incorporate a Spatial Anomaly Focus Enhancer (SAFE) to focus on IAFF features and allow the model to comprehensively capture the correlation between the IAFF, its surrounding information, and overall image information. This approach prevents the model from overfitting and ensures effective learning of the scarce and subtle IAFF information. With these components, CFNet provides a highly accurate solution capable of detecting even tiny and ambiguous IAFFs while precisely distinguishing subtle differences. In our experiments, the proposed model demonstrates significant performance improvements over existing models, with each component effectively minimizing missed IAFFs and achieving accurate classification. Our main contributions are as follows:

- We propose a novel model inspired by the diagnostic approach employed by specialists to enhance the classification performance of tiny and ambiguous IAFFs. To the best of our knowledge, this is the first model capable of effectively identifying all known types of IAFF features, regardless of prior surgical history or the presence of pathological fractures, while minimizing False Negatives. Furthermore, we are the first to utilize all major radiographic views (AP, ER, IR, LT) ensuring high accuracy in recognizing IAFFs across various imaging conditions.

- We introduce the DCCE to overcome the challenges of information loss in small fractures and the limited contextual understanding encountered in conventional classification models. DCCE comprises two branches: one branch captures the overall characteristics of the femur and identifies potential IAFF regions across the entire X-ray image, while the other understands and focuses on IAFF features with surrounding details in high-resolution images, thereby extracting complementary information.
- We propose the LPFN to effectively learn the subtle features of small and ambiguous IAFFs and prevent the misclassification of noise and artifacts as IAFFs by leveraging complementary information. This approach preserves the unique meaning and perspective of the extracted features, integrating them without interference across different levels. By doing so, the model can utilize information from multiple levels and learn complex features and correlations that are difficult to capture independently.
- We incorporate SAFE to minimize missed IAFFs and mitigate model bias toward regions unrelated to IAFFs. This approach captures comprehensive contextual information and long-range dependencies within the input, addressing the limitations of traditional Convolutional Neural Networks (CNNs) [16] and emphasizing anomalous regions. Consequently, it ensures that even subtle differences are not overlooked while preventing the model from overfitting to normal regions and backgrounds.

The remainder of this paper is organized as follows. Section 2 reviews related works, and Section 3 provides a detailed description of our proposed model. Section 4 presents the dataset utilized for the experiments, experiment details, evaluation metrics, and experimental results. Finally, the discussion and conclusion are provided in Sections 5 and 6, respectively.

2. Related Works

2.1. AFF and IAFF Classification and Detection

Recent advancements in AI-based medical image processing have shown considerable potential in aiding doctors, and various studies have been conducted on its applications. However, there has been very little research conducted on IAFF and AFF. To date, there are only two deep learning-based studies on AFF and a single study on IAFF. The studies on AFF aim to classify AFF and Normal Femur Fractures (NFFs) to ensure appropriate treatment decisions. Ref. [17] compared two approaches: an automated method using downsampling alone and an intervention method that focuses on fracture features by manually cropping the fracture region. The effectiveness of these approaches was evaluated using VGG [18], ResNet [19], and an Inception Network [20]. Ref. [21] adopted a multimodal approach by integrating X-ray data with Electronic Health Records (EHR) data containing AFF risk factors. They explored the impact of probability fusion, feature fusion, and learned feature fusion methods [22,23] on the features extracted from each datum.

Ref. [12] conducted the only study focused on IAFF, employing an ensemble method [24] to classify D-IAFF and normal cases. They manually cropped IAFF regions from the X-ray images and applied a Sobel filter [25] to extract edge information. Subsequently, they utilized pretrained DenseNet [26], EfficientNet [27], and MobileNet [28] for the ensemble method and achieved promising results. However, they only addressed D-IAFF with relatively distinct features in the AP view, which may have led to model bias, and it remained uncertain as to whether this approach is applicable to other radiographic views or different types of IAFF. Moreover, the manual crop processing is labor intensive and time-consuming. Additionally, all of the aforementioned studies excluded patients with a history of surgeries, such as implants or joint replacements, as well as those with pathological fracture traces. Given that AFF and IAFF predominantly occur in elderly patients, many who require prediagnosis may have undergone prior surgeries or experienced fractures. Therefore, these models face significant limitations in their practical use in real medical settings.

2.2. Bone Fracture Classification

Numerous automated classification models have been proposed to enhance the reliability of bone fracture classification. In machine learning-based approaches, Ref. [29] utilized Radiomics [30] to extract and analyze mathematical and statistical information from X-ray images using techniques such as Recursive Feature Elimination (RFE), Sequential Forward Selection (SFS), Least Absolute Shrinkage and Selection Operator (LASSO), and Ridge methods. They then compared the performance of these feature selection methods using XGBoost (XGB) [31], LightGBM [32], and Logistic Regression [33]. Ref. [34] employed various edge detection methods to extract bone information and classify different types of fractures and normal cases in X-ray images using Random Forest [35] and AdaBoost [36]. Ref. [37] adopted k-Nearest Neighbors (kNN) [38], Naive Bayes (NB) [39] and Support Vector Machine (SVM) [40] to evaluate the effectiveness of different machine learning methods for classifying femoral neck fracture in pelvic X-ray images. Additionally, Ref. [41] integrated machine learning with deep learning methods to classify tibial and fibula fractures. They extracted features from X-ray images using CNNs and applied the ensemble method with the results from machine learning models to obtain the final outcome. While these machine learning approaches demonstrate utility for identifying simple patterns, they often struggle with noise and data complexity, resulting in insufficient accuracy for practical application in clinical settings.

In recent years, deep learning-based models have achieved significant advancements in bone fracture classification, demonstrating promising performance. LSNet [42] proposed a lightweight CNN architecture combined with machine learning algorithms to classify wrist fractures in X-ray images. They utilized a SqueezeNet-based [43] Siamese network [44] for feature extraction and a machine learning classifier for the final decision. SFNet [45] employed an enhanced Canny edge detection algorithm [46] to localize fracture regions and applied multiscale feature fusion, combining both the localized results and original X-ray images to generate richer features. Ref. [47] introduced a curriculum learning [48] approach that integrates medical knowledge for proximal femur fracture classification in X-ray images. They assigned difficulty levels to each training sample based on discrepancies between medical guidelines and expert annotations. Subsequently, they enhanced performance by using a CNN to gradually increase the difficulty of training examples during training. Ref. [49] applied meta-learning with paired image and text data to classify femur fractures in pelvic X-ray images. They proposed an algorithm using an encoder–decoder structure to learn features from both image and radiology report modalities, thereby compensating for information that might be missing from the image alone. Ref. [50] adopted a domain adaptation approach that leveraged information and features extracted from CT images to enhance the classification of Osteoporotic Vertebral Fracture (OVF) in X-ray images. They employed a feature-level mix-up module to reduce domain discrepancies between CT and X-ray images, enabling the model to focus on semantically consistent features such as outlines and textures across both domains. Compared to conventional machine learning-based methods, most of these models achieved good performance. However, these studies and approaches do not consider and learn the characteristics of small and ambiguous fractures, limiting their practical applicability in IAFF classification.

2.3. Tiny Lesion Classification and Detection

In medical imaging, low-resolution and contrast problems can hinder the accurate identification of tiny lesions, potentially leading to overlooked lesion details due to inaccurate information at the pixel level. In addition, these lesions may resemble the surrounding tissues, structures, or noise in terms of shape and brightness, making it challenging to detect and distinguish them. In an attempt to address these problems, several models have been proposed. Ref. [51] proposed a two-stage model for detecting microaneurysms in retinal images and microcalcifications in mammograms. The first stage adopted a deep cascade decision tree [52] to eliminate certain background information and mitigate the class imbalance between small lesions and the backgrounds. In the second stage, a CNN was

employed to extract and learn features from the remaining portion of the image. Ref. [53] cropped input images into patches of varying sizes to capture local features of tiny lesions. They then utilized a multiscale network in combination with an ensemble method to improve detection accuracy. PESA R-CNN [54] proposed a model that utilizes perihematomal edema (PHE) [55] information to reduce the risk of missing tiny and low-contrast brain hemorrhages. They developed a semi-supervised Center Surround Difference U-Net to segment PHE and hemorrhage regions, generating an expanded Region of Interest (RoI) based on these segmentations. They then gathered information through networks of different depths corresponding to the RoI size and integrated with equal weight for the final result. Ref. [56] employed complementary networks to detect and classify Focal Cerebra Ischemia (FCI) and Lacunar Infarction (LACI). They used a U-Net-based [57] primary network to detect tiny lesions, and the results were further refined by a correction network to identify lesion types.

In studies focused on fractures, Fracture R-CNN [58] integrated clinical diagnostic knowledge into Faster R-CNN [59] to reduce missed tiny skull fractures. They identified potential fracture regions and selected appropriate anchors for varying fracture sizes using an adaptive anchoring Region Proposal Network (RPN). Additionally, they improved detection accuracy by removing skull suture interference with an anti-interference head (A-Head) module. Ref. [60] proposed a two-stage network to classify triquetral avulsion fractures and Segond fractures. They identified RoIs using a Deep Convolutional Neural Network (DCNN), automatically cropped these regions, and subsequently classified them into different fracture types using pretrained models. Ref. [61] developed an R-CNN-based model for detecting arm fractures. They adopted a backbone network with a feature pyramid [62] structure to extract multiscale features, which were subsequently integrated to generate RoIs. To accommodate tiny fractures, they constrained the minimum RoI size and expanded it to ensure that small fractures were captured. Ref. [63] introduced a three-step algorithm for classifying small rib fractures in CT scans. First, they used a U-Net to segment bones and remove vertebrae and scapula based on their shape and location information, which interfered with fracture detection. Subsequently, they employed 3D DenseNet combined with an inception mechanism to capture multiscale features. These studies are optimized for specific tasks and image types, such as rib fracture classification in CT images or microaneurysms detection in retinal images, limiting their generalizability and performance when applied to other tasks. Therefore, it is crucial to develop a model optimized for X-ray-based IAFF diagnosis to minimize patient risk and provide reliable support to clinicians in real medical settings.
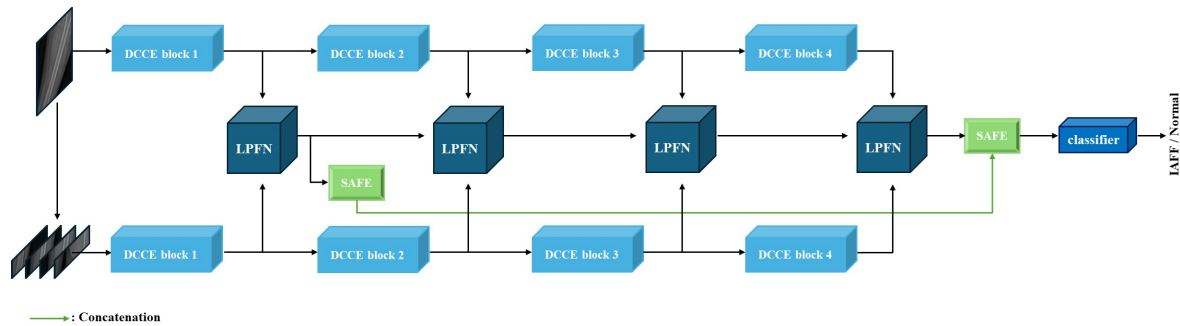
3. Proposed Model

3.1. Model Overview

The proposed CFNet is a universal approach that can be applied to a wide range of classification models, inspired by the diagnostic methods used by specialists for identifying IAFF. The classification performance of the proposed model is majorly enhanced by three key components, DCEE, LPFN, and SAFE, as illustrated in Figure 4.

CFNet consists of two input branches. Each branch of DCCE is composed of a feature extractor selected from classification models and is divided into four blocks based on the level of features being extracted. The first input branch processes the entire X-ray image, while the second input branch sequentially receives four high-resolution X-ray slices of equal size, divided from the top to the bottom of the original image. Features at various levels are then extracted through each block.

These features are integrated at different levels by our LPFN. It fuses the feature maps extracted from the corresponding blocks of each branch, enabling the model to learn richer features and correlations that are difficult to obtain individually. The fused results are then further combined with the output of the subsequent LPFN. Consequently, the classifier utilizes information that has been aggregated from all LPFN outputs. This approach ensures

that features from various levels complement each other without interference, making them effectively utilized for accurate predictions.

Figure 4. The overview of our proposed CFNet model. It is mainly enhanced by DCCE, LPFN, and SAFE. In the two input branches, DCCE extracts the complementary overall and detailed features of the X-ray images at different feature levels. The features extracted from DCCE are then fused at different levels in the LPFN, preserving unique perspectives. Additionally, SAFE is applied to the results of the first and fourth LPFNs to learn the correlations between positional information within the features and emphasize anomalous regions.

Despite these advancements, there remains a possibility of missing tiny IAFFs, and the high proportion of normal regions and background information may limit the model's ability to learn the characteristics of IAFF. To address this problem, we incorporate SAFE into the results of the first and last LPFN to emphasize anomalous regions and comprehensively learn the relationships among positional information. The results from the two SAFE modules are then combined and fed into the classifier of the selected model to generate the final classification result. By adopting and integrating these modules, the proposed model prevents misclassification and minimizes the risk of missing tiny and ambiguous IAFFs, thereby achieving high classification performance.

3.2. Dual Context-Aware Complementary Extractor (DCCE)

In medical image classification, many models either rely on a single entire image or crop it into small patches for training. However, since IAFFs are tiny and lack distinct features, relying solely on a single entire image can lead to a significant loss or even a complete disappearance of IAFF information as the model deepens. Furthermore, the patch-based approach only utilizes information from a highly limited region, making it unable to utilize surrounding contextual information. Additionally, the severe class imbalance between normal and IAFF patches often leads to a model bias toward the majority class, hindering the accurate learning of IAFF characteristics. To address these limitations and effectively extract features across different feature levels, we propose the DCCE.

DCCE serves as the feature extractor of CFNet and is compatible with various classification models. It comprises two branches. The first branch processes the entire femoral X-ray to extract overall features, which emulates the approach specialists use when reviewing the entire femur. This branch captures the overall structure of the femur, including key IAFF characteristics such as curvature. It also identifies suspicious regions and distributions of IAFF, enabling the model to leverage location-based features. The second branch sequentially processes data that have been divided into four equal segments from top to bottom, based on the height of the entire image. This approach preserves high resolution, enabling a detailed analysis of IAFF boundaries and patterns. Unlike patch-based approaches, this strategy enables the utilization of surrounding information related to the IAFF, preserving contextual information and allowing the model to capture IAFF details more accurately through comprehensive analysis. This branch simulates how specialists zoom in and identify areas where IAFFs frequently occur, thereby capturing additional information about tiny IAFFs that may be missed or insufficiently addressed in the first branch.

In addition, each DCCE branch is organized into four blocks based on the level of features being extracted. To define these blocks, we first divide the total number of layers in the selected feature extractor into four equal groups. Within these divided groups, each DCCE block is defined using the nearest unit block or stage of the feature extractor as a division point. The first DCCE block focuses on extracting low-level features such as the edges and textures of the femur. The second and third blocks capture the shape and structural characteristics of the femur, extracting intermediate-level features, and the last block extracts high-level features, including contextual and semantic information.

DCCE overcomes the challenges of information loss and limited contextual understanding found in conventional methods by extracting both the overall features of the input image and IAFF characteristics in conjunction with surrounding context through two separate branches. This approach ensures that the model captures even ambiguous and small critical features without overlooking them. Additionally, the four DCCE block structures enable level-wise feature extraction at multiple levels and perspectives, allowing the model to accurately distinguish between IAFFs, anatomical deformities, and normal regions while recognizing subtle differences. Consequently, DCCE provides complementary and rich information to LPFN and SAFE, significantly enhancing the accuracy of IAFF detection.

3.3. Level-Wise Perspective-Preserving Fusion Network (LPFN)

The features extracted from different inputs each contain unique meanings and information. If these features are integrated randomly without considering their respective levels, their inherent meanings and correlations may be distorted, leading to the loss of useful patterns and relationships that the model could learn. It can hinder the effective utilization of key information extracted from each branch, such as the femur's overall structural characteristics, IAFF positional information, and detailed IAFF features. Therefore, it is essential to integrate the features appropriately according to their levels.
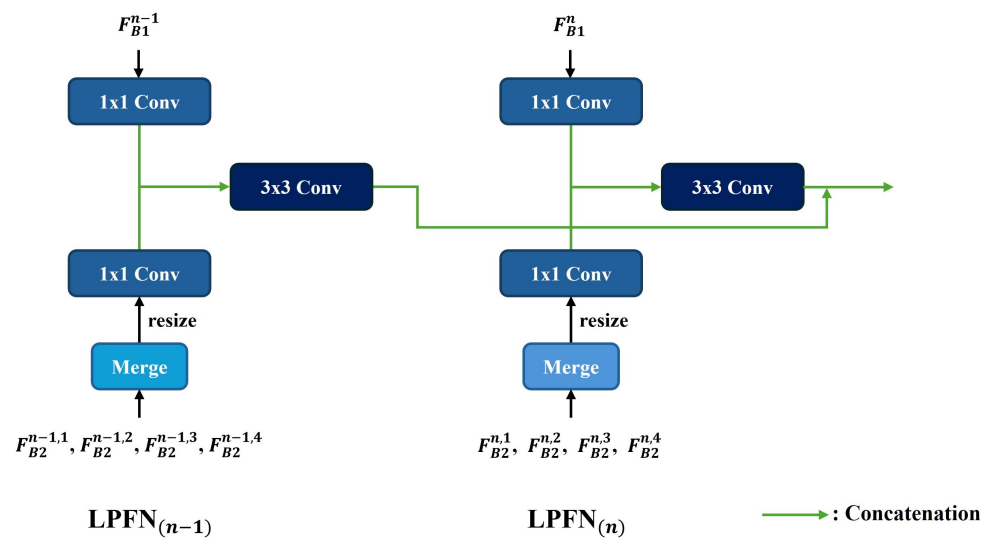
To mitigate this challenge, we propose the Level-wise Perspective-preserving Fusion Network (LPFN), which integrates features extracted from the DCCE at different levels without interference. LPFN employs 1×1 convolution to align the channel dimensions of feature maps and preserve essential information from the outputs of the DCCE blocks extracted in the same sequence from each branch. For the second branch, which processes four inputs per data sample, the feature maps are sequentially merged according to their order and then consolidated into a single feature map. Subsequently, this result is resized to match the dimensions of the feature map from the first branch. To minimize information loss and preserve the unique perspectives of each feature, we fuse the two feature maps through concatenation. This approach expands the dimensionality of the feature map, allowing the model to learn from more comprehensive information. The resultant feature map is then processed with a 3×3 convolutional layer, followed by Batch Normalization [64], which ensures the stable learning of complex relationships and features that would be difficult to capture independently from each branch's results alone. Moreover, this process enhances the model's representation capabilities and enables the comprehensive use of information from different levels. The result from each LPFN is then fused with the subsequent LPFN results, and ultimately, the classifier utilizes the information fused across all levels of LPFN results. The training procedure of the LPFN is illustrated in Figure 5, and the result of the nth LPFN is represented in Equation (1):

$$B_1^n = \mathrm{Conv}_{1 \times 1}\left(F_{\mathrm{B1}}^n\right),$$

$$B_2^n = \mathrm{Conv}_{1 \times 1}\left(\mathrm{Resize}\left(\mathrm{Merge}\left(F_{\mathrm{B2}}^{n,1}, F_{\mathrm{B2}}^{n,2}, F_{\mathrm{B2}}^{n,3}, F_{\mathrm{B2}}^{n,4}\right)\right)\right) \tag{1}$$

$$LPFN_{(n)} = \left(\mathrm{Conv}_{3 \times 3}\left(B_1^n \| B_2^n\right)\right) \| LPFN_{(n-1)}$$

F_{B1}^n represents the feature map extracted from the first branch of the nth DCCE block, while $F_{\mathrm{B2}}^{n,i}$ denotes the feature map extracted from the ith slice (where i = 1, 2, 3, 4) in the second

branch of the nth DCCE block. B_1^n and B_2^n refer to the feature maps from the first and second branches after processing for fusion, respectively. $\|$ represents concatenation, and $LPFN_{(n)}$ is the output of the nth LPFN.

Figure 5. Training procedure of the $n-1$th and nth Level-wise Perspective-preserving Fusion Network (LPFN). The feature maps extracted from the DCCE blocks of each branch are fused at different levels without interference, followed by the extraction of complex relationships and features using a 3×3 convolution. The result is then fused into the subsequent LPFN, and this process is consistently applied to all LPFNs.

LPFN operates analogously to how specialists compare and analyze both the overall femur information and the details of regions where IAFFs frequently occur. By comprehending the overall characteristics of the image, LPFN aids in reducing the misinterpretation of noise and artifacts as IAFFs and provides positional information that indicates a higher probability of IAFF presence. Based on this guide, the model accurately learns the fine details of small and ambiguous IAFF from high-resolution information, ultimately enhancing prediction accuracy and sensitivity through the utilization of complementary information.

3.4. Spatial Anomaly Focus Enhancer (SAFE)

IAFF features are often ambiguous and resemble typical anatomical deformations, making them easy to overlook. Additionally, due to their small size, the model may be biased toward normal regions and backgrounds. In such cases, the model may fail to accurately understand the unique characteristics of IAFF, leading to an increased rate of False Negatives where IAFF is misclassified as a normal case. Therefore, it is crucial to emphasize IAFF features and understand their relationships with the surrounding information. While CNN models are adept at learning local patterns through convolutional kernels, they are limited in capturing long-range dependencies and broader contextual information. To overcome this limitation, we propose SAFE, a self-attention mechanism-based [65] approach.

SAFE treats each position of $LPFN_{(n)}$ as a unique vector and learns the relationships between positions. To achieve this, spatial information is consolidated into a single dimension, converting the feature map into a sequence format (Equation (2)). A linear projection (Equation (3)) is then applied to derive three vectors: query (Q), key (K), and value (V) (Equation (4)). This reflects the input information directly, preserves the relationships between positional information, and flexibly transforms the dimensions based on data complexity and feature characteristics. Subsequently, SAFE is computed as shown in Equation (5):

$$LPFN_{(n)seq} = Flatten(LPFN_{(n)}) \tag{2}$$

$$linear(A) = XA^T + b \tag{3}$$

$$Q = XA_q^T + b_q, \quad K = XA_k^T + b_k, \quad V = XA_v^T + b_v \tag{4}$$

$$SAFE(Q, K, V) = LPFN_{(n)seq} + linear\left(\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V\right) \tag{5}$$

In Equation (3), X represents the input vector or matrix, A is the weight matrix, and b denotes the bias vector. In Equation (5), d_k refers to the dimensions of K, Q represents the feature vector of a specific location in the image, and K contains information about all other locations. The similarity between Q and K is computed using the dot product to identify their interrelationships, assigning higher weights to more important positions. Subsequently, by applying the Softmax function [66] with a 'soft-assignment' approach, the values in the output vector are transformed into probabilities ranging from 0 to 1 with a total sum of 1. This prevents problems of divergence or convergence to 0, enabling the model to comprehensively learn relationships across multiple regions. V represents the actual information at each location, and by weighting V according to this probability distribution, more important information is emphasized with higher weights, allowing the model to focus on important features. This result is combined with the input sequence ($LPFN_{(n)seq}$) after linear transformation to generate the final outcome ($SAFE(Q, K, V)$), enabling efficient computation without distortion of the similarity calculation results.

This approach preserves the original input information and enables the model to comprehensively understand the relationship between the input image context, surrounding information of IAFF, and IAFF itself by learning the correlations between each position in the image and all other positions. In addition, by assigning higher weights to anomalous regions and emphasizing IAFF, it prevents the model from overfitting to regions unrelated to IAFF. As a result, CFNet effectively focuses more on relevant regions and enhances the model's ability to accurately detect even subtle differences.

Loss

We employ the Cross-Entropy loss function (L_{CE}) to effectively model the mutually exclusive probability distributions between the IAFF and normal classes. The L_{CE} measures the difference between the model's predicted probability distribution and the target probability distribution, evaluating how closely the model's output aligns with the target distribution. The equation is as follows:

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(p_i) \tag{6}$$

where C indicates the number of classes, y_i denotes the target label, and p_i represents the probability predicted by the model for class i. This loss function ensures that the predicted distribution aligns with the target distribution during training, thereby enhancing classification accuracy.

4. Experiments

4.1. Dataset

The University Hospital (UH) dataset was collected at Kyungpook National University Hospital (KNUH) between August 2010 and November 2022. It comprises 794 X-ray images from 236 patients, including 430 images from 92 patients with IAFF and 364 images from 144 patients in the normal group. The IAFF cases are further categorized into D-IAFF and S-IAFF, and three orthopedic specialists reviewed and classified the data into normal or IAFF types. The dataset comprises images of both the left and right femurs, obtained from different radiographic views, including Anteroposterior (AP), External Rotation (ER), Internal Rotation (IR), and Lateral (LT). To ensure data independence for each patient, reflecting real clinical settings, the training and evaluation sets are split by patient. Among the 794 collected images, 666 images (IAFF: 354, Normal: 312) were randomly selected for

5-fold cross-validation training, and the remaining 128 images (IAFF: 76, Normal: 52) were used for evaluation. This dataset was approved by the KNUH Institutional Review Board under approval number KNUH202402007-HE001 on 26 February 2024.

4.2. Data Preprocessing

To enhance the data processing efficiency and optimize memory usage, we converted Digital Imaging and Communication in Medicine (DICOM) files into Numerical Python (NumPy) format. The data were then preprocessed using two different methods (Figure 6). (1) Crop (C): To reduce unnecessary information in the image, such as left/right markers and knee implants, we cropped the images to a size of 2200 $\times$ 2200, corresponding to the smallest data dimension. To preserve S-IAFF characteristics, only the bottom portion of the images was cropped for height, while both sides were symmetrically cropped for width. (2) Automated Extraction and Alignment (AEA): We employed a segmentation model to generate femur masks from the original X-ray images. After evaluating several models, U-Net++ was selected for its superior performance. To further refine the masks, the connectivity of pixel values in the generated mask was computed to eliminate noise, and the inlier set was extracted using the RANdom SAmple Consensus (RANSAC) algorithm [67]. The Hough transform [68] was then applied to determine the rotation angle for vertically aligning the mask. Additionally, the histogram was analyzed to exclude the knee and pelvis regions, ensuring that only the RoI of the femur was extracted from the mask. After preprocessing, the (C) images were resized to 1024 $\times$ 1024, and the (AEA) images were resized to 1024 $\times$ 256 using bilinear interpolation. Finally, the pixel values were normalized to the range [0, 1].

(a) Original   (b) Crop   (c) AEA
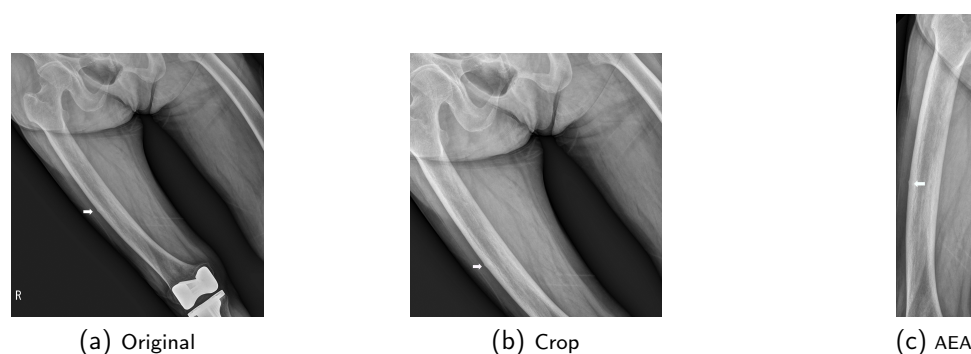
Figure 6. Examples of original X-ray image and preprocessed results. (**a**) Original input image. (**b**) Crop preprocessing result shows that the knee implant, markers, and some noise have been removed from the original image. (**c**) AEA preprocessing result indicates that all unnecessary information has been eliminated, retaining only the femur region essential for training. The arrow indicates the location of the IAFF.

4.3. Training Details

We conducted experiments using a 5-fold cross-validation approach to evaluate generalization performance and ensure an accurate comparison of results. For the experiments, We utilized models pretrained on ImageNet [69], fine-tuning each model without freezing any layers for 300 epochs with a batch size of 16 for each fold. We employed a Stochastic Gradient Descent (SGD) optimizer [70], with a learning rate set to 0.1×10^{-4} and a momentum of 0.9. To enhance model robustness, flipping and rotation augmentation were applied. All methods were implemented in PyTorch, and experiments were conducted on a single NVIDIA RTX A6000 GPU (48 GB).

4.4. Evaluation Metrics

We evaluate the performance of the proposed model using several widely recognized classification metrics: accuracy, F1-score, AUROC, AUPRC, precision, recall, and specificity. These metrics provide a comprehensive assessment of the model's ability to classify IAFF

and normal cases. In these metrics, True Positive (TP) refers to cases where the model correctly predicts positive (IAFF) samples, True Negative (TN) represents cases where the model correctly identifies negative (normal) samples, False Positive (FP) indicates instances where the model incorrectly predicts negative samples as positive, and False Negative (FN) denotes cases where positive samples are misclassified as negative. All metrics range from 0 to 1, with values closer to 1 indicating higher performance.

Accuracy: This metric evaluates the overall correctness of the model's predictions by calculating the ratio of correctly predicted cases (both TP and TN) to the total number of cases:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

F1 score: The *F1 score* is the harmonic mean of precision and recall, providing a balance between these two metrics. It is particularly useful when precision and recall are in a trade-off. A value closer to 1 suggests that both precision and recall are high, while a value closer to 0 indicates a deficiency in one or both metrics:

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{8}$$

AUROC: *AUROC* represents the area under the Receiver Operating Characteristic (ROC) curve, where the x-axis plots the False Positive Rate (FPR), and the y-axis plots the True Positive Rate (TPR). This metric evaluates the model's ability to distinguish between positive and negative classes across various thresholds. A higher *AUROC* indicates that the model can effectively reduce False Positives while maximizing True Positives at various threshold levels.

AUPRC: *AUPRC* measures the area under the Precision–Recall (PR) curve, where precision is plotted on the y-axis and recall on the x-axis. This metric focuses on the model's ability to identify positive samples, excluding negative class performance. A higher *AUPRC* represents a model's effectiveness in identifying positive cases while minimizing False Positives. An *AUROC* or *AUPRC* value of 1 indicates an ideal model, signifying perfect classification, while 0.5 suggests performance equivalent to random guessing. Values below 0.5 indicate the model performs worse than random chance, reflecting a tendency to misclassify instances.

Precision: *Precision* measures the proportion of True Positive cases among all cases predicted as positive, highlighting the significance of False Positives. A high precision value implies fewer False Positives:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Recall: *Recall* measures the proportion of True Positive cases among all actual positive instances, focusing on minimizing False Negatives. A higher recall suggests that the model is less likely to miss positive cases:

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

Specificity: *Specificity* quantifies the model's ability to correctly identify negative cases, with a focus on minimizing False Positives. A high specificity indicates that the model effectively avoids misclassifying negative samples as positive:

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

4.5. Comparison of Classification Performance with Other State-of-the-Art Models

We evaluate the performance and suitability of the proposed model for IAFF classification by comparing it with state-of-the-art models and also utilizing them as our baseline models. As shown in Table 1, baseline models with crop preprocessing, ResNet-50, a widely used CNN model, exhibit relatively poor performance across all metrics, failing to accurately capture the features of small and ambiguous IAFFs. The latest model, MobileNetV4 [71], is highly lightweight, making it challenging to adequately capture small and ambiguous IAFF features. As a result, it shows very low performance, similar to ResNet. Although DenseNet-121, EfficientNet-B2, ConvNeXt V2 [72], and RDNet [73] show improved performance over traditional CNN models such as ResNet-50 and GoogLeNet [20], their results are still insufficient. Vision Transformer-based [74] models, such as RepViT [75] and FastViT [76], despite their recent success and remarkable performance on various tasks, show similar or even inferior performance compared to traditional CNN models due to the limited amount of data. In contrast, EdgeNeXt [77], FocalNet [78], and VGG16 outperform other models, with VGG16 achieving the highest performance among the state-of-the-art models. However, these baseline models face challenges in distinguishing noise, artifacts, and common deformations from IAFFs, resulting in high recall but lower precision and specificity performance.

Even with AEA preprocessing, ResNet-50 and MobileNetV4 still show relatively lower performance compared to others. In contrast, EfficientNet-B1, EfficientNet-B3, and FastViT, which perform poorly with crop preprocessing, show improved performance and achieve results comparable to or surpassing DenseNet-121 and EfficientNet-B2. ConvNeXt V2 and RDNet also demonstrate enhanced performance, attaining results comparable to EdgeNeXt, FocalNet, and VGG16. Notably, VGG16 achieves the highest performance among all models, even with AEA preprocessing. While AEA preprocessing reduces noise and artifacts, leading to a relatively higher precision due to decreased misclassifications, these models still struggle to capture IAFF characteristics accurately. Consequently, they tend to misclassify IAFFs as normal cases or confuse regular deformations and normal regions with IAFFs, leading to lower recall and specificity.

When comparing the effects of different preprocessing methods, it is observed that most baseline models, with the exception of EfficientNet-B2, EdgeNeXt, and FocalNet, achieve approximately 4% to 6% improvement in accuracy and F1-score with AEA preprocessing compared to the crop method. This finding suggests that excessive information from normal regions and backgrounds can hinder the model's ability to learn IAFF characteristics. Moreover, the three aforementioned models also demonstrate enhanced performance in distinguishing between IAFF and normal cases, resulting in increased AUROC and AUPRC, as well as higher precision and specificity performance compared to crop preprocessing results. Despite these improvements, these models still tend to misclassify small and ambiguous IAFFs as normal cases, resulting in relatively lower recall performance.

Table 1. Comparison of classification performance of the proposed approach and state-of-the-art methods.

Model	Method		Accuracy	F1 Score	AUROC	AUPRC	Precision	Recall	Specificity
ResNet-50	C	Baseline	0.6719	0.7145	0.7123	0.7842	0.7395	0.6921	0.6423
		Proposed	0.7703	0.8201	0.8012	0.8102	0.7706	0.8816	0.6077
	AEA	Baseline	0.7313	0.7765	0.7656	0.7974	0.7709	0.7868	0.6500
		Proposed	0.8276	0.8627	0.8158	0.9362	0.8664	0.8606	0.7708
GoogLeNet	C	Baseline	0.7500	0.7909	0.7987	0.8410	0.7924	0.8026	0.6731
		Proposed	0.7844	0.8235	0.8028	0.8311	0.8010	0.8474	0.6923
	AEA	Baseline	0.7891	0.8191	0.8561	0.8884	0.8343	0.8089	0.7615
		Proposed	0.8262	0.8609	0.8272	0.8357	0.8214	0.9046	0.7115

Table 1. *Cont.*

Model	Method		Accuracy	F1 Score	AUROC	AUPRC	Precision	Recall	Specificity
DenseNet-121	C	Baseline	0.7922	0.8284	0.8457	0.8670	0.8147	0.8447	0.7154
		Proposed	0.8231	0.8547	0.8751	0.8965	0.8493	0.8640	0.6731
	AEA	Baseline	0.7938	0.8139	0.8766	0.9148	0.8764	0.7605	0.8423
		Proposed	0.8644	0.8946	0.9027	0.9444	0.8953	0.8982	0.8063
EfficientNet-B1	C	Baseline	0.7672	0.8102	0.7879	0.7971	0.7841	0.8395	0.6615
		Proposed	0.7872	0.8346	0.8152	0.8590	0.7963	0.8772	0.6454
	AEA	Baseline	0.8156	0.8419	0.8693	0.9046	0.8589	0.8263	0.8000
		Proposed	0.8359	0.8615	0.8994	0.9253	0.8670	0.8586	0.8029
EfficientNet-B2	C	Baseline	0.8234	0.8570	0.8584	0.8771	0.8257	0.8921	0.7231
		Proposed	0.8448	0.8839	0.8895	0.9200	0.8373	0.9364	0.6875
	AEA	Baseline	0.8047	0.8287	0.8657	0.9018	0.8644	0.7974	0.8154
		Proposed	0.8216	0.8537	0.8733	0.9063	0.8598	0.8503	0.7697
EfficientNet-B3	C	Baseline	0.7625	0.8083	0.8115	0.8338	0.7760	0.8447	0.6423
		Proposed	0.7969	0.8395	0.8232	0.8414	0.7907	0.8947	0.6538
	AEA	Baseline	0.8203	0.8525	0.8840	0.9096	0.8328	0.8737	0.7423
		Proposed	0.8483	0.8794	0.9170	0.9384	0.8642	0.8955	0.7692
ConvNeXt V2	C	Baseline	0.8000	0.8408	0.8418	0.8773	0.8035	0.8842	0.6769
		Proposed	0.8281	0.8599	0.8883	0.9160	0.8317	0.8904	0.7372
	AEA	Baseline	0.8563	0.8800	0.8999	0.9295	0.8756	0.8868	0.8115
		Proposed	0.8750	0.8967	0.9168	0.9381	0.8817	0.9123	0.8205
EdgeNeXt	C	Baseline	0.8625	0.8876	0.9065	0.9237	0.8655	0.9132	0.7885
		Proposed	0.8908	0.9131	0.9280	0.9452	0.8913	0.9376	0.8189
	AEA	Baseline	0.8625	0.8840	0.9230	0.9465	0.8875	0.8816	0.8346
		Proposed	0.9035	0.9241	0.9403	0.9591	0.9201	0.9309	0.8563
FocalNet	C	Baseline	0.8610	0.8853	0.8957	0.9025	0.8649	0.9079	0.7923
		Proposed	0.8869	0.9121	0.9091	0.9278	0.8846	0.9422	0.7952
	AEA	Baseline	0.8594	0.8817	0.9086	0.9311	0.8778	0.8868	0.8192
		Proposed	0.8908	0.9127	0.9264	0.9572	0.9177	0.9091	0.8594
MobileNetV4	C	Baseline	0.7141	0.7463	0.7614	0.8098	0.7902	0.7079	0.7231
		Proposed	0.7813	0.8204	0.8030	0.8304	0.7999	0.8421	0.6923
	AEA	Baseline	0.7438	0.7929	0.7687	0.8111	0.7654	0.8263	0.6231
		Proposed	0.8125	0.8501	0.8234	0.8374	0.8098	0.8947	0.6923
RDNet	C	Baseline	0.8391	0.8658	0.8898	0.9124	0.8561	0.8763	0.7846
		Proposed	0.8750	0.8959	0.9251	0.9390	0.8849	0.9079	0.8269
	AEA	Baseline	0.8594	0.8824	0.9169	0.9460	0.8802	0.8868	0.8192
		Proposed	0.8906	0.9054	0.9241	0.9540	0.9306	0.8816	0.9039
FastViT	C	Baseline	0.7672	0.8136	0.8084	0.8449	0.7778	0.8553	0.6385
		Proposed	0.7928	0.8394	0.8215	0.8607	0.9001	0.8841	0.6410
	AEA	Baseline	0.8359	0.8636	0.9058	0.9329	0.8549	0.8737	0.7808
		Proposed	0.8730	0.8941	0.8300	0.9489	0.8881	0.9013	0.8317
RepViT	C	Baseline	0.7188	0.7635	0.7664	0.8068	0.7626	0.7684	0.6462
		Proposed	0.7422	0.7891	0.7743	0.8210	0.7679	0.8132	0.6385
	AEA	Baseline	0.7578	0.8000	0.7619	0.7765	0.7848	0.8158	0.6731
		Proposed	0.7852	0.8234	0.7874	0.8328	0.8126	0.8421	0.7019
VGG16	C	Baseline	0.8797	0.8990	0.9348	0.9521	0.8959	0.9026	0.8462
		Proposed	**0.9167**	**0.9346**	**0.9584**	**0.9745**	**0.9284**	**0.9409**	**0.8750**
	AEA	Baseline	0.9000	0.9146	0.9565	0.9718	0.9246	0.9079	0.8885
		Proposed	**0.9310**	**0.9456**	**0.9694**	**0.9854**	**0.9455**	**0.9455**	**0.9063**

By integrating our proposed model with these baselines, we observed notable improvements across all metrics and models, regardless of the preprocessing method. Notably, ResNet-50 shows significant improvements, with accuracy and F1-score increasing by approximately 9% with crop preprocessing and 6% with AEA preprocessing. Specifically, when integrating CFNet with VGG16 and applying AEA preprocessing, it outperforms all other models across every metric, achieving accuracy, F1-score, AUROC (Figure 7), and AUPRC scores of 0.931, 0.9456, 0.9694, and 0.9854, respectively. Additionally, the proposed method demonstrates a significant improvement in recall performance across all models compared to the baselines, showing its effectiveness in mitigating the False Negative problem. Furthermore, even when applying crop preprocessing, the proposed method accurately learns and classifies IAFFs without bias, despite the abundance of normal regions and background information. The results of this experiment validate that the proposed method is applicable to various models and significantly enhances performance, thereby demonstrating the superiority and suitability of our approach for IAFF classification.
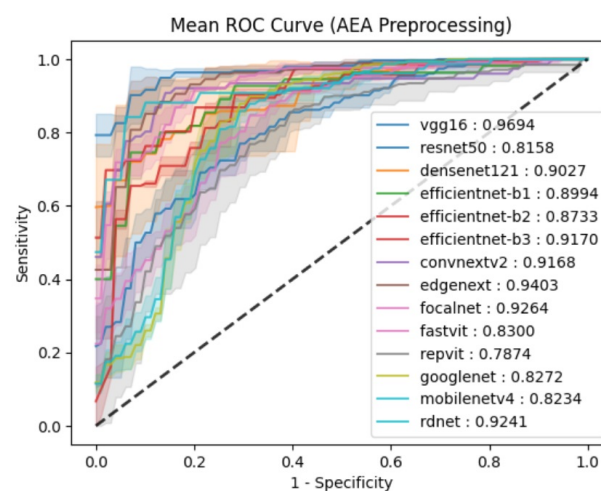
Figure 7. Comparison of AUROC results under AEA preprocessing.

4.6. Experiments on DCCE and LPFN Group Configurations

We evaluate the performance of the CFNet across different configurations of the DCCE and LPFN groups, utilizing VGG16 as the baseline due to its superior performance in previous experiments. In accordance with our proposed methodology, we first count the total number of layers and divide them into n (where n = 3, 4, 5) equal groups. Subsequently, each DCCE block is defined based on the nearest unit block or stage as a division point. As shown in Table 2, the 3-group configuration exhibits lower performance across all metrics for both preprocessing methods compared to the 4-group and 5-group configurations, primarily due to the reduced extraction and learning of feature details in the 3-group setup. Nevertheless, due to the effectiveness of the SAFE component, recall shows excellent performance. The 5-group configuration extracts and integrates features at a more detailed level, utilizing relatively more information compared to the 4-group configuration. Consequently, while it exhibits slightly higher performance in accuracy and F1-score, it shows lower performance in AUROC and AUPRC, indicating a lack of robust across all thresholds. Furthermore, although the increased sensitivity to IAFF in the 5-group configuration enhances precision and recall performance slightly, it also leads to a rise in False Positives, thereby reducing specificity, as the model misclassifies more normal cases as IAFF. Ultimately, while the 5-group configuration demonstrates comparable or better performance than the 4-group setup, the increased training time and resource consumption renders it less practical. Therefore, for a more balanced and practical approach, we select the 4-group configuration for DCCE and LPFN in our final model.

Table 2. Performance comparison based on the configuration of DCCE and LPFN groups.

# of Groups	Accuracy	F1-Score	AUROC	AUPRC	Precision	Recall	Specificity
			Crop preprocessing				
3	0.9080	0.9275	0.9472	0.9654	0.9242	0.9309	0.8687
4	0.9167	0.9346	0.9584	0.9745	0.9284	0.9409	0.8750
5	0.9177	0.9350	0.9563	0.9710	0.9229	0.9478	0.8659
			AEA preprocessing				
3	0.9230	0.9404	0.9604	0.9776	0.9406	0.9418	0.8938
4	0.9310	0.9456	0.9692	0.9854	0.9455	0.9455	0.9063
5	0.9354	0.9513	0.9627	0.9834	0.9432	0.9600	0.9000

4.7. Performance Comparison Based on the Number of Input Slices

We evaluate the performance based on the number of image slices fed into the second branch. Following the approach utilized in our proposed method, we divided the image into n slices (where $n = 3, 4, 5$) from top to bottom based on the height of the image prior to inputting them into the second branch. We utilize VGG16 as the baseline, with all conditions kept the same as the proposed model, except for the number of input slices. As shown in Table 3, when using 3 slices, the model can capture a relatively broad range of context and IAFF characteristics. However, due to the lower resolution of each slice than $n = 4$ or 5, it struggles to capture the fine details of IAFF, resulting in slightly lower overall performance compared to the 4-slice configuration. Moreover, when using 5 slices, the performance further decreases, showing similar or even lower results compared to the 3-slice configuration. This decline in performance is likely due to the height of the sliced images being too short relative to their width, limiting the model's ability to utilize contextual information and features surrounding the IAFF in the femur. As a result, the understanding of the relationship between the IAFF features and the surrounding information diminishes, making it difficult to distinguish IAFFs from other deformities or normal regions, which in turn leads to a decline in performance. Based on these experimental results, we conclude that the optimal configuration for our model is to use 4 DCCE and LPFN groups, along with 4 input slices for the second branch.

Table 3. Performance comparison based on the number of slices input into the second branch of the DCCE.

# of Slices	Accuracy	F1-Score	AUROC	AUPRC	Precision	Recall	Specificity
			Crop preprocessing				
3	0.9119	0.9307	0.9471	0.9659	0.9231	0.9394	0.8646
4	0.9167	0.9346	0.9584	0.9745	0.9284	0.9409	0.8750
5	0.9057	0.9259	0.9426	0.9642	0.9183	0.9345	0.8564
			AEA preprocessing				
3	0.9253	0.9413	0.9637	0.9821	0.9379	0.9452	0.8906
4	0.9310	0.9456	0.9692	0.9854	0.9455	0.9455	0.9063
5	0.9218	0.9384	0.9613	0.9807	0.9351	0.9418	0.8875

4.8. Classification Performance Analysis Using Confusion Matrix

To implement the IAFF classification model in real medical settings, achieving high accuracy is essential; however, it is equally important to conduct a thorough analysis of False Positives (Type I error) and False Negatives (Type II error). In particular, Type II errors where an actual IAFF is missed can have severe consequences, as patients may not receive timely treatment. Therefore, we evaluate these errors and the overall performance of the model using a confusion matrix. VGG16 is set as the baseline, and its results are compared with CFNet. As shown in Table 4, VGG16 with crop preprocessing achieves a

True Positive count of 69 and a True Negative count of 44, indicating overall acceptable performance. However, with 8 False Positives and 7 False Negatives, it relatively frequently misses IAFFs or misclassifies normal cases as IAFFs, especially in data with small and ambiguous IAFF features. This occurs because crop preprocessing includes unnecessary information outside the femur region, causing the model to misinterpret noise and artifacts in normal data as IAFF or to overlook IAFF features due to the distraction of irrelevant details. However, CFNet shows improved performance over the baseline model, with 6 False Positives and 4 False Negatives. As shown in Table 5, using AEA preprocessing, VGG16 reduces False Positives to 6, but the False Negatives remain unchanged compared to Table 4. This result indicates that using only the femur region as input reduces extraneous information, leading to fewer misclassifications. However, small and ambiguous IAFFs continued to be missed, resulting in a high False Negative rate. In contrast, CFNet improves overall performance by leveraging complementary, rich features and understanding the surrounding context. This allows it to better distinguish IAFF from other information in the input image. Additionally, SAFE emphasizes anomalous regions, enhancing the identification of these areas and minimizing both False Negatives and False Positives. As a result, CFNet based on VGG16 achieves the best performance with 5 False Positives and 3 False Negatives, showcasing exceptional accuracy and reliability. These results highlight the potential of CFNet for practical application in real medical settings.

Table 4. Comparison of the confusion matrix between VGG16 and CFNet under crop preprocessing. Bold highlights the best performance.

Crop Preprocessing		VGG16		CFNet	
		Predicted class			
		P	N	P	N
Actual class	P	69	7	**72**	**4**
	N	8	44	**6**	**46**

Table 5. Comparison of the confusion matrix between VGG16 and CFNet under AEA preprocessing. Bold highlights the best performance.

AEA Preprocessing		VGG16		CFNet	
		Predicted class			
		P	N	P	N
Actual class	P	69	7	**73**	**3**
	N	6	46	**5**	**47**

We also analyze the cases in which False Negatives and False Positives occur in CFNet. Typically, IAFF manifests in the lateral cortex of the femur and exhibits characteristics of cortical buckling. However, in the LT view, IAFF may present with entirely different features, such as fine line patterns, or in some instances, no noticeable features at all, as illustrated in Figure 8a. When no discernible features are present, specialists rely on information from other radiographic views to make a diagnosis. While the model accurately classifies data with different features, False Negatives are observed in cases where no features are visible in the LT direction. Moreover, the femur is susceptible to a variety of deformities and diseases beyond IAFF, some of which display characteristics nearly identical to IAFF as shown in Figure 8b. It is observed that False Positives occur in such cases, where the model mistakenly identifies these deformities as IAFF.

(a) FN　　　　　　　　　　　　　　　　　　　　　(b) FP

Figure 8. Example of False Negative (FN) and False Positive (FP). FN occurs in image where no discernible IAFF features are present, while FP appears in the case of femoral deformities that closely resemble IAFF features. The arrow indicates the location of the IAFF.

4.9. Analysis of Classification Performance and Robustness Across Different X-Ray Radiographic Views

The shapes and characteristics of IAFFs vary depending on the radiographic view, with certain views showing nearly absent distinguishing features, making it challenging for the model to learn, and leading to potential misclassifications. In this experiment, we evaluate the classification performance and robustness of the model across various radiographic views. We use VGG16 as the baseline model and compare the performance using accuracy and F1-score metrics. As shown in Table 6, the baseline results with crop preprocessing achieve relatively high performance in AP view, where IAFF features are more distinguishable from normal regions. However, the model shows lower performance in the LT view, where the distinguishing features are either insufficient or less prominent. With AEA preprocessing, the removal of irrelevant information allows the model to focus on essential features, leading to an overall improvement in the baseline performance. Nonetheless, performance in the LT view remains lower compared to other views due to the fact that the LT view exhibits completely different characteristics. Additionally, the limited data make it difficult for the baseline model to learn these features effectively. In contrast, the proposed method significantly improves performance across all radiographic views. With crop preprocessing, the proposed method improves accuracy by approximately 6.6% and F1-score by 4.1% in the ER view, where the baseline performance is initially low. In the LT view, the proposed method achieves notable improvements, with a 5.4% increase in accuracy and a 6.6% increase in F1-score. Similarly, AEA preprocessing with the proposed method also demonstrates overall performance enhancements, with the LT view accuracy increasing by 5.4% and F1-score by 5.7%. These experimental results validate that the proposed method effectively captures even insufficient and ambiguous features, as well as very subtle differences. It also demonstrates robustness and high performance across all radiographic views regardless of the preprocessing method, highlighting the model's suitability and superiority.

Table 6. Comparison of classification performance across different X-ray radiographic views. Bold highlights the best performance.

Model	AP		ER		IR		LT	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
Crop preprocessing								
Baseline	0.9317	0.9349	0.8383	0.8959	0.9130	0.9418	0.8357	0.8241
CFNet	**0.9561**	**0.9593**	**0.9043**	**0.9373**	**0.9217**	**0.9517**	**0.8893**	**0.8900**
AEA preprocessing								
Baseline	0.9366	0.9339	0.9217	0.9481	0.8956	0.9317	0.8537	0.8511
CFNet	**0.9610**	**0.9657**	**0.9391**	**0.9583**	**0.9217**	**0.9504**	**0.9073**	**0.9082**

4.10. Performance Analysis Based on Parameter and Execution Time

We apply our proposed CFNet to baseline models to evaluate performance in terms of the number of parameters, as well as execution (training and inference) times. For comparison, we select ResNet-50 as a representative CNN model, VGG16 as a high-performance model, MobileNetV4 as the latest lightweight model, and FastViT as a Vision Transformer-based model. All models are compared under AEA preprocessing conditions. The training time is measured over 300 epochs, including both training and validation phases, while inference time is measured based on processing a total of 128 test samples. As shown in Table 7, the ResNet-50-based model requires a relatively large number of parameters and shows longer execution time, while demonstrating the lowest performance. The MobileNetV4-based model demonstrates the fewest parameters and shortest execution time, making it suitable for deployment in medical devices. However, its low parameter count limits the model representation, reducing its ability to accurately identify small and ambiguous IAFF features, resulting in low performance similar to ResNet-50-based CFNet. In contrast, FastViT achieves relatively high performance with roughly half the parameters of the ResNet-50-based model. Notably, VGG16-based model demonstrated even more robust and superior performance, with a parameter count and execution time comparable to FastViT. Additionally, its short inference time allows for prompt assistance to medical professionals, underscoring its suitability for real clinical application.

Table 7. Performance comparison based on the number of parameters and execution time.

Base Model	# of Parameters	Training Times	Test Times	Accuracy	F1-Score	AUROC	AUPRC
ResNet-50	51,084,928	5 h 4 m	3 m 3 s	0.8276	0.8627	0.8847	0.9362
MobileNetV4	19,753,219	2 h 54 m	2 m 20 s	0.8125	0.8501	0.8234	0.8374
FastViT	28,471,936	4 h 1 m	2 m 32 s	0.8730	0.8941	0.8275	0.9489
VGG16	30,173,147	4 h 12 m	2 m 36 s	0.9310	0.9456	0.9692	0.9854

4.11. Ablation Study

We evaluate the contribution of each component of CFNet to the overall classification performance. Table 8 presents the results of our ablation study based on VGG16, which achieved the highest performance in previous experiments. As shown in the results, each of our novel components provides significant performance enhancement to the baseline model. While the baseline model shows decent performance with both preprocessing methods, it struggles with noise, artifacts, and deformations, often misclassifying them as IAFF and failing to capture fine details, resulting in lower precision and specificity. By adding DCCE and LPFN to the baseline, the model enables to learn both the overall characteristics of the femur and IAFF features in conjunction with the surrounding contextual information. This integration reduces errors in misclassifying normal regions as IAFF, leading to an improvement in precision. However, False Negatives persist, resulting in relatively low recall performance. When applying SAFE to the baseline, the model effectively learns long-range dependencies and relationships between IAFF features and broader contextual information, leading to notable improvements in AUROC and AUPRC. Additionally, SAFE emphasizes anomalous regions and captures tiny IAFFs, reducing False Negative rates and improving recall performance. However, False Positives still affect the model's specificity performance. In contrast, our proposed model, which integrates the strengths of each component, accurately learns IAFF characteristics while minimizing the misclassification of normal cases as IAFFs. Consequently, the proposed method achieves high performance across precision, recall, and specificity. Furthermore, the proposed approach clearly distinguishes between IAFF and normal cases, resulting in superior AUROC and AUPRC performance results, achieving the highest performance across all metrics. These findings demonstrate that our proposed method offers high reliability in IAFF classification and holds significant potential for practical application in real clinical settings.

Table 8. Ablation Study to analyze the impact of each component of the proposed method on overall performance. Bold highlights the best performance.

Model	Accuracy	F1-Score	AUROC	AUPRC	Precision	Recall	Specificity
			Crop preprocessing				
Baseline	0.8797	0.8990	0.9348	0.9521	0.8959	0.9026	0.8462
Baseline + DCCE + LPFN	0.9004	0.9214	0.9415	0.9631	0.9227	0.9212	0.8646
Baseline + SAFE	0.9080	0.9278	0.9475	0.9667	0.9200	0.9364	0.8594
CFNet	**0.9167**	**0.9346**	**0.9584**	**0.9745**	**0.9284**	**0.9409**	**0.8750**
			AEA preprocessing				
Baseline	0.9000	0.9146	0.9565	0.9718	0.9246	0.9079	0.8885
Baseline + DCCE + LPFN	0.9234	0.9391	0.9644	0.9827	0.9451	0.9227	0.9063
Baseline + SAFE	0.9224	0.9385	0.9651	0.9834	0.9410	0.9364	0.8984
CFNet	**0.9310**	**0.9456**	**0.9692**	**0.9854**	**0.9455**	**0.9455**	**0.9063**

5. Discussion

In this study, we experimentally demonstrate that applying the IAFF diagnostic method commonly used by specialists to CFNet significantly enhances overall performance. We also show that integrating DCCE with LPFN, along with applying SAFE, is highly effective for classifying tiny and ambiguous IAFFs. CFNet can extract features from femoral X-ray images using various state-of-the-art models. In this process, using only a single entire image may result in the loss of critical details of tiny IAFFs, while relying exclusively on small patches can hinder accurate learning due to the restricted context information. To address these challenges, we propose DCCE, which consists of two branches to capture both overall features and fine-grained details along with the surrounding context of IAFFs. This approach mirrors the clinical process, where specialists first review the entire X-ray image and then focus on regions where IAFFs commonly occur. Additionally, to integrate features extracted from DCCE at different levels while preserving their unique significance, we propose LPFN. LPFN prevents the misclassification of noise and artifacts as IAFF, enhances sensitivity by capturing detailed information from high-resolution images, and provides complementary information. LPFN operates similarly to how clinicians analyze and compare overall structures and frequently occurring regions. However, since IAFF features are often ambiguous and difficult to distinguish from deformations, there remains a risk of missing IAFF, and their small size can lead to model bias towards normal and background regions. To address this, we incorporate SAFE, which captures long-range dependencies that are challenging for CNNs alone and helps understand spatial relationships more effectively. Furthermore, SAFE assigns higher weights to abnormal regions, emphasizing IAFF features and preventing the model from overfitting to normal regions.

Despite the promising results, there are still limitations and room for improvement. Experimental results show that in cases where IAFF features are either entirely absent such as in some LT images or in a very early stage, the model encounters difficulty accurately classifying IAFF. As a result, there is a risk of IAFF going unnoticed by both patients and medical professionals, potentially delaying treatment. This could lead to progression to a complete fracture or the significant worsening of IAFF, making treatment more complex and challenging. To prevent this problem, based on previous studies that indicate a close association between IAFF and conditions such as osteoporosis and femoral curvature, we plan to collect relevant meta-data and modify CFNet to support multimodal learning alongside image data, further improving classification accuracy in all cases. Additionally, the proposed method has demonstrated significantly improved performance with limited data. However, to ensure the generalizability of our approach, more data are required. Currently, there are no publicly or privately available IAFF datasets, so we are actively collecting additional data from KNUH to validate the generalization performance of the proposed method. Additionally, to demonstrate the model's robustness through external validation, we are collecting data from other university hospitals with varying X-ray

conditions. The goal of the proposed method is to apply it in real medical settings. We will validate the model's generalizability and robustness through the collection of additional data and further experiments.

6. Conclusions

In this article, we propose CFNet, a novel approach to accurately classify even tiny and ambiguous IAFFs without missing any. In our model, DCCE recognizes contextual information and extracts complementary information on both the overall femur features and detailed regions at multiple levels, addressing problems of information loss and limited contextual understanding. By introducing LPFN, our model preserves the unique meaning of features at each level, enabling seamless feature fusion without interference. This design supports the stable learning of complex relationships and enhances classification accuracy and sensitivity. In addition, SAFE comprehensively captures spatial dependencies and emphasizes anomalous regions, minimizing missed IAFFs and preventing overfitting to normal regions. Experimental results demonstrate that each component of CFNet contributes meaningfully to the model's improved overall performance.

For future work, the current approach will divide the selected model into four equal segments based on the number of layers and assign DCCE blocks according to adjacent unit blocks or stages. We plan to enhance the DCCE module by incorporating techniques such as receptive field analysis or filter response evaluation. These methods will allow for a more precise assessment of feature map levels extracted at each layer and enable a more adaptive assignment of DCCE blocks tailored to each model. Additionally, since the femur can present with various diseases, deformities, and fractures that may mimic IAFF characteristics, we aim to incorporate a Large Language Model (LLM) [79] to leverage extensive domain knowledge, enabling CFNet to analyze predictions and offer insights into potential conditions beyond IAFF. This approach is expected to increase the model's reliability, making it a valuable tool in clinical settings for accurate patient risk assessment and for guiding preventive measures.

Author Contributions: Conceptualization, J.C., M.L., H.J. and C.-W.O.; Methodology, J.C. and J.L.; Software, J.C., J.L. and D.K.; Validation, J.C. and J.L.; Formal analysis, J.C., D.K. and H.J.; Investigation, J.C.; Resources, J.-H.L., J.-W.K. and C.-W.O.; Data curation, J.-H.L. and C.-W.O.; Writing—original draft, J.C. and H.J.; Writing—review & editing, J.C., M.L. and H.J.; Visualization, J.C. and H.J.; Supervision, M.L., S.J., H.J. and C.-W.O.; Project administration, M.L., S.J., H.J. and C.-W.O.; Funding acquisition, M.L. and H.J. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: This study was approved by the Kyungpook National University Hospital (KNUH) Institutional Review Board under approval number KNUH202402007-HE001 on 26 February 2024.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets presented in this article are not readily available due to the inclusion of personally identifiable information. Ethical and legal constraints imposed by our institutional review board prevent us from sharing the raw data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Larsen, M.S.; Schmal, H. The enigma of atypical femoral fractures: A summary of current knowledge. *EFORT Open Rev.* **2018**, *3*, 494–500. [CrossRef] [PubMed]
2. Shane, E.; Burr, D.; Abrahamsen, B.; Adler, R.A.; Brown, T.D.; Cheung, A.M.; Cosman, F.; Curtis, J.R.; Dell, R.; Dempster, D.W.; et al. Atypical subtrochanteric and diaphyseal femoral fractures: Second report of a task force of the American Society for Bone and Mineral Research. *J. Bone Miner. Res.* **2014**, *29*, 1–23. [CrossRef] [PubMed]
3. Ikeda, S.; Sakai, A.; Tanaka, H.; Takeuchi, Y.; Ohnishi, H.; Murakami, H.; Saito, M.; Ito, M.; Nakamura, T. Atypical femoral fractures were associated with deterioration of bone quality and curvature of femoral shaft. *Orthop. Proc. Bone Jt.* **2013**, *95*, 67.
4. Velasco, S.; Kim, S.; Bleakney, R.; Jamal, S.A. The clinical characteristics of patients with hip fractures in typical locations and atypical femoral fractures. *Arch. Osteoporos.* **2014**, *9*, 171. [CrossRef] [PubMed]
5. Targownik, L.E.; Leslie, W.D. The relationship among proton pump inhibitors, bone disease and fracture. *Expert Opin. Drug Saf.* **2011**, *10*, 901–912. [CrossRef]
6. Girgis, C.M.; Sher, D.; Seibel, M.J. Atypical femoral fractures and bisphosphonate use. *N. Engl. J. Med.* **2010**, *362*, 1848–1849. [CrossRef]
7. Adler, R.A.; El-Hajj Fuleihan, G.; Bauer, D.C.; Camacho, P.M.; Clarke, B.L.; Clines, G.A.; Compston, J.E.; Drake, M.T.; Edwards, B.J.; Favus, M.J.; et al. Managing osteoporosis in patients on long-term bisphosphonate treatment: Report of a task force of the American Society for Bone and Mineral Research. *J. Bone Miner. Res.* **2016**, *31*, 16–35. [CrossRef]
8. Selga, J.; Nuñez, J.; Minguell, J.; Lalanza, M.; Garrido, M. Simultaneous bilateral atypical femoral fracture in a patient receiving denosumab: Case report and literature review. *Osteoporos. Int.* **2016**, *27*, 827–832. [CrossRef]
9. Pearce, O.; Edwards, T.; Al-Hourani, K.; Kelly, M.; Riddick, A. Evaluation and management of atypical femoral fractures: An update of current knowledge. *Eur. J. Orthop. Surg. Traumatol.* **2021**, *31*, 825–840. [CrossRef]
10. Charoenngam, N.; Rittiphairoj, T.; Jaroenlapnopparat, A.; Mettler, S.K.; Ponvilawan, B.; Okoli, U.; Ungprasert, P.; Marangoz, M.S. Mortality risk after atypical femoral fracture: A systematic review and meta-analysis. *Endocr. Pract.* **2022**, *28*, 1072–1077. [CrossRef]
11. Prasarn, M.L.; Ahn, J.; Helfet, D.L.; Lane, J.M.; Lorich, D.G. Bisphosphonate-associated femur fractures have high complication rates with operative fixation. *Clin. Orthop. Relat. Res.* **2012**, *470*, 2295–2301. [CrossRef] [PubMed]
12. Kim, T.; Moon, N.H.; Goh, T.S.; Jung, I.D. Detection of incomplete atypical femoral fracture on anteroposterior radiographs via explainable artificial intelligence. *Sci. Rep.* **2023**, *13*, 10415. [CrossRef] [PubMed]
13. Shane, E.; Burr, D.; Ebeling, P.R.; Abrahamsen, B.; Adler, R.A.; Brown, T.D.; Cheung, A.M.; Cosman, F.; Curtis, J.R.; Dell, R.; et al. Atypical subtrochanteric and diaphyseal femoral fractures: Report of a task force of the American Society for Bone and Mineral Research. *J. Bone Miner. Res.* **2010**, *25*, 2267–2294. [CrossRef]
14. Lee, Y.K.; Lee, Y.J.; Lee, N.K.; Nho, J.H.; Koo, K.H. Low positive predictive value of bone scan to predict impending complete fracture among incomplete atypical femoral fracture. *J. Korean Med. Sci.* **2018**, *33*, e157. [CrossRef]
15. Akgun, U.; Canbek, U.; Aydogan, N.H. Reliability and diagnostic utility of radiographs in patients with incomplete atypical femoral fractures. *Skelet. Radiol.* **2019**, *48*, 1427–1434. [CrossRef]
16. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
17. Zdolsek, G.; Chen, Y.; Bögl, H.P.; Wang, C.; Woisetschläger, M.; Schilcher, J. Deep neural networks with promising diagnostic accuracy for the classification of atypical femoral fractures. *Acta Orthop.* **2021**, *92*, 394–400. [CrossRef]
18. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NA, USA, 27–30 June 2016 ; pp. 770–778.
20. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
21. Schilcher, J.; Nilsson, A.; Andlid, O.; Eklund, A. Fusion of electronic health records and radiographic images for a multimodal deep learning prediction model of atypical femur fractures. *Comput. Biol. Med.* **2024**, *168*, 107704. [CrossRef]
22. Huang, S.C.; Pareek, A.; Seyyedi, S.; Banerjee, I.; Lungren, M.P. Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines. *NPJ Digit. Med.* **2020**, *3*, 136. [CrossRef]
23. Holste, G.; Partridge, S.C.; Rahbar, H.; Biswas, D.; Lee, C.I.; Alessio, A.M. End-to-end learning of fused image and non-image features for improved breast cancer classification from mri. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3294–3303.
24. Dietterich, T.G. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 1–15.
25. Kanopoulos, N.; Vasanthavada, N.; Baker, R.L. Design of an image edge detection filter using the Sobel operator. *IEEE J. Solid-State Circuits* **1988**, *23*, 358–367. [CrossRef]
26. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
27. Tan, M. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.

28. Howard, A.G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.

29. Park, J.Y.; Lee, S.H.; Kim, Y.J.; Kim, K.G.; Lee, G.J. Machine learning model based on radiomics features for AO/OTA classification of pelvic fractures on pelvic radiographs. *PLoS ONE* **2024**, *19*, e0304350. [CrossRef]

30. Van Griethuysen, J.J.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H.J. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.* **2017**, *77*, e104–e107. [CrossRef]

31. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.

32. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.

33. Hosmer,D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2013.

34. Sahin, M.E. Image processing and machine learning-based bone fracture detection and classification using X-ray images. *Int. J. Imaging Syst. Technol.* **2023**, *33*, 853–865. [CrossRef]

35. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

36. Freund, Y.; Schapire, R.; Abe, N. A short introduction to boosting. *J. Jpn. Soc. Artif. Intell.* **1999**, *14*, 1612.

37. Açıcı, K.; Sümer, E.; Beyaz, S. Comparison of different machine learning approaches to detect femoral neck fractures in X-ray images. *Health Technol.* **2021**, *11*, 643–653. [CrossRef]

38. Peterson, L.E. K-nearest neighbor. *Scholarpedia* **2009**, *4*, 1883. [CrossRef]

39. Rish, I. An empirical study of the naive Bayes classifier. In Proceedings of the IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle, WA, USA, 4–6 August 2001; Volume 3, pp. 41–46.

40. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13*, 18–28. [CrossRef]

41. Tesfaw, A.; Teshager, A.; Mirolgn, A.; Ambachew, E. Bone Fracture Classification Using Heterogeneous Ensemble Machine Learning and Deep Learning Algorithms. *SSRN* **2022**. [CrossRef]

42. Anwar, T.; Anwar, H. Lsnet: A novel cnn architecture to identify wrist fracture from a small X-ray dataset. *Int. J. Inf. Technol.* **2023**, *15*, 2469–2477. [CrossRef]

43. Iandola, F.N. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. *arXiv* **2016**, arXiv:1602.07360.

44. Koch, G.; Zemel, R.; Salakhutdinov, R. Siamese neural networks for one-shot image recognition. In Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 7–9 July 2015; Volume 2, pp. 1–30.

45. Yadav, D.P.; Sharma, A.; Athithan, S.; Bhola, A.; Sharma, B.; Dhaou, I.B. Hybrid SFNet model for bone fracture detection and classification using ML/DL. *Sensors* **2022**, *22*, 5823. [CrossRef]

46. Rong, W.; Li, Z.; Zhang, W.; Sun, L. An improved CANNY edge detection algorithm. In Proceedings of the 2014 IEEE International Conference on Mechatronics and Automation, Tianjin, China, 3–6 August 2014; pp. 577–582.

47. Jiménez-Sánchez, A.; Mateus, D.; Kirchhoff, S.; Kirchhoff, C.; Biberthaler, P.; Navab, N.; González Ballester, M.A.; Piella, G. Medical-based deep curriculum learning for improved fracture classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, 13–17 October 2019*; Proceedings, Part VI 22; Springer: Berlin/Heidelberg, Germany, 2019; pp. 694–702.

48. Bengio, Y.; Louradour, J.; Collobert, R.; Weston, J. Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 41–48.

49. Lee, C.; Jang, J.; Lee, S.; Kim, Y.S.; Jo, H.J.; Kim, Y. Classification of femur fracture in pelvic X-ray images using meta-learned deep neural network. *Sci. Rep.* **2020**, *10*, 13694. [CrossRef]

50. Wang, Y.; Bai, T.; Li, T.; Huang, L. Osteoporotic Vertebral Fracture Classification in X-rays Based on a Multi-modal Semantic Consistency Network. *J. Bionic Eng.* **2022**, *19*, 1816–1829. [CrossRef]

51. Bria, A.; Marrocco, C.; Tortorella, F. Addressing class imbalance in deep learning for small lesion detection on medical images. *Comput. Biol. Med.* **2020**, *120*, 103735. [CrossRef]

52. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]

53. Savelli, B.; Bria, A.; Molinara, M.; Marrocco, C.; Tortorella, F. A multi-context CNN ensemble for small lesion detection. *Artif. Intell. Med.* **2020**, *103*, 101749. [CrossRef] [PubMed]

54. Chang, J.; Choi, I.; Lee, M. PESA R-CNN: Perihematomal Edema Guided Scale Adaptive R-CNN for Hemorrhage Segmentation. *IEEE J. Biomed. Health Inform.* **2022**, *27*, 397–408. [CrossRef] [PubMed]

55. Wang, K.Y.; Wu, C.H.; Zhou, L.Y.; Yan, X.H.; Yang, R.L.; Liao, L.M.; Ge, X.M.; Liao, Y.s.; Li, S.J.; Li, H.Z.; et al. Ultrastructural changes of brain tissues surrounding hematomas after intracerebral hemorrhage. *Eur. Neurol.* **2015**, *74*, 28–35. [CrossRef] [PubMed]

56. Liu, Y.; Li, X.; Li, T.; Li, B.; Wang, Z.; Gan, J.; Wei, B. A deep semantic segmentation correction network for multi-model tiny lesion areas detection. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 89. [CrossRef]

57. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015*; Proceedings, part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.

58. Lin, X.; Yan, Z.; Kuang, Z.; Zhang, H.; Deng, X.; Yu, L. Fracture R-CNN: An anchor-efficient anti-interference framework for skull fracture detection in CT images. *Med. Phys.* **2022**, *49*, 7179–7192. [CrossRef]

59. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]

60. Ren, M.; Yi, P.H. Deep learning detection of subtle fractures using staged algorithms to mimic radiologist search pattern. *Skelet. Radiol.* **2022**, *51*, 345–353. [CrossRef]

61. Guan, B.; Zhang, G.; Yao, J.; Wang, X.; Wang, M. Arm fracture detection in X-rays based on improved deep convolutional neural network. *Comput. Electr. Eng.* **2020**, *81*, 106530. [CrossRef]

62. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.

63. Yao, L.; Guan, X.; Song, X.; Tan, Y.; Wang, C.; Jin, C.; Chen, M.; Wang, H.; Zhang, M. Rib fracture detection system based on deep learning. *Sci. Rep.* **2021**, *11*, 23513. [CrossRef]

64. Ioffe, S. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* **2015**, arXiv:1502.03167.

65. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.

66. Nwankpa, C.; Ijomah, W.; Gachagan, A.; Marshall, S. Activation functions: Comparison of trends in practice and research for deep learning. *arXiv* **2018**, arXiv:1811.03378.

67. Fischler, M.A.; Bolles, R.C. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]

68. Hough, P.V. Method and Means for Recognizing Complex Patterns. U.S. Patent 3,069,654, 18 December 1962.

69. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.

70. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.

71. Qin, D.; Leichner, C.; Delakis, M.; Fornoni, M.; Luo, S.; Yang, F.; Wang, W.; Banbury, C.; Ye, C.; Akin, B.; et al. MobileNetV4-Universal Models for the Mobile Ecosystem. *arXiv* **2024**, arXiv:2404.10518.

72. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.

73. Kim, D.; Heo, B.; Han, D. DenseNets Reloaded: Paradigm Shift Beyond ResNets and ViTs. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 395–415.

74. Dosovitskiy, A. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

75. Wang, A.; Chen, H.; Lin, Z.; Han, J.; Ding, G. Repvit: Revisiting mobile cnn from vit perspective. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 15909–15920.

76. Vasu, P.K.A.; Gabriel, J.; Zhu, J.; Tuzel, O.; Ranjan, A. FastViT: A fast hybrid vision transformer using structural reparameterization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 5785–5795.

77. Maaz, M.; Shaker, A.; Cholakkal, H.; Khan, S.; Zamir, S.W.; Anwer, R.M.; Shahbaz Khan, F. Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 3–20.

78. Yang, J.; Li, C.; Dai, X.; Gao, J. Focal modulation networks. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 4203–4217.

79. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [CrossRef]