https://doi.org/10.31436/iiumej.v27i2.4341

Efficiency-Aware Multi-Class Spinal Disorder Classification Using CBAM-Enhanced Lightweight CNNs with Dual-Branch Fusion

TEDDY SURYA GUNAWAN[1*], NURUL JANNAH[2], MIRA KARTIWI[3], NOREHA ABDUL MALIK[1]

[1]*Electrical and Computer Engineering Dept., International Islamic University Malaysia, Malaysia*
[2]*Mechatronics Engineering Dept., International Islamic University Malaysia, Malaysia*
[3]*Information Systems Dept., International Islamic University Malaysia, Malaysia*

**Corresponding author: tsgunawan@iium.edu.my*

(Received: 6 March 2026; Accepted: 25 April 2026; Published online: 10 May 2026)

ABSTRACT: Spinal X-rays are still often read through manual measurements, yet the patients who most need timely assessment cannot afford delay, inconsistency, or heavy computational pipelines. Motivated by this clinical tension, this study proposes an efficiency-aware deep learning framework for three-class spinal disorder classification that asks a practical question rarely centered in prior work: not only which model is most accurate, but which model is accurate enough, light enough, and fast enough to matter in real screening settings. Using a public dataset of 338 subjects, five lightweight backbones, CBAM-enhanced variants, and a dual-branch fusion model were evaluated through stratified 5-fold cross-validation under multiple balancing strategies, with performance measured by accuracy, precision, recall, F1-score, parameter count, FLOPs, model size, latency, and throughput. The results reveal an unexpected pattern: bigger models do not win. MobileNetV3Small delivers the strongest efficiency-performance balance, reaching an F1-score of 0.962 with only 1.0 million parameters, while the best overall result is achieved by the Fusion_MNv3_MNAS model under augmentation-only training, with an F1-score of 0.976. Ablation findings further show that attention and fusion are not universally beneficial, but become most effective when paired with sufficient data-driven regularization, and that fine-tuning about 30% of backbone parameters yields the most favorable adaptation. Taken together, these findings show that performance in spinal X-ray classification depends less on model size alone than on the fit between architecture and training strategy. The study therefore offers a concrete and clinically relevant message: lightweight, well-regularized models can match or surpass heavier alternatives while remaining more practical for scalable deployment.

ABSTRAK: Radiograf tulang belakang masih kerap dinilai melalui pengukuran manual, sedangkan pesakit yang memerlukan rawatan awal tidak dapat menanggung kelewatan, ketidakselarasan, atau kebergantungan pada sistem pengiraan yang berat. Berpunca pada masalah klinikal ini, kajian ini mengemukakan satu rangka kerja pembelajaran mendalam berpaksikan kecekapan bagi pengelasan tiga kelas gangguan tulang belakang, dengan menumpukan persoalan praktikal yang jarang diberi perhatian dalam kajian terdahulu, iaitu bukan pada model mana yang paling tepat, tetapi model mana yang cukup tepat, cukup ringan, dan cukup pantas bagi persekitaran saringan klinikal. Dengan menggunakan satu set data awam yang melibatkan 338 subjek, lima model asas ringan, varian yang dipertingkatkan dengan CBAM, serta model gabungan dwi-cabang telah dinilai melalui pengesahan silang berstrata lima lipatan di bawah beberapa strategi pengimbangan kelas, dengan prestasi diukur menggunakan ketepatan, kejituan, keboleh ingatan, skor F1, bilangan parameter, FLOPs, saiz model, kependaman, dan kadar pemprosesan. Dapatan kajian menunjukkan satu corak yang tidak dijangka, iaitu model yang lebih besar tidak semestinya memberi prestasi terbaik.

MobileNetV3Small memperlihatkan keseimbangan paling kukuh antara kecekapan dan prestasi dengan mencapai skor F1 sebanyak 0.962 hanya menggunakan 1.0 juta parameter, manakala prestasi keseluruhan terbaik dicapai oleh model Fusion_MNv3_MNAS di bawah latihan berasaskan augmentasi sahaja dengan skor F1 sebanyak 0.976. Analisis ablasi seterusnya menunjukkan bahawa mekanisme perhatian dan gabungan tidak sentiasa memberikan manfaat secara menyeluruh, sebaliknya menjadi paling berkesan apabila dipadankan dengan regularisasi berasaskan data yang mencukupi, dan penalaan halus sekitar 30% parameter rangka asas menghasilkan penyesuaian yang terbaik. Secara keseluruhan, dapatan ini menunjukkan bahawa prestasi dalam pengelasan sinar-X tulang belakang kurang bergantung pada saiz model semata-mata, sebaliknya lebih dipengaruhi oleh kesesuaian antara seni bina model dan strategi latihan. Oleh itu, kajian ini membawa mesej yang jelas dan signifikan dari sudut klinikal, iaitu model ringan yang diregularisasikan dengan baik mampu menandingi malah mengatasi model yang lebih berat, iaitu kekal lebih praktikal bagi pelaksanaan berskala dalam persekitaran saringan klinikal.

KEYWORDS: *Spinal Disorder Classification, Deep Learning, Lightweight CNN, Attention Mechanism (CBAM), Efficiency-Aware Evaluation*

1. INTRODUCTION

A spinal X-ray may look routine, but the decision drawn from it can shape the rest of a patient's life. Conditions such as scoliosis and spondylolisthesis often begin as subtle structural abnormalities on radiographs, yet have significant outcomes. When diagnosis is delayed, the consequences can extend far beyond posture or back pain, progressing into deformity, functional limitation, and more invasive treatment. Scoliosis is commonly identified through abnormal lateral curvature of the spine, whereas spondylolisthesis is marked by the forward displacement of one vertebra over another. In daily clinical practice, both conditions are still largely assessed through manual radiographic measurements, including Cobb angle estimation and vertebral slippage grading. These methods remain clinically valuable, but they are time-consuming, operator-dependent, and vulnerable to inter- and intra-observer variation, which may affect diagnostic consistency and treatment planning [1, 2]. The problem, therefore, is not merely whether spinal disorders can be seen on X-rays, but whether they can be assessed quickly, consistently, and at scale.

Recent progress in deep learning has made automated spinal image analysis increasingly feasible. Convolutional neural networks (CNNs), particularly when used through transfer learning, have shown a strong ability to extract discriminative patterns directly from radiographs and to support accurate classification and detection of spinal abnormalities [3, 4]. Prior studies have reported high performance on three-class spinal classification tasks [1], while others have shown that model selection and optimization strategy can substantially affect the final outcome [5]. At the same time, related work has expanded into more specialized tasks such as spondylolisthesis detection, grading, and segmentation [6-8]. Yet an important gap remains. Much of the existing literature is driven by the pursuit of higher accuracy alone, as if the best model is simply the one with the highest number. In practice, that assumption is incomplete. A model that is accurate but computationally heavy, slow, or difficult to deploy may be of limited value in real clinical screening settings. Moreover, much prior work treats scoliosis and spondylolisthesis as separate problems, whereas real screening scenarios often require a unified decision across multiple classes.

This gap becomes even more important when lightweight models are considered. Compact CNN architectures are attractive because they reduce memory demand, model size, and inference cost, all of which matter in resource-constrained or high-throughput clinical

https://doi.org/10.31436/iiumej.v27i2.4341

environments. However, lightweight efficiency often comes with a representational cost: smaller networks may overlook subtle anatomical cues that distinguish a mildly abnormal spine from a normal one. Attention mechanisms provide a promising way to address this tension. Modules such as Squeeze-and-Excitation and the Convolutional Block Attention Module (CBAM) help networks emphasize the most informative channels and spatial regions, thereby strengthening feature discrimination without a prohibitive computational burden [9-11]. In other medical imaging tasks, including chest X-ray analysis and retinal disease classification, such mechanisms have improved performance with relatively modest overhead [12-14]. Even so, their role in lightweight spinal X-ray classification remains insufficiently understood, particularly when the evaluation extends beyond accuracy to include efficiency, robustness, and deployment relevance.

Against this background, this study asks a practical but underexplored question: in spinal X-ray classification, what kind of model is not only accurate but also efficient enough to matter? To answer this question, this paper proposes an efficiency-aware deep learning framework for three-class classification of spinal disorders from X-ray images. Five lightweight baseline architectures, namely EfficientNetB0, MobileNetV2, MobileNetV3Small, MNASNet, and DenseNet121, are systematically evaluated and then contrasted with three CBAM-augmented variants and a dual-branch fusion model. Rather than reporting performance solely by accuracy, the study adopts a broader evaluation lens that includes F1-score, parameter count, floating-point operations (FLOPs), model size, latency, and throughput. To improve the credibility of the findings, the experiments are conducted using stratified 5-fold cross-validation, multiple class-balancing strategies, and targeted ablation studies on attention design, fine-tuning depth, fusion strategy, and data balancing. This design allows the study to move beyond leaderboard-style comparison and instead uncover why certain models work better, under what conditions, and at what computational cost.

The contribution of this work extends beyond applying another deep learning model to spinal imaging. Its main contributions are as follows:

- *It reframes the central question from accuracy alone to practical value*. The study proposes an efficiency-aware transfer learning framework for unified three-class spinal disorder classification and compares nine models using both effectiveness and efficiency metrics, including F1-score, parameter count, FLOPs, model size, latency, and throughput, under stratified 5-fold cross-validation and multiple class-balancing strategies, judging models by how lightly and quickly they operate as well as how well they predict.
- *It shows that attention is not magic, but conditional*. By integrating CBAM into lightweight CNNs, this work demonstrates that attention mechanisms do not automatically improve performance in every setting; their benefit becomes convincing only when paired with sufficient data-driven regularization, especially augmentation-based training.
- *It introduces a lightweight fusion strategy that is both strong and purposeful*. The proposed Fusion_MNv3_MNAS model combines complementary feature representations from two efficient backbones to improve robustness and achieve the best overall classification performance, while still avoiding the heavy computational burden of larger architectures.
- *It explains why performance changes, not just which model wins*. Through targeted ablation studies on attention components, fine-tuning depth, fusion design, and balancing strategy, this work shows that performance in spinal X-ray classification depends less on model size alone than on the fit between architecture, regularization, and training strategy.

https://doi.org/10.31436/iiumej.v27i2.4341

2. RECENT ADVANCES IN SPINAL DISORDER CLASSIFICATION

Deep learning has changed the landscape of spinal X-ray analysis. What was once dominated by manual measurements and clinician-dependent interpretation is now increasingly supported by models that can learn diagnostic patterns directly from radiographs. This shift is important because disorders such as scoliosis and spondylolisthesis are not minor visual irregularities. They are clinically significant conditions that can alter posture, function, pain, and long-term quality of life if not recognized early. In recent years, the literature has expanded rapidly across classification, detection, and segmentation tasks for spinal abnormalities [1, 6, 15]. At the same time, lightweight architectures and attention mechanisms have gained momentum because they enable models that are not only accurate but also efficient enough for practical deployment. Even so, the literature remains fragmented. Some studies emphasize predictive accuracy almost exclusively, others focus on single diagnostic tasks, and still others adopt architectures whose deployment costs are rarely discussed. As a result, the field has progressed quickly, but not always coherently. This section, therefore, reviews prior work through three connected themes: deep learning for spinal X-ray classification, lightweight architectures and efficiency-aware evaluation, and attention mechanisms for feature enhancement. The aim is not merely to summarize earlier studies, but to show what the field has achieved, what it has overlooked, and why a more balanced framework is now needed.

2.1. Deep Learning for Spinal X-Ray Classification

The earliest breakthroughs in this area established an important fact: spinal disorder classification from X-ray images can be automated with remarkably high performance. A notable study demonstrated the feasibility of three-class classification among normal spine, scoliosis, and spondylolisthesis using transfer learning on a dataset of 338 subjects, reporting a maximum accuracy of 98.02% [1]. That work was influential not only because of its strong performance, but also because it recognized that accuracy alone can be misleading in imbalanced datasets and highlighted F1-score as a more reliable measure. Using the same benchmark dataset, subsequent work explored different architectures and optimization strategies and reported an accuracy of 99.01% with AlexNet using stochastic gradient descent with momentum [5]. Together, these studies showed that deep learning was not merely promising in spinal X-ray classification. It was already highly competitive. Yet they also raised a deeper question. When several models report near-perfect results on a relatively small dataset, the issue is no longer whether deep learning works, but how robust those results are, how much they depend on the training setup, and whether the gains remain meaningful beyond a single benchmark split.

The literature has also expanded beyond simple classification into detection and segmentation tasks. A Faster R-CNN framework with a ResNet backbone was introduced for automated scoliosis diagnosis, showing that region-based deep learning models could assist or potentially replace parts of manual radiographic assessment [15]. More recent studies pushed toward larger datasets and more specialized objectives. For instance, a YOLOv8-based segmentation model was developed for spondylolisthesis detection using over 10,000 images, achieving a precision of 96.77% [7]. Another study combined deep learning with geometric analysis for grading spondylolisthesis and reported accuracies above 94% on both internal and external datasets [8]. In parallel, lightweight architectures such as EfficientNet and MobileNet variants were explored in musculoskeletal imaging tasks, suggesting that compact models can remain highly competitive while reducing computational burden [2, 14, 16]. These developments point to a clear trend: spinal imaging research is moving toward automated systems that are faster, broader, and more clinically relevant.

https://doi.org/10.31436/iiumej.v27i2.4341

However, the dominant pattern in prior work remains accuracy-centric. Many studies report high performance, often using single train-test splits, without systematically examining computational efficiency, fold-to-fold variance, or deployment constraints [1, 5]. In addition, a large part of the literature remains task-specific, focusing on either scoliosis or spondylolisthesis in isolation rather than on unified multi-class screening. This matters because real clinical workflows do not always present a single isolated question. They often require a model to distinguish among several plausible conditions within one decision framework. The literature has therefore shown that automated spinal analysis is possible, but it has not yet fully established what makes such systems robust, efficient, and practical for real-world clinical use.

2.2. Lightweight Architectures and Efficiency-Aware Evaluation

As deep learning matured, a more practical concern came into view: a model that performs well in a paper may still be unsuitable in practice if it is too large, too slow, or too resource-intensive to deploy. This concern has driven the rise of lightweight convolutional neural network architectures that seek to preserve strong predictive performance while reducing memory demand and inference cost. EfficientNet introduced compound scaling to coordinate network depth, width, and input resolution more systematically [17]. MobileNetV2 reduced computational load through inverted residual blocks and linear bottlenecks [18]. MobileNetV3 extended this efficiency-driven philosophy by integrating squeeze-and-excitation modules and neural architecture search for mobile and embedded environments [19]. MNASNet pushed the idea further by explicitly optimizing latency on target platforms [20]. DenseNet121, although architecturally different, improved feature reuse and gradient flow through dense connectivity while still remaining relatively compact compared with heavier classical backbones [21]. Together, these models represent a broader shift in deep learning design, from raw capacity toward purposeful efficiency.

This shift is especially meaningful in medical imaging. Clinical systems are not judged solely by peak accuracy but also by their ability to operate reliably under practical constraints. Several studies on chest X-ray analysis, retinal disease detection, and histopathology have shown that lightweight CNNs can achieve performance comparable to larger models while requiring far less computation [22, 23]. In structured imaging domains, this finding is particularly striking because it challenges the common intuition that deeper and larger models are always superior. In some cases, compact models perform competitively precisely because the visual domain is constrained and clinically relevant patterns are more structured than those in natural image benchmarks [23]. EfficientNet-based models, for example, have shown favorable accuracy-efficiency trade-offs in clinical classification settings [24]. These findings suggest that model selection should not be driven solely by accuracy, but by the balance between predictive performance and operational feasibility.

Despite this growing recognition, efficiency-aware evaluation is still underused in spinal X-ray analysis. Many studies continue to report classification performance without accompanying measures such as parameter count, floating-point operations, model size, latency, or throughput. This omission is not trivial. In healthcare, a model that is accurate but difficult to scale or too slow for practical integration may have limited clinical value. Recent discussions in medical artificial intelligence increasingly emphasize that deployable systems must be not only accurate, but also interpretable, scalable, and computationally feasible [25, 26]. Within this context, efficiency-aware evaluation is no longer a secondary consideration. It is becoming part of a responsible model assessment.

https://doi.org/10.31436/iiumej.v27i2.4341

2.3. Attention Mechanisms in Medical Imaging

If lightweight models solve one problem, they can also create another. Their compactness often comes at the cost of reduced representational capacity, which may limit their ability to capture subtle diagnostic cues. In spinal radiographs, such cues can be clinically decisive. A mild scoliotic curvature or a slight vertebral displacement may be easy to miss if the model does not learn where to focus. Attention mechanisms have emerged as an effective approach to this problem because they enable networks to emphasize informative features without significantly increasing computational cost. One of the earliest influential designs was the Squeeze-and-Excitation Network, which introduced channel-wise attention by adaptively recalibrating feature responses [10]. The Convolutional Block Attention Module extended this idea by combining channel attention with spatial attention, enabling a model to learn both which features matter and where the most relevant image regions are located [9, 11]. This combination makes CBAM particularly attractive for lightweight CNNs, where efficiency remains essential.

The usefulness of attention mechanisms has already been demonstrated across a broad range of medical imaging tasks. Attention-guided CNNs have improved thoracic disease classification from chest X-rays by directing the model toward clinically relevant regions [12]. Attention modules have also enhanced performance and interpretability in general image classification settings [27]. Attention U-Net improved segmentation by emphasizing salient anatomical structures [13], while related attention-based designs in retinal imaging improved classification accuracy with relatively small computational overhead [28]. Across these tasks, the message is consistent: attention can make models more discriminative without substantially increasing their weight. This is especially valuable when model efficiency is a design constraint rather than an afterthought.

In addition to attention mechanisms, fusion and multi-branch architectures have been investigated to improve robustness by combining complementary feature representations. Multi-branch networks can capture different feature scales and contextual patterns, and their outputs can be aggregated through strategies such as feature concatenation or logit averaging [29, 30]. These designs are particularly appealing in medical imaging, where limited data and subtle class boundaries often mean that no single representation is sufficient. Yet despite the growing success of attention and fusion strategies in other imaging domains, their application to spinal X-ray classification remains limited. This gap is important because it suggests that the field has explored many useful ideas independently, but has not yet fully tested how they interact within a unified and efficiency-aware framework.

2.4. Research Gap Analysis and Study Positioning

The literature on spinal X-ray analysis has reached an interesting stage. The field has already shown that deep learning can work, sometimes with strikingly high reported performance. Early studies demonstrated that transfer learning could classify normal spine, scoliosis, and spondylolisthesis with very high accuracy on a dataset of 338 subjects [1, 5]. Other studies expanded the scope to include scoliosis detection, spondylolisthesis segmentation, and grading using larger datasets and more specialized pipelines [7, 8, 15]. At first glance, this progress suggests the problem is nearly solved. Yet a closer reading shows that the story is less complete. High accuracy has been reported, but often under narrow evaluation settings, with limited discussion of robustness, efficiency, or deployment cost. In other words, the field has produced strong answers to the question of whether spinal disorder classification is possible, but weaker answers to the more important question of which solutions are reliable and practical enough to matter in real screening settings.

Three gaps remain especially clear. First, much of the literature is still driven by an accuracy-centric evaluation culture. Reported results are often based on single train-test splits, and therefore say little about statistical stability or sensitivity to data partitioning [1, 5]. This becomes particularly important when the benchmark dataset is relatively small, because even strong headline results may reflect favorable splits rather than consistently robust behavior. Second, computational efficiency is rarely treated as a core outcome. Most prior studies emphasize predictive performance while omitting measures such as parameter count, floating-point operations, model size, and inference latency, even though these factors strongly influence whether a model can be deployed in real clinical environments. Third, prior work is frequently task-specific. Many studies address scoliosis alone, spondylolisthesis alone, or a specialized subtask, such as segmentation or grading [2, 7, 8, 15], whereas practical screening requires a unified framework that can distinguish among multiple clinically plausible classes.

Recent progress in lightweight architecture design and attention mechanisms points toward a more meaningful direction, but important combinations remain underexplored. Lightweight backbones such as MobileNet and EfficientNet suggest that strong performance does not always require heavy networks [2, 14, 16, 24]. Attention mechanisms such as SE-Net and CBAM have shown that feature selection can be made more discriminative without incurring prohibitive computational costs [9, 14]. Fusion strategies have also shown promise in other medical imaging tasks by combining complementary representations to improve robustness [29, 30]. However, these elements have seldom been studied together in spinal X-ray classification under a unified and efficiency-aware protocol. This distinction is important. The novelty of the present study does not lie in claiming that CBAM alone is new, or that fusion alone is unprecedented. Rather, its novelty lies in bringing together lightweight backbones, CBAM-based enhancement, dual-branch fusion, class-balancing strategies, stratified 5-fold cross-validation, and efficiency-aware evaluation into a single coherent framework, and then analyzing not only which model performs best but also why performance changes across architectures and training conditions.

Table 1 sharpens this point by comparing representative prior studies not only in terms of task and best result, but also in terms of evaluation setting, efficiency reporting, and practical limitations. The contrast is revealing. Prior studies have made important contributions, but they largely optimize one dimension at a time: accuracy, detection capability, geometric precision, or task specialization. What remains missing is a framework that jointly addresses unified three-class classification, robustness through cross-validation, and deployability through explicit efficiency analysis. This is precisely where the present study is positioned.

To address these gaps, this study proposes an efficiency-aware deep learning framework for three-class classification of spinal disorders from X-ray images. The framework integrates lightweight CNN backbones with CBAM-based attention and a dual-branch fusion strategy, and evaluates them using stratified 5-fold cross-validation, multiple class-balancing strategies, and targeted ablation studies. More importantly, it treats model performance as something to interpret rather than merely announce. The central claim is therefore stronger than a simple accuracy comparison: in spinal X-ray classification, performance is governed not only by architecture but also by the interaction among architecture, regularization, and evaluation design. By jointly addressing effectiveness, efficiency, and robustness, this study aims to move the field from isolated high-scoring models toward more credible, explainable, and clinically deployable solutions.

Table 1. Comparison of Representative Deep Learning Approaches for Spinal X-Ray Analysis

Study	Main Task	Dataset	Approach	Best Reported Result	Evaluation Setting	Efficiency Reported	Key Limitation
Fraiwan *et al*., 2022 [1]	Three-class classification	338 subjects	Transfer learning CNN	98.02% accuracy	Fixed split	No	No cross-validation; limited generalizability analysis; no efficiency evaluation
Güneş *et al*., 2024 [5]	Three-class classification	338 subjects	AlexNet + SGDM	99.01% accuracy	Fixed split	No	Single-split evaluation; accuracy-centric comparison; no efficiency metrics
Chen *et al*., 2022 [15]	Scoliosis detection/classification	Spine X-rays	Faster R-CNN + ResNet	Strong automated detection performance	Task-specific evaluation	No	Heavy architecture; high computational cost; limited real-time practicality
Vephasayanant *et al*., 2024 [7]	Spondylolisthesis detection/segmentation	10,616 images	YOLOv8 segmentation	96.77% precision	Task-specific evaluation	No	Segmentation task; not directly comparable to three-class classification; no efficiency analysis
Xu *et al*., 2025 [8]	Spondylolisthesis diagnosis/grading	4,400 images	Deep learning + geometric analysis	96.1% accuracy	Internal and external datasets	No	Multi-stage pipeline; increased complexity; limited deployment feasibility
Maaliw, 2023 [2]	Scoliosis severity/Cobb angle estimation	Spine X-rays	Deep learning regression (SCOLIONET)	High correlation with ground truth	Regression setting	No	Regression-only task; not a multi-class classifier; no efficiency evaluation
This study	Three-class classification	338 subjects	CBAM-enhanced lightweight CNNs + dual-branch fusion	Comprehensive effectiveness-efficiency evaluation	Stratified 5-fold cross-validation with ablation studies	Yes	Addresses unified classification, robustness, and deployment-oriented evaluation

3. RESEARCH METHODOLOGY

This study was designed to answer a practical question with scientific discipline: which deep learning strategy can classify spinal disorders accurately while remaining efficient enough for realistic deployment? To answer that question, a structured experimental framework was developed that moves from data preparation to evaluation in a transparent and reproducible manner. The overall pipeline consists of dataset preparation, image preprocessing, model

https://doi.org/10.31436/iiumej.v27i2.4341

development, training under multiple class-balancing strategies, and performance assessment using both effectiveness and efficiency metrics. Rather than treating methodology as a sequence of disconnected technical steps, the framework was intentionally designed to address specific challenges in spinal X-ray classification, namely, limited data, class imbalance, fairness in comparison, and deployment relevance.

3.1. Experimental Setup

All experiments were conducted on a high-performance computing platform to ensure consistency, reproducibility, and fair comparison across all evaluated models. The implementation was developed using the PyTorch deep learning framework, selected for three practical reasons. First, it offers strong native CUDA support for GPU-accelerated training on Windows platforms. Second, its eager execution mode facilitates transparent debugging and flexible model development. Third, it has become one of the most widely adopted frameworks in contemporary deep learning research, making the implementation easier to inspect, reproduce, and extend. The complete hardware and software configuration used in this study is summarized in Table 2.

Table 2. Hardware and Software Configuration

Component	Specification
Operating System	Windows Server Data Center 2022 (Version 21H2, Build 20348.4893)
System Architecture	64-bit
CPU	Intel Core i9-14900K (32 threads)
RAM	64 GB DDR5
Storage	2 TB SSD + 4 TB HDD
GPU	NVIDIA GeForce RTX 4090 (24 GB VRAM)
CUDA Version	12.4
Python	3.11.15
PyTorch	2.3.1
torchvision	0.18.1
NumPy	1.26.4
Environment	Conda (Scoliosis)
Reproducibility Seed	42
Deterministic Setting	`cudnn.deterministic = True`
Benchmark Setting	`cudnn.benchmark = False`

Reproducibility was treated as a methodological requirement rather than an afterthought. A fixed random seed of 42 was applied across Python, NumPy, and PyTorch, including both CPU and CUDA operations. In addition, deterministic behavior was enforced by setting `torch.backends.cudnn.deterministic = True` and disabling runtime performance benchmarking through `torch.backends.cudnn.benchmark = False`. These settings reduce non-deterministic variation during training and help ensure that experimental outcomes can be reproduced consistently across repeated runs. This is particularly important in a study comparing multiple lightweight architectures, where small performance differences should reflect model behavior rather than random execution.

To further strengthen transparency and reproducibility, the full source code of the proposed framework has been made publicly available at the project repository available at https://github.com/tsgunawan/SpinalDisordersClassification. The code is released under the MIT License, allowing unrestricted use, modification, and redistribution with minimal legal restriction. By making both the implementation and configuration explicit, this study aims to support not only result verification but also future extension by other researchers.

https://doi.org/10.31436/iiumej.v27i2.4341

3.2. Dataset

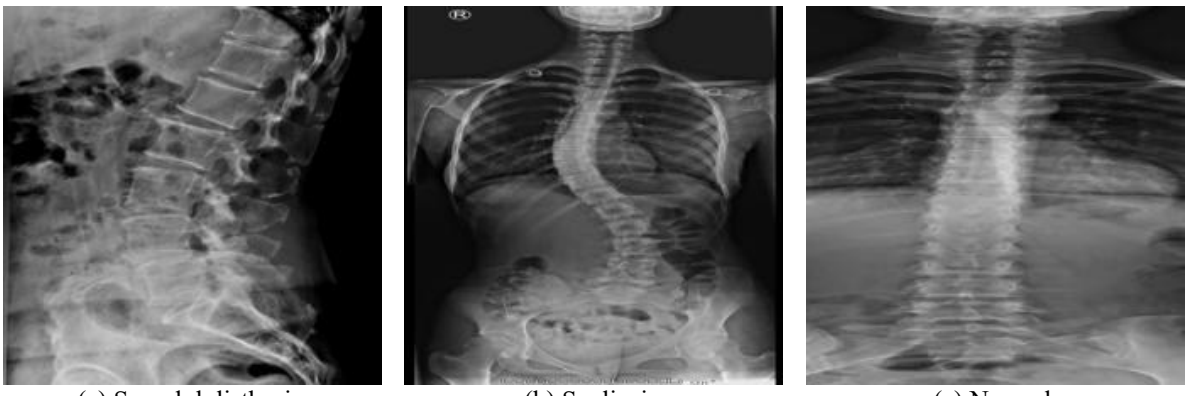
This study uses a publicly available spinal X-ray dataset introduced in [1], which has been widely adopted in prior work on automated classification of spinal disorders. The dataset contains 338 subjects and reflects a clinically meaningful screening scenario involving three diagnostic categories: Normal, Scoliosis, and Spondylolisthesis. Although the dataset is modest in size, this is precisely what makes it important. Small clinical datasets are common in real medical AI research, and they create a demanding setting in which model performance depends not only on architecture, but also on how carefully the experiment is designed. In this study, the limited dataset size was therefore treated not as a weakness to be ignored, but as a condition that required stricter evaluation and stronger methodological control.

To avoid inheriting bias from earlier evaluation protocols, the original dataset splits used in prior studies were not reused. Instead, all available samples were pooled and repartitioned using stratified 5-fold cross-validation. This strategy preserves the class proportion in each fold while making fuller use of the available data. Within each fold, the training portion was further divided into training and validation subsets using an 80:20 stratified split. This design provides a more reliable estimate of generalization performance than a single train-test split, especially when the dataset is small and imbalanced. It also allows performance to be reported as mean and standard deviation across folds, which gives a more credible picture of both central tendency and variability.

Table 3. Class Distribution of the Spine X-Ray Dataset

Class	**Subjects**	**Female**	**Male**	**Proportion**
Normal	71	40	31	21.0%
Scoliosis	188	151	37	55.6%
Spondylolisthesis	79	49	30	23.4%
Total	338	240	98	100%

As shown in Table 3, the dataset is inherently imbalanced, with scoliosis forming the majority class. This class distribution is clinically plausible, but it also introduces an important learning challenge. Without proper control, the model may become biased toward the dominant category and underperform on minority classes that are equally important from a diagnostic perspective. For this reason, multiple class-balancing strategies were incorporated into the experimental design, as described in Section 3.4, so that model performance could be examined under different responses to imbalance rather than under a single fixed assumption.



(a) Spondylolisthesis (b) Scoliosis (c) Normal

Figure 1. Sample Images from Each Class in the Spinal X-Ray Dataset

https://doi.org/10.31436/iiumej.v27i2.4341

Figure 1 presents representative X-ray images from the three diagnostic classes. These examples offer a useful visual reminder that the task is not trivial. The differences among classes are medically meaningful but, in some cases, visually subtle, especially when abnormality is mild or structural change is localized. The figure, therefore, illustrates why strong feature extraction and robust generalization are essential for deep learning-based spinal disorder classification. The challenge is not simply to detect obvious deformity, but to learn stable discriminative patterns across images that vary in anatomy, severity, and visual presentation.

3.3. Preprocessing and Augmentation

In a small medical imaging dataset, preprocessing is not a minor technical detail. It is the stage that determines whether models learn stable diagnostic patterns or become distracted by irrelevant variation. For that reason, this study adopts a unified preprocessing and augmentation pipeline across all evaluated models. The purpose is twofold: first, to ensure a fair comparison among architectures, and second, to fully leverage transfer learning from ImageNet-pretrained backbones. By standardizing the inputs before model-specific learning begins, the experiment reduces avoidable bias and keeps the comparison focused on what matters most, namely how each architecture responds to the same diagnostic problem.

3.3.1. Input Standardization

All input images were resized to a fixed resolution of 224 × 224 pixels to ensure compatibility with standard ImageNet-pretrained CNN architectures, including EfficientNet, MobileNet, and DenseNet variants [17, 19]. A common normalization scheme was then applied using the ImageNet mean and standard deviation values, namely mean = [0.485, 0.456, 0.406] and standard deviation = [0.229, 0.224, 0.225]. This choice aligns the input distribution with the pretrained model weights, thereby supporting more stable convergence during training [31]. More importantly, it eliminates inconsistencies that might otherwise arise from model-specific preprocessing choices. In other words, every model begins from the same visual footing, making the comparison more credible and methodologically cleaner.

3.3.2. Training Data Augmentation

Because the dataset is limited in size, the models must learn from clinically plausible variation rather than from repeated exposure to nearly identical images. To improve generalization and reduce overfitting, stochastic data augmentation was applied during training. The augmentation pipeline includes the following transformations:

- `RandomRotation` (±20°)
- `RandomResizedCrop` (scale = [0.85, 1.15])
- `ColorJitter` (brightness = 0.1, contrast = 0.1)
- `RandomHorizontalFlip`

These transformations introduce controlled changes in orientation, scale, and visual intensity while preserving the essential anatomical structure of the spine. The objective is not to aggressively distort the data, but to teach the model that clinically relevant patterns should remain recognizable even when the image presentation varies. This kind of augmentation has been widely shown to improve robustness in medical image classification by effectively increasing the diversity of training samples without collecting new data [32].

https://doi.org/10.31436/iiumej.v27i2.4341

3.3.3. Augmentation-Only Strategy

For the `augmentation_only` balancing strategy, the augmentation pipeline was expanded further to increase data diversity without relying on synthetic resampling. Two additional transformations were incorporated: `RandomAffine` and `RandomPerspective`. These operations simulate geometric perturbations and mild viewpoint variation, encouraging the model to develop more durable feature representations under realistic visual shifts. Unlike resampling-based techniques, this strategy works entirely with transformed versions of real images. As a result, it avoids the risk of generating artificial samples that may look mathematically plausible but anatomically unrealistic. This makes augmentation-only learning particularly attractive in medical imaging, where fidelity to real structure is critical.

3.3.4. Validation and Testing Protocol

For validation and testing, no augmentation was applied. All images were passed through a deterministic preprocessing pipeline that resized them to 224 × 224 pixels, converted them to tensor format, and normalized them using ImageNet statistics. This ensures that evaluation metrics reflect true model generalization rather than performance inflated by augmented inputs. The distinction is important: augmentation is used only to improve learning during training, not to make evaluation easier. In this way, the reported results remain a more faithful reflection of how the model behaves on unseen radiographs.

3.4. Class Balancing Strategies

Class imbalance is one of the most persistent challenges in medical image classification. When one category appears much more frequently than others, a model can achieve deceptively strong overall accuracy by learning the majority class well while performing poorly on clinically important minority classes. In a screening context, that imbalance is not a statistical inconvenience. It is a diagnostic risk. Formally, given a dataset $\mathcal{D} = [(x_i, y_i)]_{i=1}^{N}$ with class prior distribution $P(y)$, imbalance occurs when $P(y_k) \neq P(y_j)$ for classes $k \neq j$. Under such conditions, empirical risk minimization tends to bias the decision boundary toward the majority class, leading to poorer minority-class recognition and weaker generalization [33]. To address this issue, this study evaluates three complementary balancing strategies that act at different levels of the learning pipeline: loss-level adjustment, representation-level resampling, and input-level transformation. This design allows the study to examine not only whether balancing helps, but also which type of balancing works best for each model.

3.4.1. Class Weighting (class_weight)

The first strategy addresses the imbalance by modifying the loss function. Each class is assigned a weight inversely proportional to its frequency, so that minority classes contribute more strongly during optimization. Let C denote the number of classes, and w_c represent the weight assigned to the class c. The weight is defined as:

$$w_c = \frac{N}{C \cdot N_c} \tag{1}$$

where N is the total number of samples and N_c is the number of samples in class c. The weighted cross-entropy loss is then formulated as:

$$\mathcal{L}_{weighted} = -\sum_{i=1}^{N} \sum_{c=1}^{C} w_c y_{i,c} \log \hat{y}_{i,c} \tag{2}$$

where $y_{i,c}$ is the ground-truth indicator, and $\hat{y}_{i,c}$ is the predicted probability.

This strategy does not change the dataset itself. Instead, it changes the extent to which each class influences learning. Its main advantage is stability. It preserves the original data manifold,

https://doi.org/10.31436/iiumej.v27i2.4341

avoids synthetic sample generation, and provides a transparent baseline for imbalanced learning [33]. For that reason, class weighting is used in this study as the reference strategy against which the other balancing approaches are compared.

3.4.2. Feature-Space SMOTE (smote_features)

The second strategy extends the Synthetic Minority Over-sampling Technique (SMOTE) from the image domain into the learned feature space. In classical SMOTE, new minority samples are generated by interpolating between neighboring examples:

$$x_{\text{new}} = x_i + \lambda(x_j - x_i), \quad \lambda \sim \mathcal{U}(0,1) \tag{3}$$

where x_i and x_j are minority-class samples. While this is often useful for tabular data, direct interpolation in pixel space is problematic for medical X-ray images because it can produce anatomically implausible structures.

To avoid that problem, SMOTE is applied here not to raw images, but to deep feature embeddings $z = f_\theta(x) \in \mathbb{R}^d$, where f_θ denotes a pretrained CNN backbone and $d \in [576, 1280]$ depending on the architecture. Synthetic embeddings are generated as

$$z_{new} = z_i + \lambda(z_j - z_i) \tag{4}$$

This creates a two-stage learning process:

- Learn a feature extractor f_θ from input images
- Apply SMOTE in the feature space to obtain balanced embeddings
- Train a classifier $g_\phi(z)$ on the augmented feature set

By operating in a semantically structured latent space, feature-space SMOTE is more likely to preserve meaningful class relationships while improving minority representation [34]. However, this benefit comes with a trade-off. The training pipeline becomes decoupled, and the method is not directly compatible with end-to-end architectures that rely on attention modules or fusion mechanisms. This limitation is important and is made explicit in this study.

3.4.3. Augmentation-Only Strategy (augmentation_only)

The third strategy addresses class imbalance solely through stochastic augmentation, without explicit resampling or loss reweighting. Let $\mathcal{T}$ denote a distribution over valid image transformations such as rotation, scaling, or intensity adjustment. Augmented samples are produced as

$$x' = T(x), \quad T \sim \mathcal{T} \tag{5}$$

where the transformation preserves the semantic class label y.

This strategy enlarges the effective support of the input distribution and encourages the model to learn more invariant representations. In this study, geometric and photometric transformations, including affine transformations and perspective distortions, are used to enrich the training set. Unlike SMOTE-based methods, augmentation-only learning works directly with physically plausible image variations and therefore maintains anatomical realism. It also preserves end-to-end training, making it compatible with all baseline, attention-enhanced, and fusion-based architectures. This is one reason why augmentation-based learning is often especially valuable in small medical imaging datasets [32].

https://doi.org/10.31436/iiumej.v27i2.4341

3.4.4. Strategy Comparison and Rationale

The three balancing strategies explored in this study reflect three different philosophies of correction:

- **Class weighting** changes the optimization objective.
- **Feature-space SMOTE** augments the learned representation space.
- **Augmentation-only** enriches the input distribution.

A deliberate design decision in this work is the exclusion of pixel-space SMOTE. Although straightforward in principle, pixel-level interpolation can generate anatomically inconsistent X-ray images, potentially distorting the very structures the model is expected to learn. In contrast, feature-space SMOTE operates on semantically meaningful embeddings, while augmentation-based learning preserves physical plausibility through valid transformations of real images. This distinction is not only technical, but methodological. In medical imaging, a balancing strategy must not merely increase minority representation. It must do so without breaking the anatomical logic of the data. The downstream consequences of these three strategies, and their interaction with model architecture, are examined further through the ablation studies presented in Section 5.

3.5. Baseline Models

A fair benchmark should not compare models that all think the same way. It should compare models that represent genuinely different design philosophies, especially when the goal is to understand not only which model performs best, but why. For that reason, five convolutional neural network architectures were selected as baseline models: EfficientNetB0, MobileNetV2, MobileNetV3Small, MNASNet, and DenseNet121. These models span a meaningful range of modern design strategies, including compound scaling, inverted residual learning, hardware-aware search, platform-aware latency optimization, and dense feature reuse. Together, they create a spectrum from compact mobile-oriented networks to deeper feature-reuse architectures, making them well-suited for studying the trade-off between diagnostic performance and computational efficiency in spinal X-ray classification.

This selection was also made with transparency in mind. During the early model consideration, NAS-inspired lightweight networks were of particular interest due to their relevance to efficient deployment. In the final benchmark, MNASNet was used as the representative NAS-based architecture. This choice was made deliberately because MNASNet provides stable pretrained support within the PyTorch and torchvision environment used in this study, enabling a consistent and reproducible implementation across experiments. Thus, MNASNet serves here as the practical latency-aware NAS-based baseline for comparison. This clarification is important because benchmark credibility depends not only on what models are selected, but also on whether those selections are technically reproducible and openly justified.

Table 4 summarizes the selected architectures in terms of embedding dimension, approximate parameter count, ImageNet Top-1 accuracy, and design philosophy. EfficientNetB0 uses compound scaling to jointly balance network depth, width, and image resolution, achieving strong predictive performance with relatively modest complexity [17]. MobileNetV2 reduces computational burden through inverted residual blocks and linear bottlenecks, making it one of the most influential lightweight backbones in efficient vision modeling [18]. MobileNetV3Small further advances this idea by combining hardware-aware neural architecture search with squeeze-and-excitation modules, producing a model optimized for low-resource environments [19]. MNASNet extends the neural architecture search

https://doi.org/10.31436/iiumej.v27i2.4341

paradigm by directly targeting inference latency on deployment hardware [20]. DenseNet121, in contrast, follows a different logic: each layer receives information from all previous layers, encouraging feature reuse and smoother gradient flow while maintaining reasonable parameter efficiency [21]. These five backbones, therefore, do more than provide variety. They allow the benchmark to test whether spinal X-ray classification benefits more from compact efficiency, richer connectivity, or a carefully balanced compromise between the two.

Table 4. Baseline CNN Architectures and Their Characteristics

Model	Torchvision Function	Embedding Dimension	Parameters (Millions)	ImageNet Top-1 Accuracy (%)	Design Philosophy
EfficientNetB0	efficientnet_b0	1280	~5.3	77.1	Compound scaling of depth, width, and resolution
MobileNetV2	mobilenet_v2	1280	~3.4	71.9	Inverted residuals with linear bottlenecks
MobileNetV3Small	mobilenet_v3_small	576	~2.5	67.7	Hardware-aware NAS with squeeze-and-excitation
MNASNet	mnasnet1_0	1280	~4.4	73.5	Platform-aware neural architecture search
DenseNet121	densenet121	1024	~8.0	74.4	Dense connectivity with feature reuse

To ensure that backbone differences, rather than classifier design, drive the comparison, all baseline models were adapted using the same transfer learning framework. Let $f_\theta(x)$ denote the pretrained backbone that maps an input image x to a feature vector $z \in \mathbb{R}^d$, where d is the embedding dimension of the selected model. A lightweight classifier head $g_\phi(z)$ is then appended for three-class classification:

$$g_\phi(z) = W_2\,\sigma\big(Dropout(W_1 z)\big) \tag{6}$$

where $W_1 \in \mathbb{R}^{128\times d}$, $W_2 \in \mathbb{R}^{3\times 128}$, and $\sigma(\cdot)$ denotes the ReLU activation function. This corresponds to the following architecture:

- `Linear`$(d \rightarrow 128)$
- `ReLU`
- `Dropout(0.3)`
- `Linear`$(128 \rightarrow 3)$

This unified classifier head serves an important methodological purpose. It removes unnecessary variation at the output stage and ensures that performance differences can be interpreted primarily as consequences of backbone behavior rather than custom head engineering. In a benchmark study, this kind of control is essential because it keeps the comparison honest.

https://doi.org/10.31436/iiumej.v27i2.4341

The training strategy follows a two-stage transfer learning paradigm. In the first stage, all backbone parameters θ are frozen, and only the classifier head parameters ϕ are optimized. This allows each model to retain the general visual representations learned from ImageNet while adapting its final decision layer for spinal X-ray classification, reducing the risk of overfitting. In the second stage, described in Section 3.7, a subset of higher-level backbone layers is unfrozen to permit domain-specific refinement. This staged approach is especially important for small medical datasets, where training a deep network end-to-end from scratch is rarely practical and often unstable [23]. In effect, the model is first taught to speak the language of the dataset, and only then allowed to refine how it listens to the anatomy.

3.6. Proposed Models

The proposed models were designed around a simple but important idea: in spinal X-ray classification, a model should not become heavier merely to become stronger. Instead, it should learn to focus better. Guided by this principle, this study develops an efficiency-aware, attention-enhanced deep learning framework that combines lightweight convolutional backbones, the Convolutional Block Attention Module (CBAM), and a dual-branch fusion strategy for three-class classification of spinal disorders. As illustrated in Figure 2 and formalized in Algorithm 1, the framework aims to improve discriminative feature learning while preserving the computational efficiency needed for realistic deployment. The central design choice is therefore not brute-force depth, but selective enhancement. Rather than adding complexity indiscriminately, the model is encouraged to attend to the most informative spinal features and, when beneficial, to combine complementary viewpoints from two lightweight backbones.

3.6.1. CBAM-Augmented Backbones

The first component of the proposed framework strengthens conventional transfer learning models by inserting an attention module between the pretrained backbone and the classifier head. The overall pipeline follows a simple sequence:

$$\text{Input Image} \rightarrow \text{Frozen Backbone} \rightarrow \text{CBAM} \rightarrow \text{GAP} \rightarrow \text{Classifier Head} \quad (7)$$

Given an input X-ray image $x \in \mathbb{R}^{3\times224\times224}$, a pretrained backbone network f_θ extracts a feature map $F \in \mathbb{R}^{C\times H\times W}$. Because the dataset is limited, the backbone is initially frozen to preserve the general visual knowledge learned from ImageNet and to reduce overfitting during early training. The extracted features are then refined by CBAM, which applies channel attention followed by spatial attention. This ordering matters. The model is first encouraged to decide **which** features are important, and only then to determine **where** the relevant evidence lies in the image. In spinal radiographs, this is particularly valuable because pathological evidence is often subtle, localized, and easily diluted within a larger anatomical field.

The *channel attention module* aggregates global context using both average pooling and max pooling, followed by a shared multilayer perceptron and sigmoid activation:

$$M_c = \sigma\left(MLP\big(AvgPool(F)\big) + MLP\big(MaxPool(F)\big)\right) \quad (8)$$

The resulting channel-refined feature map is computed as $F_c = M_c \odot F$, where $\odot$ denotes element-wise multiplication. This mechanism helps the network amplify diagnostically informative channels while suppressing less relevant responses.

Next, the *spatial attention module* identifies the image regions that deserve emphasis. It applies channel-wise pooling and then a convolution with a 7×7 kernel:

https://doi.org/10.31436/iiumej.v27i2.4341

$$M_s = \sigma\left(\text{Conv}_{7\times7}([\text{AvgPool}(F_c); \text{MaxPool}(F_c)])\right) \quad (9)$$

The final refined feature map is then obtained as $F' = M_s \odot F_c$. This sequential attention process allows the model to emphasize both globally important features and spatially localized patterns, such as abnormal spinal curvature, vertebral misalignment, or subtle radiographic distortions associated with pathology.

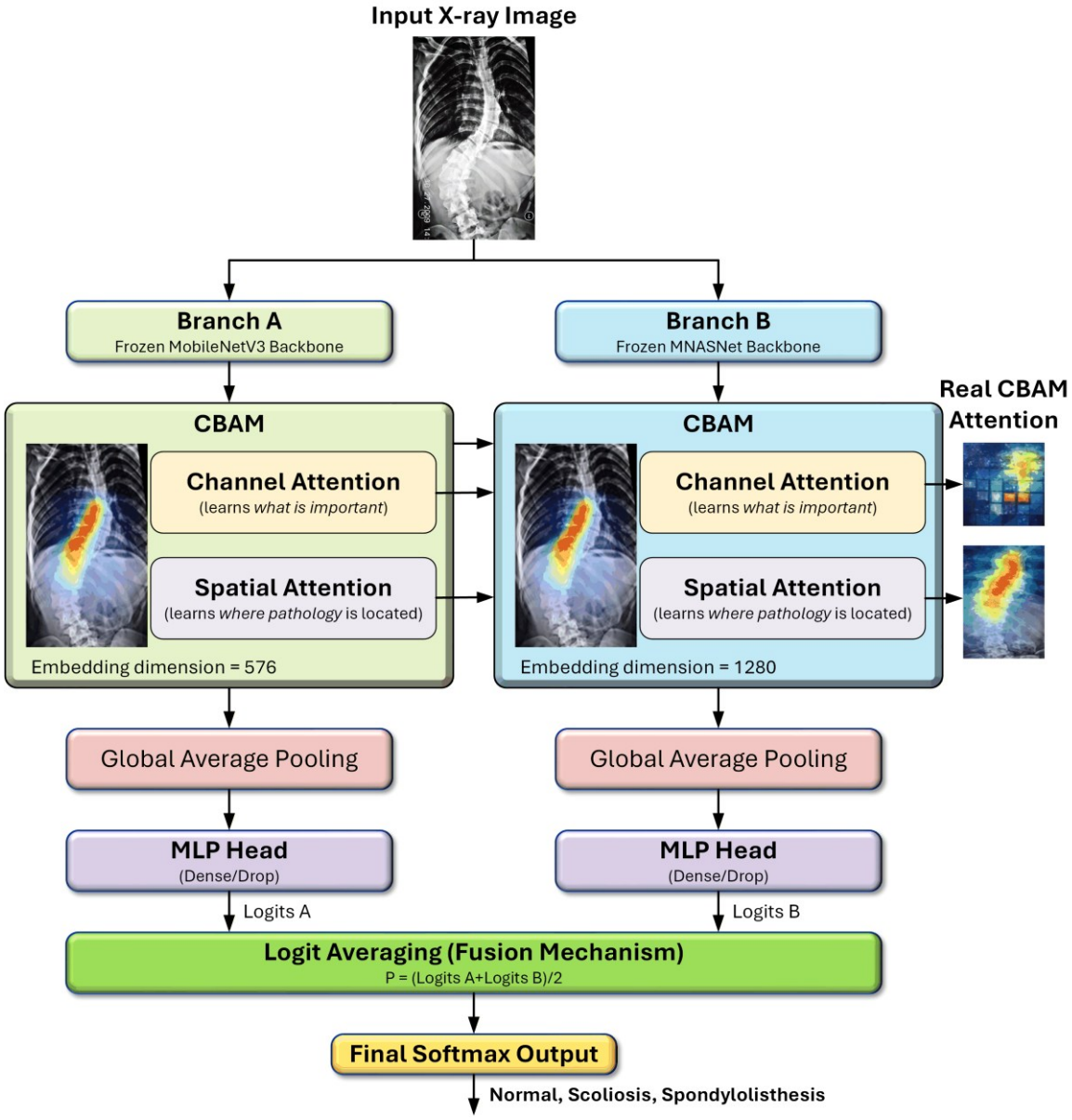


Figure 2. Proposed CBAM-Enhanced Dual-Branch Fusion Model for Multi-Class Spinal X-Ray Classification

https://doi.org/10.31436/iiumej.v27i2.4341

Algorithm 1. Inference Procedure of the Proposed CBAM-Based Dual-Branch Fusion Framework

Algorithm 1 Forward Pass of the Proposed Dual-Branch CBAM-Enhanced Fusion Model

Require: Input spine X-ray image $x \in \mathbb{R}^{3\times224\times224}$
Require: Frozen pretrained backbones f_A and f_B
Require: CBAM modules CBAM_A and CBAM_B
Require: Classifier heads g_A and g_B
Ensure: Predicted probability vector $\hat{\mathbf{y}} \in \mathbb{R}^3$ for {Normal, Scoliosis, Spondylolisthesis}

Branch A: CBAM-MobileNetV3Small
$F_A \leftarrow f_A(x)$ ▷ $F_A \in \mathbb{R}^{C_A\times H_A\times W_A}$
$F'_A \leftarrow \text{CBAM}(F_A)$ ▷ Attention-refined feature map
$\mathbf{z}_A \leftarrow \text{GAP}(F'_A)$ ▷ $\mathbf{z}_A \in \mathbb{R}^{576}$
$\mathbf{l}_A \leftarrow g_A(\mathbf{z}_A)$ ▷ $\mathbf{l}_A \in \mathbb{R}^3$

Branch B: CBAM-MNASNet
$F_B \leftarrow f_B(x)$ ▷ $F_B \in \mathbb{R}^{C_B\times H_B\times W_B}$
$F'_B \leftarrow \text{CBAM}(F_B)$ ▷ Attention-refined feature map
$\mathbf{z}_B \leftarrow \text{GAP}(F'_B)$ ▷ $\mathbf{z}_B \in \mathbb{R}^{1280}$
$\mathbf{l}_B \leftarrow g_B(\mathbf{z}_B)$ ▷ $\mathbf{l}_B \in \mathbb{R}^3$

Fusion by Logit Averaging
$\mathbf{l}_{\text{fuse}} \leftarrow 0.5 \cdot \mathbf{l}_A + 0.5 \cdot \mathbf{l}_B$

Final Prediction
$\hat{\mathbf{y}} \leftarrow \text{Softmax}(\mathbf{l}_{\text{fuse}})$
return $\hat{\mathbf{y}}$

function $\text{CBAM}(F)$
  $M_c \leftarrow \sigma\Big(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))\Big)$ ▷ Channel attention map
  $F_c \leftarrow M_c \odot F$ ▷ Channel-refined feature map
  $M_s \leftarrow \sigma\Big(\text{Conv}_{7\times7}([\text{AvgPool}(F_c); \text{MaxPool}(F_c)])\Big)$ ▷ Spatial attention map
  $F_s \leftarrow M_s \odot F_c$ ▷ Final refined feature map
  return F_s
end function

function $g(\mathbf{z})$
  $\mathbf{h} \leftarrow \text{ReLU}(W_1\mathbf{z} + \mathbf{b}_1)$
  $\tilde{\mathbf{h}} \leftarrow \text{Dropout}(\mathbf{h}, p = 0.3)$
  $\mathbf{l} \leftarrow W_2\tilde{\mathbf{h}} + \mathbf{b}_2$
  return $\mathbf{l}$
end function

Notation: f_A denotes the frozen MobileNetV3Small backbone and f_B denotes the frozen MNASNet backbone. $GAP(\cdot)$ is global average pooling, $\sigma(\cdot)$ is the sigmoid activation, and $\odot$ denotes element-wise multiplication. The classifier heads g_A and g_B share the same multilayer perceptron structure: Linear($d \rightarrow 128$) → ReLU → Dropout(0.3) → Linear($128 \rightarrow 3$).

After attention refinement, global average pooling converts the feature map into a compact vector $z \in \mathbb{R}^C$, where $C \in [576, 1024, 1280]$ depends on the backbone. A lightweight classifier head then produces the class logits. To keep the comparison controlled and fair, the same head design is used across all attention-enhanced models. In this study, CBAM is integrated into three backbones: MobileNetV3Small, MNASNet, and EfficientNetB0. These models were selected because they offer different trade-offs between compactness and

representational strength. The goal is not simply to test whether CBAM improves performance, but to examine under which architectural conditions attention becomes most useful.

3.6.2. Dual-Branch Fusion Model

Attention helps a model focus. Fusion helps it hesitate less. To further improve robustness and generalization, this study proposes a dual-branch fusion architecture named **Fusion_MNv3_MNAS**. As shown in Figure 2 and described in Algorithm 1, the same input image is processed through two parallel CBAM-enhanced branches:

- **Branch A**: CBAM-MobileNetV3Small, which provides a compact and efficient representation with $d = 576$.
- **Branch B**: CBAM-MNASNet, which provides a richer latency-aware representation with, $d = 1280$.

Each branch independently produces a logit vector:

$$\mathbf{l}_A = f_1(x), \mathbf{l}_B = f_2(x), \mathbf{l}_A, \mathbf{l}_B \in \mathbb{R}^3 \tag{10}$$

The final prediction is obtained through *logit-level fusion*:

$$\hat{\mathbf{y}} = \text{Softmax}\left(\frac{1}{2}\mathbf{l}_A + \frac{1}{2}\mathbf{l}_B\right) \tag{11}$$

This design is motivated by a simple insight from ensemble learning: when two competent learners make different kinds of mistakes, combining them can reduce variance and improve reliability. Here, the two branches are intentionally lightweight but not redundant. MobileNetV3Small favors compact, aggressively efficient feature extraction, whereas MNASNet contributes a somewhat richer representation shaped by latency-aware neural architecture search. Their diversity is therefore functional rather than cosmetic. The fusion model is not merely larger than either branch alone; it is more balanced because it merges two different perspectives on the same anatomy.

This distinction is important for the framework's novelty. The contribution is not the isolated use of CBAM or fusion, both of which have appeared in other domains. Rather, the contribution lies in combining lightweight backbones, attention refinement, and dual-branch fusion within a unified spinal X-ray classification framework that is explicitly evaluated under efficiency-aware criteria. In other words, the architecture is designed not just to score well, but to reveal how complementary lightweight models can cooperate without drifting into the cost profile of a heavy network.

From an efficiency standpoint, the proposed fusion model maintains a favorable parameter-performance balance and remains lighter than heavier baselines such as DenseNet121. Unlike conventional ensembles that require separately trained models and duplicated inference workflows, the proposed architecture integrates both branches into a single unified framework, allowing efficient parallel feature extraction and coordinated prediction. By embedding CBAM within each backbone, the model becomes more selective about the spinal patterns it amplifies. By combining two heterogeneous lightweight branches, it becomes more robust to representation bias and less dependent on a single architectural viewpoint. The resulting design, therefore, aims to achieve what is often difficult in medical AI: improve discrimination, strengthen robustness, and preserve practical deployment.

3.7. Training Protocol

Training on a small medical imaging dataset requires restraint as much as ambition. If the model is adapted too little, it may remain tied to generic ImageNet features that are not

sufficiently sensitive to spinal anatomy. If it is adapted too aggressively, it may overfit the limited training data and lose the stability gained from pretraining. To navigate this tension, this study adopts a two-phase training protocol that combines frozen feature extraction with selective fine-tuning. The strategy is designed to preserve the broad visual knowledge already learned by the pretrained backbone while gradually allowing the model to specialize in spinal X-ray interpretation. In this sense, the model is not forced to relearn vision from the beginning. It is first grounded, then refined.

3.7.1. Phase 1: Frozen Backbone Training

In the first phase, all backbone parameters θ were frozen, and only the classifier head parameters ϕ were optimized. This allows the model to immediately leverage the generic visual representations learned from large-scale ImageNet pretraining while minimizing the risk of overfitting to the relatively small spinal X-ray dataset. The model was trained for 50 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4}. The optimization objective can be expressed as:

$$\phi^* = \arg\min_{\phi} \mathcal{L}\big(g_\phi\big(f_\theta(x)\big), y\big), \quad \theta \; fixed \tag{12}$$

This phase has a practical purpose: it teaches the classifier head to interpret the backbone features before the backbone itself is allowed to adapt. For small medical datasets, this is often the safest starting point because it prevents the network from immediately drifting toward unstable domain-specific fitting.

3.7.2. Phase 2: Partial Fine-Tuning

After initial convergence, a second training phase was introduced to adapt the higher-level feature representations to the target domain. In the main benchmark protocol, the top 20% of backbone layers were unfrozen, while the lower-level layers remained fixed to preserve generic feature extraction capabilities. The model was then trained for an additional 30 epochs using a reduced learning rate of 1×10^{-5}, allowing gradual and controlled refinement of domain-relevant features:

$$(\theta', \phi') = \arg\min_{\theta', \phi} \mathcal{L}\big(g_\phi\big(f_{\theta'}(x)\big), y\big) \tag{13}$$

where $\theta' \subset \theta$ denotes the subset of unfrozen parameters.

This staged optimization helps reduce catastrophic forgetting while enabling the network to capture finer anatomical cues that are specific to spinal radiographs. It is especially important in this study because the abnormalities of interest, such as mild curvature or vertebral displacement, may be too subtle to be fully represented by generic pretrained features alone.

For transparency, the 20% unfreezing fraction was used as the **default benchmark setting** across models to maintain a uniform and fair experimental protocol during the main comparative evaluation. A later ablation study in Section 5.2 showed that, for some models, especially CBAM-MobileNetV3Small, unfreezing approximately 30% of the backbone yields slightly better adaptation. This does not invalidate the main benchmark. Rather, it shows that the globally fixed protocol used for fair comparison is not always identical to the model-specific optimum discovered through later analysis. Accordingly, the benchmark results should be interpreted as results under a controlled common protocol, while the ablation study identifies the more refined task-specific optimum. This clarification directly addresses the potential ambiguity between the benchmark default and the later fine-tuning analysis.

3.7.3. Regularization and Training Control

To improve generalization and reduce the risk of overfitting, three training control mechanisms were incorporated:

- *Early stopping* with a patience of 10 epochs, monitored on validation loss, to terminate training when no further improvement is observed.
- *Adaptive learning rate scheduling* using `ReduceLROnPlateau`, which reduces the learning rate by a factor of 0.5 when validation loss fails to improve for 5 consecutive epochs.
- *Model checkpointing*, in which the model state achieving the best validation loss is saved to preserve the strongest generalizing configuration.

These mechanisms were included not simply to stabilize optimization, but to ensure that performance differences among models reflect learning quality rather than accidental overtraining. In a benchmark study involving multiple lightweight architectures, such control is essential for fair comparison.

3.8. Cross-Validation Strategy

A small dataset can produce impressively high numbers for the wrong reasons. One favorable train-test split may overestimate a model's true performance, while another may unfairly underestimate it. To avoid that problem, this study adopts a *stratified 5-fold cross-validation protocol* over the full dataset of 338 subjects. The purpose is not only to increase statistical reliability, but also to ensure that the reported results reflect repeated and balanced evaluation rather than a single lucky partition. Stratification preserves the original class distribution within each fold, so that all three diagnostic categories remain proportionally represented during both training and evaluation.

Let the dataset be denoted as $\mathcal{D} = [(x_i, y_i)]_{i=1}^{N}$, where $N = 338$. The data are partitioned into $K = 5$ mutually exclusive folds $[\mathcal{D}]_{k=1}^{5}$, such that:

$$\mathcal{D} = \bigcup_{k=1}^{5} \mathcal{D}_k \, , \mathcal{D}_i \cap \mathcal{D}_j = \emptyset \text{ for } i \neq j \tag{14}$$

At each iteration k, one fold $\mathcal{D}_k$ is reserved for evaluation, while the remaining four folds are used for model development:

$$\mathcal{D}_{\text{train}}^{(k)} = \mathcal{D} \setminus \mathcal{D}_k, \quad \mathcal{D}_{\text{val}}^{(k)} = \mathcal{D}_k \tag{15}$$

Within each iteration, the development portion is further internally stratified into *80% training and 20% validation* subsets. This additional split supports model selection, learning-rate adjustment, and early stopping without leaking information from the held-out evaluation fold. As a result, the evaluation remains separated from optimization, which is crucial for preserving the credibility of the final performance estimates.

Performance is aggregated across all folds using the mean and standard deviation:

$$\bar{M} = \frac{1}{K} \sum_{k=1}^{K} M^{(k)} \, , \sigma_M = \sqrt{\frac{1}{K} \sum_{k=1}^{K} \left(M^{(k)} - \bar{M} \right)^2} \tag{16}$$

where $M^{(k)}$ denotes the evaluation metric for fold k, such as accuracy or F1-score. Reporting results as $\bar{M} \pm \sigma_M$ serves two purposes at once. The mean captures the central performance level, while the standard deviation reveals how stable that performance remains across data

https://doi.org/10.31436/iiumej.v27i2.4341

partitions. In a dataset of this size, both are necessary. A high mean without stability can be misleading, whereas a slightly lower mean with low variability may represent the more trustworthy model. This is why cross-validation in this study is not merely a procedural choice. It is part of the argument for credibility.

3.9. Evaluation Metrics

A model that is accurate but impractical is difficult to trust in a real clinical pipeline. Conversely, a model that is fast but diagnostically weak is of little use to clinicians. For that reason, this study adopts a dual evaluation framework that treats performance from two angles at once: *effectiveness*, which reflects predictive quality, and *efficiency*, which reflects computational cost. This distinction is central to the present work. The objective is not merely to identify the model with the highest score, but to determine which model offers the most credible balance between diagnostic capability and deployment feasibility. In this sense, evaluation is not the last step of the framework. It is part of the paper's argument.

3.9.1. Effectiveness Metrics

Model effectiveness is evaluated using standard classification metrics derived from the confusion matrix. For each class c, let TP_c, FP_c, FN_c, and TN_c denote the numbers of true positives, false positives, false negatives, and true negatives, respectively.

The overall accuracy is defined as:

$$\text{Accuracy} = \frac{\sum_c TP_c}{\sum_c (TP_c + FP_c + FN_c + TN_c)} \tag{17}$$

Accuracy provides a useful global summary of correctness, but on an imbalanced dataset, it can also be misleading. A model may achieve a high overall score simply by favoring the majority class, even while underperforming on clinically important minority classes. For that reason, this study gives particular importance to *weighted metrics* that account for class distribution and therefore offer a more faithful view of model behavior across all diagnostic categories.

The class-wise *precision*, *recall*, and *F1-score* are defined as:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \quad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \quad \text{F1}_c = \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{18}$$

The *weighted averages* are computed as:

$$\text{Metric}_{\text{weighted}} = \sum_{c=1}^{C} w_c \cdot \text{Metric}_c, w_c = \frac{N_c}{N} \tag{19}$$

where N_c is the number of samples in class c, and N is the total number of samples. This weighting ensures that each class contributes in proportion to its presence in the dataset, thereby reducing bias toward the majority class.

In this study, weighted precision, weighted recall, weighted F1-score, and accuracy are reported consistently across folds, with **means ± standard deviations**. This consistency is important because a metric without variability can appear more certain than it truly is. Reporting dispersion across folds allows the reader to judge not only how well a model performs on average but also how stable that performance is across different data partitions.

To provide a more concrete view of model behavior, confusion matrices are also reported. They reveal where the model succeeds, where it struggles, and which classes are most commonly confused. This is particularly useful in spinal X-ray classification, where mild scoliosis, borderline normal cases, or subtle vertebral displacement may create ambiguity that

summary scores alone cannot fully explain. In addition, *Receiver Operating Characteristic* (ROC) curves and the corresponding *Area Under the Curve* (AUC) are computed using a one-vs-rest strategy. The ROC curve traces the true positive rate against the false positive rate across decision thresholds, while the AUC summarizes the model's class-separation ability independent of any single threshold. Together, these metrics provide a fuller picture of diagnostic discrimination.

3.9.2. Efficiency Metrics

Predictive performance alone does not determine whether a model is useful in practice. In clinical deployment, a model must also be compact enough to store, fast enough to respond, and efficient enough to run within realistic hardware constraints. For this reason, computational efficiency is evaluated systematically alongside classification quality.

Three structural efficiency metrics are considered:

- *Parameter Count*, including both total and trainable parameters, to quantify model complexity and memory demand.
- *Floating Point Operations (FLOPs)*, computed using the `thop` library, to estimate the computational burden of inference.
- *Model Size*, measured in megabytes (MB) based on the serialized `state_dict`, to reflect storage requirements.

In addition to structural cost, runtime behavior is evaluated using two complementary inference metrics:

- *Latency*, defined as the time required to process a single input sample. Latency is computed as the mean inference time over T repeated runs after an initial warm-up phase:

$$\text{Latency} = \frac{1}{T}\sum_{t=1}^{T} t_{\text{inference}}^{(t)} \tag{20}$$

- *Throughput*, defined as the number of images processed per second:

$$\text{Throughput} = \frac{\text{Number of images}}{\text{Total inference time}} \tag{21}$$

To improve measurement stability, latency is recorded over 50 runs with 5 warm-up iterations to reduce transient GPU effects. All runtime measurements are performed under the same hardware conditions and with batch size 1, so that the evaluation reflects a realistic clinical screening scenario in which patient scans are processed individually rather than in large offline batches.

This efficiency analysis is not included as a secondary supplement. It is integral to the study's central claim. A model with marginally higher accuracy may not represent the better solution if it requires substantially more parameters, more memory, and more inference time. By reporting both predictive effectiveness and computational efficiency side by side, this study aims to identify models that are not only theoretically strong but also credible for real-world deployment in spinal disorder screening.

4. RESULTS AND DISCUSSION

The value of a medical AI model is revealed not only by the score it reaches, but by the trade-offs it makes to get there. For that reason, this section evaluates the proposed framework from two connected perspectives: effectiveness and efficiency. All results are reported using

stratified 5-fold cross-validation as the mean ± standard deviation, so that performance is judged not only by its absolute value but also by its stability across data partitions. This distinction matters for a dataset of 338 subjects, where a single favorable split can make a model appear more certain than it truly is. The discussion, therefore, moves beyond reporting which model scored highest. It examines which models remain credible under repeated evaluation, which architectures use computation most wisely, and where the remaining classification errors still concentrate.

4.1. Baseline Comparison

Table 5 compares the five baseline architectures under the `class_weight strategy`. The first important finding is that the top three baselines are essentially tied in predictive performance. MobileNetV3Small achieves a weighted F1-score of 0.962±0.022, while MobileNetV2 and EfficientNetB0 each achieve 0.961 with comparable variability. These differences are too small to support any strong claim of superiority based solely on accuracy. In practical terms, the baseline comparison does not identify one clearly dominant scorer. Instead, it shifts the question to something more meaningful: when predictive performance is nearly the same, which model earns that performance with the least computational burden?

Table 5. Performance and Efficiency Comparison of Baseline Models under Class-Weighted Strategy (5-fold stratified cross-validation, mean ± standard deviation)

Model	Accuracy	Precision (w)	Recall (w)	F1 (w)	Params (M)	FLOPs (M)	Size (MB)	Latency (ms)	Throughput (img/s)
EfficientNetB0	0.962 ± 0.022	0.962 ± 0.023	0.962 ± 0.022	0.961 ± 0.023	4.172	414.029	15.915	18.73	53.391
MobileNetV2	0.962 ± 0.027	0.962 ± 0.027	0.962 ± 0.027	0.961 ± 0.027	2.388	326.371	9.11	12.16	82.236
MobileNetV3Small	**0.962 ± 0.022**	**0.963 ± 0.022**	**0.962 ± 0.022**	**0.962 ± 0.022**	1.001	60.938	3.819	15.916	62.831
MNASNet	0.902 ± 0.030	0.923 ± 0.021	0.902 ± 0.030	0.905 ± 0.029	3.267	335.127	12.461	11.472	87.169
DenseNet121	0.935 ± 0.037	0.941 ± 0.034	0.935 ± 0.037	0.936 ± 0.036	7.085	2896.115	27.029	39.426	25.364

From that perspective, MobileNetV3Small stands out as the most persuasive baseline. It reaches the highest weighted F1-score while using only 1.00 million parameters, 60.94 million FLOPs, and a model size of 3.82 MB. DenseNet121, by contrast, requires 7.09 million parameters, 2896.12 million FLOPs, and 27.03 MB, yet achieves a lower weighted F1-score of 0.936±0.036. This is not a marginal efficiency advantage. It is a strong and unexpected reversal of the common assumption that larger networks naturally yield better results. In this task, the heaviest model is not the strongest. The result suggests that, when only about 270 training images are available per fold, representational excess does not translate into better generalization. What matters more is whether the architecture is well matched to the scale and structure of the data.

MobileNetV2 and EfficientNetB0 also perform strongly, but their efficiency profiles differ. MobileNetV2 offers the lowest latency among the top three baselines at 12.16 ms and the highest throughput at 82.24 images per second, making it especially attractive when rapid inference is prioritized. EfficientNetB0 reaches comparable predictive performance, but at a notably higher structural cost. This places MobileNetV3Small, MobileNetV2, and EfficientNetB0 in three distinct yet meaningful positions: MobileNetV3Small as the most compact, high-performing baseline; MobileNetV2 as the fastest, high-performing baseline; and

https://doi.org/10.31436/iiumej.v27i2.4341

EfficientNetB0 as the strongest, largest, lightweight baseline. Seen together, these results show that baseline selection in spinal X-ray classification should not be framed as a race for a single winner. It is more accurately a decision about which kind of efficiency-performance balance best fits the deployment objective.

MNASNet presents a different story. Although designed through neural architecture search with latency-aware objectives, it achieves only 0.905±0.029 in weighted F1-score, roughly 5 percentage points below the top baselines. Its inference speed is the fastest in the comparison, with a throughput of 87.17 images per second, but that advantage is not enough to offset the loss in diagnostic performance. In a clinical screening scenario, speed is valuable only if it does not come at too high a predictive cost. The result suggests that features optimized for generic mobile efficiency do not automatically transfer well to spinal radiographic morphology. In other words, being fast is not the same as being clinically sharp.

The comparison with prior studies further clarifies the contribution of the present benchmark. Earlier work on the same 338-subject dataset reported a mean three-class accuracy of 96.73% and a peak accuracy of 98.02% in [1], while [5] reported 99.01% accuracy. At first glance, these numbers may appear slightly higher than the cross-validation averages reported here. However, the comparison is not like-for-like. Those studies relied on fixed train-test splits, whereas the present work reports mean ± standard deviation under stratified 5-fold cross-validation. This difference is crucial. A single split can produce an impressive headline result, but it does not reveal how sensitive the model is to data partitioning. By contrast, the cross-validation protocol used here provides a more conservative and more credible estimate of generalization. Thus, the contribution of Table 4 is not merely another set of scores. It is a stronger answer to a more demanding question: not how well a model performs once, but how well it continues to perform as the evaluation becomes harder to game.

4.2. Proposed Models versus Baselines

Table 6 compares the proposed CBAM-enhanced and fusion-based models against the strongest baseline, MobileNetV3Small, under two training regimes: `class_weight` and `augmentation_only`. The central result is clear. Under `class_weight`, the proposed attention-enhanced models do not outperform the baseline. CBAM-MobileNetV3Small achieves a weighted F1-score of 0.951±0.024, which is slightly below the baseline MobileNetV3Small at 0.962±0.022. The same pattern appears in CBAM-MNASNet and, to a lesser extent, in CBAM-EfficientNetB0. Although these differences are modest and remain within approximately one standard deviation, their direction is consistent across architectures. This consistency is important because it shows that attention does not automatically translate into better generalization when the training data are limited.

This result is scientifically more useful than a simple improvement would have been. It shows that CBAM should not be treated as a universal performance booster. Under limited-data conditions, attention increases representational flexibility, but that added flexibility is only valuable if the training regime provides sufficient diversity to regularize it. When each fold contains only about 270 training samples, the model may gain more expressive capacity than the available supervision can reliably support. In that setting, the issue is not that attention fails in principle. Rather, the model becomes more adaptable than the data can safely guide.

The behavior of the fusion model reinforces the same point. Under `class_weight`, Fusion_MNv3_MNAS achieves a weighted F1-score of 0.951±0.033, again below the strongest baseline. Its performance converges toward that of its better branch, CBAM-MobileNetV3Small, rather than exceeding it. This behavior is consistent with the role of logit-

https://doi.org/10.31436/iiumej.v27i2.4341

level fusion. Fusion can reduce variance and soften the effect of a weaker branch, but it cannot create discriminative information that the branches themselves have not learned. Here, the weaker CBAM-MNASNet branch is partly compensated for, yet the overall gain remains limited because both branches are trained under the same restricted supervision regime. The result is therefore not surprising. Fusion improves reliability only when the underlying branch representations are sufficiently informative and sufficiently diverse.

Table 6. Proposed Models versus the Baseline under the `class_weight` and augmentation_only strategy

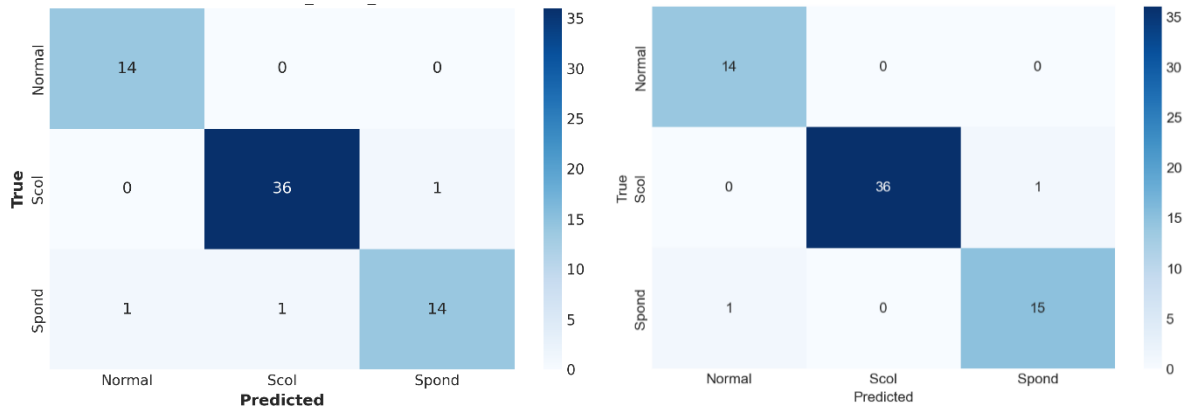
Model	`class_weight`		`augmentation_only`		Params (M)	Size (MB)	FLOPs (M)
	Accuracy	F1 (w)	Accuracy	F1 (w)			
MobileNetV3Small	0.962 ± 0.022	0.962 ± 0.022	0.953 ± 0.025	0.952 ± 0.026	1.001	3.819	61
CBAM_MobileNetV3Small	0.950 ± 0.024	0.951 ± 0.024	0.956 ± 0.026	0.955 ± 0.027	1.043	3.978	61
CBAM_MNASNet	0.891 ± 0.031	0.895 ± 0.028	0.917 ± 0.038	0.920 ± 0.036	3.472	13.243	336
CBAM_EfficientNetB0	0.959 ± 0.019	0.958 ± 0.019	0.947 ± 0.037	0.945 ± 0.038	4.377	16.696	414
Fusion_MNv3_MNAS	0.950 ± 0.034	0.951 ± 0.033	**0.976 ± 0.020**	**0.976 ± 0.020**	4.514	17.221	397

The performance changes under `augmentation_only`, and this is where the proposed framework becomes most revealing. Fusion_MNv3_MNAS reaches the highest weighted F1-score in the study, 0.976±0.020, clearly surpassing both the baseline MobileNetV3Small and all other proposed variants. This shift is the key empirical message of Table 5. The proposed models are not universally superior across all training conditions, but they become strongest when the training pipeline provides greater data diversity. Augmentation does more than rebalance the learning objective. It exposes the model to a broader range of anatomically plausible variations, thereby regularizing the additional parameters introduced by attention and fusion. Under this regime, the extra capacity of CBAM and dual-branch fusion is no longer a liability. It becomes useful.

This interaction between model design and training strategy is the main methodological insight of this section. Architecture alone does not determine performance in spinal X-ray classification. The same model can underperform under one regime and become the strongest performer under another. The implication is important: CBAM and fusion should not be evaluated in isolation, but in relation to the regularization environment in which they are trained. In the present study, augmentation-only training provides that environment more effectively than class weighting.

Figure 3 provides class-level evidence that supports this interpretation. The confusion matrices show that the remaining errors are concentrated mainly at the Normal-Scoliosis boundary. This is clinically plausible because mild scoliosis can resemble normal anatomical variation when curvature is subtle. By contrast, Spondylolisthesis is classified more consistently across models, indicating that its structural features are more visually distinct. The fusion model under `augmentation_only` reduces confusion at the Normal-Scoliosis boundary more effectively than the baseline under `class_weight`, suggesting improved sensitivity to subtle curvature-related patterns. Thus, the value of the proposed model is not merely that it raises an overall score, but that it improves discrimination in the most ambiguous part of the problem.

https://doi.org/10.31436/iiumej.v27i2.4341



(a) MobileNetV3Small under class_weight (b) Fusion_MNv3_MNAS under augmentation_only

Figure 3. Confusion Matrices (fold 5), in which residual errors concentrate on the Normal-Scoliosis boundary.

4.3. Efficiency Analysis

Figures 4 and 5 show that efficiency is not a side result in this study. It is part of the main result. When several models achieve similar predictive performance, the more important question becomes how much computation, memory, and time must be spent to obtain that performance. Viewed from this perspective, the trade-off space is highly structured rather than arbitrary.

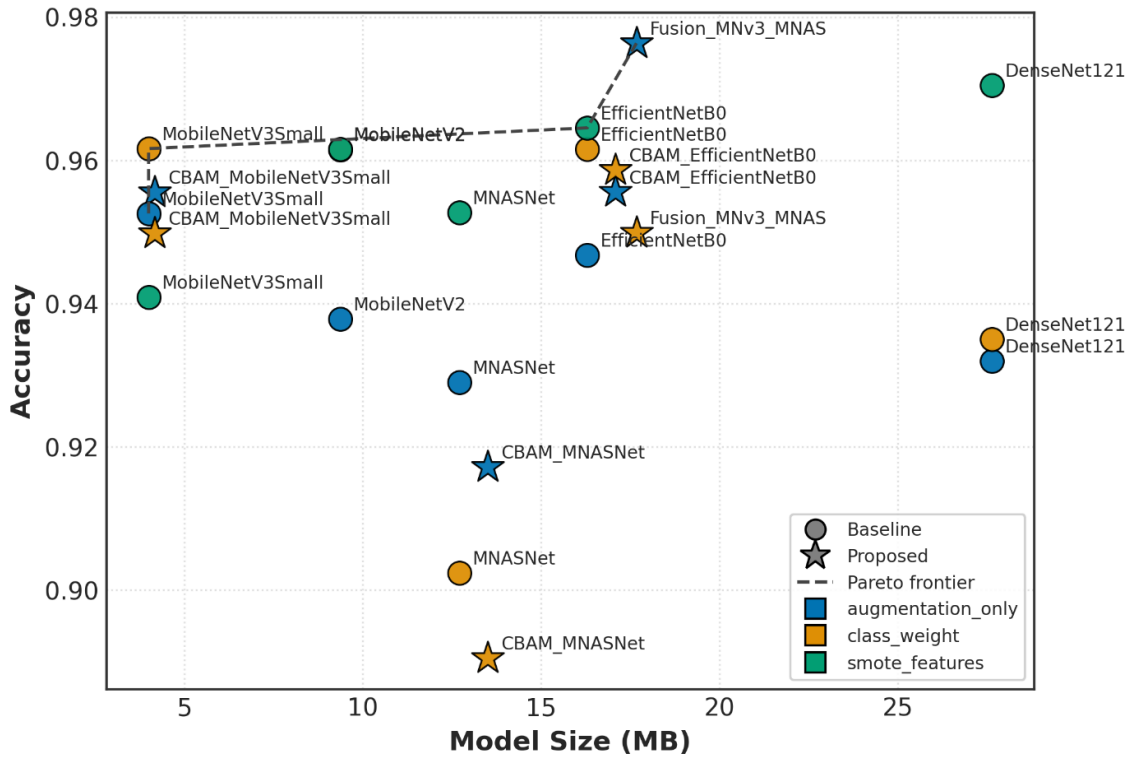


Figure 4. Model size (MB) versus accuracy for all evaluated models across three balancing strategies. The Pareto frontier highlights the smallest strong baseline, the intermediate trade-off region, and the highest-accuracy fusion configuration.

Figure 4 plots *accuracy versus model size* and reveals a compact Pareto frontier. At the smallest end of the spectrum, MobileNetV3Small remains the most storage-efficient strong

https://doi.org/10.31436/iiumej.v27i2.4341

baseline. With a model size of only 3.82 MB, it already achieves accuracy competitive with the best baseline results. At the other end, Fusion_MNv3_MNAS under `augmentation_only` achieves the highest accuracy in the study, but this gain comes at a substantially larger footprint of 17.22 MB. Between these two points, EfficientNetB0 occupies an intermediate position, offering a modest accuracy gain over the smallest model while increasing in size. The remaining models fall below this frontier and are therefore Pareto-inferior, meaning that they are larger without providing a compensating improvement in predictive performance. This is especially clear for DenseNet121, which is much larger than MobileNetV3Small yet does not deliver better accuracy on this dataset. The implication is direct: for small medical imaging datasets, more parameters do not automatically buy better generalization.

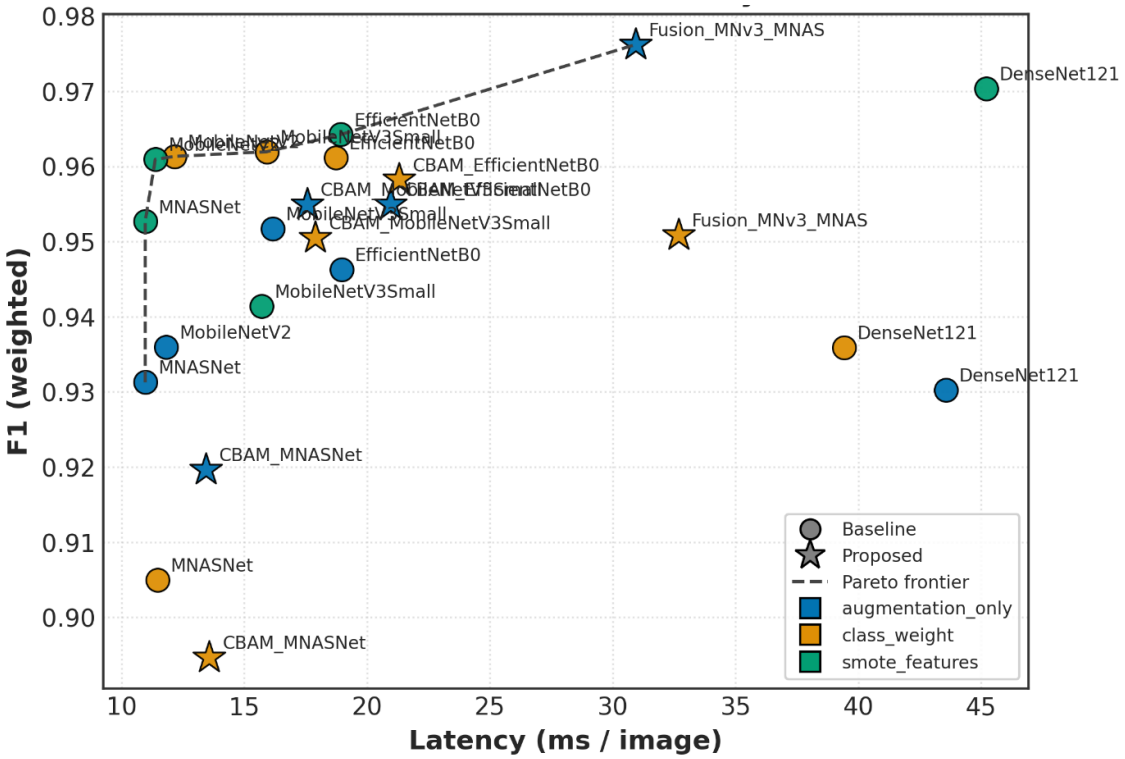


Figure 5. Inference latency (ms/image) versus weighted F1-score for all evaluated models across three balancing strategies. Lower latency is better. The figure highlights the trade-off between rapid inference and maximum predictive performance.

Figure 5 complements this view by plotting *weighted F1-score versus inference latency*, which is more relevant to deployment scenarios in which scans are processed individually. The figure shows that the best model depends on the operational constraint. If latency is the main concern, the most attractive region lies toward the left side of the plot, where lightweight baselines such as MobileNetV3Small and MobileNetV2 achieve strong weighted F1 Scores with relatively low inference times. MNASNet achieves the lowest overall latency, but its lower weighted F1-score makes that speed advantage less attractive in a clinical screening context. At the opposite end, Fusion_MNv3_MNAS under `augmentation_only` achieves the highest weighted F1-score, but with a clear latency penalty. This does not weaken the model's value. It simply defines its appropriate use case more precisely.

Taken together, Figures 4 and 5 show that no single model is optimal across all deployment scenarios. For memory-constrained or latency-sensitive settings, MobileNetV3Small remains

https://doi.org/10.31436/iiumej.v27i2.4341

the most practical choice because it combines strong predictive performance with very low storage cost. By contrast, when predictive performance is the dominant objective and a larger latency budget is acceptable, Fusion_MNv3_MNAS under `augmentation_only` becomes the preferred configuration because it extends the frontier upward. The broader lesson is that model selection in spinal X-ray classification should not be framed as a search for one universally best architecture. It should be framed as a matching problem between predictive ambition and operational constraint.

4.4. Effect of Class Balancing Strategy

Table 7 shows that model performance in this study is shaped not only by architecture, but also by how class imbalance is handled during training. This is a central result of the paper. Among the three strategies, `augmentation_only` produces the highest overall performance, with Fusion_MNv3_MNAS achieving a weighted F1-score of 0.976±0.020, the best result in the entire experimental matrix. More importantly, this is the only regime in which the proposed CBAM-enhanced and fusion-based models consistently outperform their baseline counterparts. The implication is clear: attention and fusion do not become effective simply because they are added to the architecture. They become effective when the training pipeline supplies enough data diversity to regularize the added representational capacity.

Table 7. Best-performing model under each class balancing strategy (5-fold stratified cross-validation, mean ± standard deviation)

Strategy	Best Model	Accuracy	F1 (weighted)	Key Characteristic
class_weight	MobileNetV3Small	0.962 ± 0.022	0.962 ± 0.022	Weighted cross-entropy, no synthetic data
smote_features	DenseNet121	0.971 ± 0.010	0.970 ± 0.011	SMOTE on 1024-D embeddings
augmentation_only	Fusion_MNv3_MNAS	0.976 ± 0.020	0.976 ± 0.020	Strong augmentation (geometric + photometric)

This helps explain why `augmentation_only` performs best. Class weighting changes the optimization objective, but it does not increase the visual diversity of the training data. Augmentation does. By exposing the model to a broader range of anatomically plausible geometric and photometric variation, it acts as a form of data-driven regularization. Under this regime, the CBAM module is better able to learn meaningful channel-spatial relationships instead of fitting too closely to a limited set of images. In that sense, augmentation does more than mitigate imbalance. It improves the match between the model's flexibility and the diversity of the supervision signal.

By contrast, `class_weight` provides a strong and stable baseline rather than the highest score. MobileNetV3Small achieves a weighted F1-score of 0.962±0.022 under this strategy, indicating that loss reweighting alone is sufficient to achieve competitive performance. The difference between the best `class_weight` result and the best `augmentation_only` result is approximately 1.4 percentage points in weighted F1-score. Given the cross-validation variability observed in this study, this improvement is consistent but not strongly separated at the present dataset scale. From a practical perspective, this makes `class_weight` with MobileNetV3Small the most conservative and deployment-ready configuration: it is simple, reproducible, lightweight, and strong enough to remain highly competitive without requiring a more aggressive training pipeline.

https://doi.org/10.31436/iiumej.v27i2.4341

The `smote_features` strategy produces a different pattern and must be interpreted more carefully. Its best result is obtained by DenseNet121, which improves from 0.936 to 0.970 in weighted F1-score. This suggests that feature-space oversampling can be effective for compensating class imbalance when the classifier is trained on fixed backbone embeddings. However, this setting is not directly comparable to the end-to-end strategies. In the SMOTE pipeline, oversampling is applied only after feature extraction, and only the classifier head is trained. As a result, the learning problem is structurally simpler and computationally much cheaper than full end-to-end optimization. For that reason, `smote_features` should be interpreted as a feature-level baseline rather than as an equivalent alternative to the end-to-end balancing strategies.

This difference also explains why SMOTE is not evaluated on the proposed CBAM and fusion models. Feature-space SMOTE operates on flattened embeddings after global average pooling, whereas CBAM modifies intermediate feature maps in the (C, H, W) space before pooling. There is therefore no single post-attention embedding that can be used for SMOTE interpolation without changing the learning pipeline itself. Applying SMOTE after CBAM would require retraining the model end-to-end, which defeats the purpose of the two-stage feature-space approach. Applying it directly in feature-map space would violate the assumptions of the SMOTE algorithm. Accordingly, the `smote_features` strategy is structurally incompatible with the CBAM-based and fusion-based models. This limitation is methodological rather than incidental, and the comparison must be interpreted with that constraint in mind.

Taken together, Table 7 supports a broader conclusion: class balancing strategy is not a secondary implementation detail. It changes which model becomes strongest. Under `class_weight`, the most convincing result comes from a simple lightweight baseline. Under `augmentation_only`, the proposed fusion architecture achieves the best performance in the study. The main lesson is therefore not that one balancing strategy is universally superior, but that the effect of balancing depends on the architecture it supports and the kind of learning pipeline it enables.

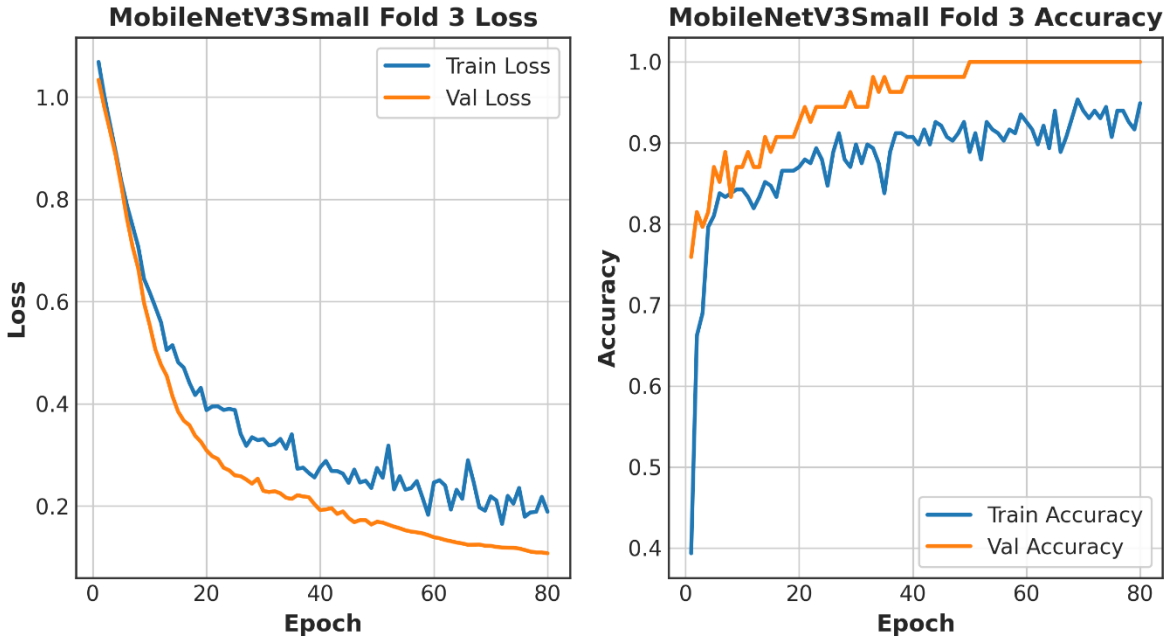
4.5. Learning Curves and Convergence

Figure 6 shows that the convergence behavior in this study is not random or unstable. It follows a clear pattern shaped by the two-phase training protocol. In *Phase 1*, where the backbone remains frozen and only the classifier head is trained at a learning rate of 1×10^{-4}, both models converge rapidly. Validation performance stabilizes within roughly 15 to 20 epochs, indicating that the pretrained ImageNet features already provide a strong starting representation for spinal X-ray classification. The early plateau, followed by one or two learning-rate reductions, suggests that optimization reaches a stable region without requiring prolonged training. This supports the design of the first phase: freezing the backbone allows efficient adaptation while limiting unnecessary parameter updates on a small dataset.

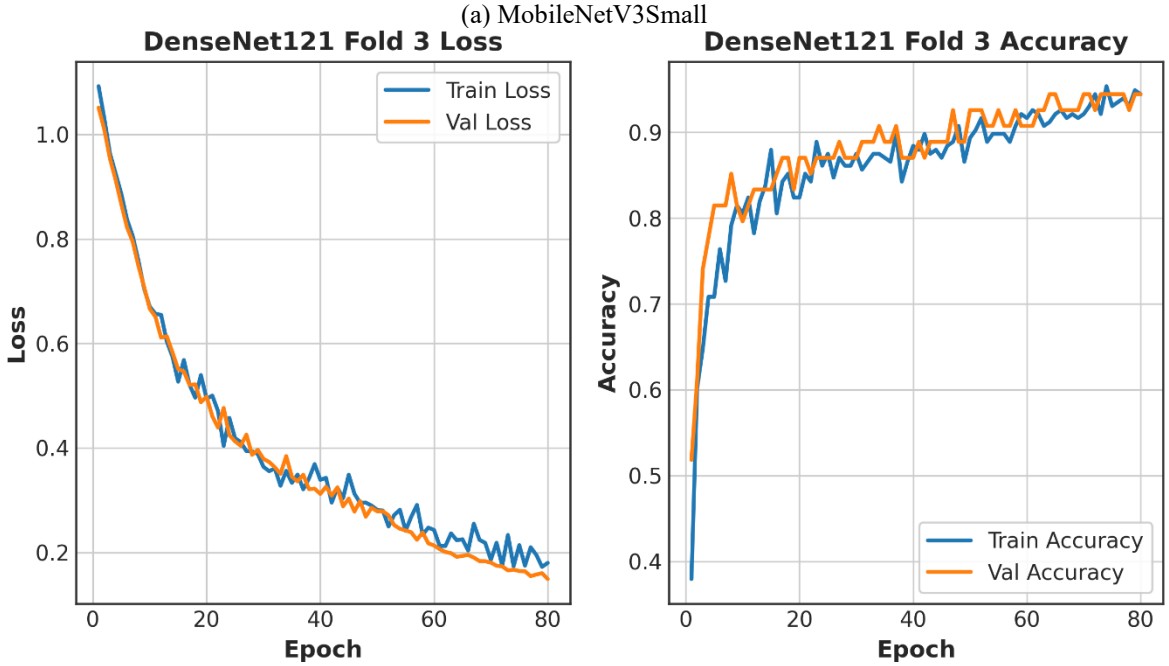
The transition to *Phase 2*, in which a subset of higher-level layers is unfrozen, and the learning rate is reduced to 1×10^{-5}, reveals a more informative contrast. For lightweight models such as MobileNetV3Small, the gain from fine-tuning is small, and the validation curve remains largely flat after unfreezing. This suggests that the pretrained features are already well aligned with the target task and that additional adaptation yields only limited benefit. In contrast, higher-capacity models, including EfficientNet-based and fusion models, show modest but more visible improvements after fine-tuning, typically on the order of 0.5 to 1 percentage point in weighted F1-score. The implication is that fine-tuning is not universally

https://doi.org/10.31436/iiumej.v27i2.4341

beneficial. Its value depends on how much unused representational capacity remains in the model and how much task-specific refinement the pretrained features still require.



(a) MobileNetV3Small



(b) DenseNet121

Figure 6. Learning curves under class_weight (fold 3). The dashed vertical line marks the Phase 1 to Phase 2 boundary. DenseNet121 exhibits clear train-validation divergence after epoch 30.

DenseNet121 provides the clearest counterexample to the assumption that more capacity is always useful. Its learning curve shows a noticeable divergence between training and validation losses beginning around epoch 30: training loss continues to decrease, while validation loss begins to rise. This is the characteristic signature of overfitting. Given the model's relatively large parameter count compared with the limited number of training samples per fold, this behavior is not surprising. More importantly, it provides direct empirical support for the broader conclusion of this paper: on small medical imaging datasets, larger architectures

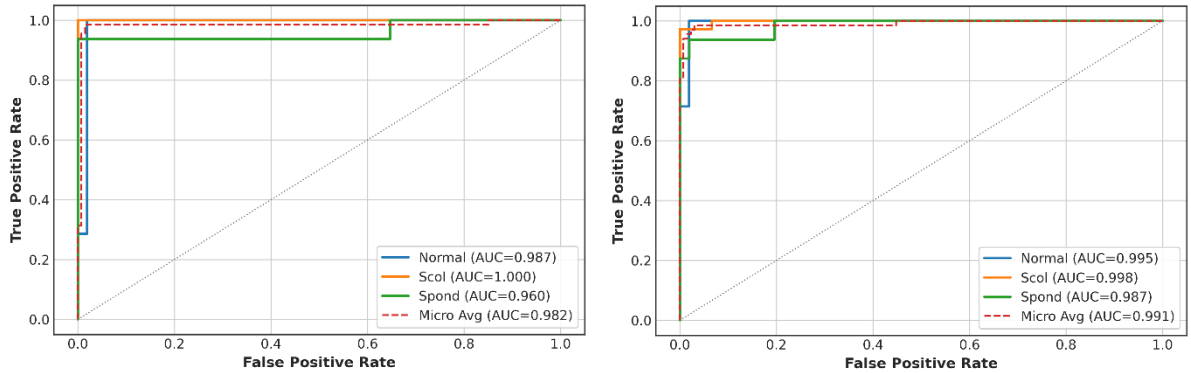
https://doi.org/10.31436/iiumej.v27i2.4341

can fit the training set more easily but generalize worse. In this setting, early stopping is not merely a routine safeguard. It functions as an active regularizer, preventing further training from degrading validation performance.

Equally important is what the learning curves do not show. None of the models exhibits unrealistically fast convergence to near-perfect training performance, which would raise concern about data leakage, trivial shortcuts, or overly aggressive preprocessing. Instead, the curves remain smooth and progressively saturating. This pattern supports the integrity of the experimental setup and suggests that the reported performance is driven by learned feature discrimination rather than accidental artifacts. Taken together, Figure 6 shows that the proposed training protocol achieves stable convergence, controlled generalization, and effective regularization, while also reinforcing a central result of the study: lightweight architectures offer the most favorable balance between learnability and generalization on this dataset.

4.6. ROC Analysis

Figure 7 provides a threshold-independent view of model behavior and confirms the main interpretation drawn from the accuracy and weighted F1 results. All high-performing models, particularly those with weighted F1-scores above 0.95, also achieve macro-average AUC values of approximately 0.98 or higher. This agreement is important because it shows that their performance is not tied to a single operating threshold. Instead, the models are learning class representations that remain well separated across a wide range of decision boundaries. In practical terms, the earlier observed performance gains reflect genuine improvements in discrimination rather than threshold-specific optimization.



(a) MobileNetV3Small (class_weight) (b) Fusion_MNv3_MNAS (augmentation_only)

Figure 7. Per-class one-vs-rest ROC curves. Spondylolisthesis achieves the highest per-class AUC across both models.

The ROC profiles of MobileNetV3Small under `class_weight` and Fusion_MNv3_MNAS under `augmentation_only` are especially informative. Their curves are both strong, but the fusion model shows a slightly more favorable separation, consistent with its higher weighted F1-score. This supports the interpretation that the gain from the fusion model is real but selective: it does not arise from arbitrary threshold tuning, but from improved class separability under a more suitable regularization regime.

At the class level, Spondylolisthesis remains the easiest category to distinguish, with per-class AUC values approaching 0.99 across the strongest models. This result is consistent with the confusion-matrix analysis and is clinically plausible. Spondylolisthesis involves a relatively

https://doi.org/10.31436/iiumej.v27i2.4341

pronounced geometric displacement of vertebral bodies, which produces a stronger global structural cue than the more subtle curvature changes seen in early or mild scoliosis. Such macroscopic patterns are more likely to survive resizing and to remain accessible to convolutional features transferred from large-scale pretraining.

The main residual difficulty remains at the Normal-Scoliosis boundary. The ROC curves for these two classes remain slightly weaker than those of Spondylolisthesis, indicating that the task is limited less by gross pathology detection than by discrimination of subtle borderline cases. This again aligns with the earlier confusion-matrix analysis. Mild scoliosis may preserve large regions of apparently normal anatomy, making the distinction dependent on finer geometric cues that are harder to preserve at standard input resolution. Thus, the ROC analysis does not introduce a new conclusion. It strengthens an existing one: the core challenge in this dataset is not recognizing obvious abnormality, but separating normal spines from mild curvature patterns with high reliability.

4.7. Summary of Key Findings

The results of this study converge on four main findings. *First*, among all evaluated backbones, MobileNetV3Small is the strongest efficiency-oriented baseline. It achieves top-tier predictive performance with only 1.0 million parameters and 61 million FLOPs, showing that, on this dataset, lightweight design is not a compromise but an advantage. The comparison with DenseNet121 makes this especially clear: substantially greater model capacity does not translate into better generalization when training samples are limited.

Second, the effectiveness of the proposed CBAM-enhanced and fusion-based models is conditional rather than universal. Under `class_weight`, they do not outperform their plain-backbone counterparts. Under `augmentation_only`, however, Fusion_MNv3_MNAS becomes the best-performing configuration in the study, achieving a weighted F1-score of 0.976. The implication is that attention and fusion are not inherently beneficial in isolation. Their value emerges when the training regime provides enough data diversity to regularize the added representational capacity.

Third, `augmentation_only` is the most effective end-to-end balancing strategy in this study. Unlike `smote_features`, it operates within the same training pipeline as the proposed architectures and does not rely on a separate feature-level oversampling stage. More importantly, it consistently improves models with the greatest representational flexibility, indicating that data diversity is a more effective enabler of attention and fusion than loss reweighting alone. In this setting, augmentation functions not only as a balancing tool, but also as the key regularizer that allows more expressive architectures to generalize.

Finally, all analyses converge on the same residual source of error: the Normal–Scoliosis boundary. Spondylolisthesis is consistently well separated, while mild scoliosis remains difficult to distinguish from normal anatomy. The remaining challenge in spinal X-ray classification is therefore not the detection of obvious pathology but the reliable recognition of subtle curvature-related variation.

These findings are based on a single public dataset of 338 subjects. Although stratified 5-fold cross-validation strengthens the estimate relative to a fixed split, generalization to broader clinical populations, different acquisition settings, or more diverse radiographic characteristics remains to be established. Accordingly, the contribution is best understood as a robust, efficiency-aware methodological benchmark rather than a final deployment claim.

https://doi.org/10.31436/iiumej.v27i2.4341

5. ABLATION STUDIES

The main results presented in Section 4 identify which models perform best. The ablation studies explain why. This distinction is important because the proposed framework combines several interacting design choices, including attention, fine-tuning, fusion, and class balancing. Without controlled ablation, an observed gain could be incorrectly attributed to the wrong component. For that reason, four targeted ablation groups are conducted, each varying one factor while holding the others fixed. The purpose is not to generate more tables, but to isolate which design choices genuinely contribute to performance, under which training conditions, and with what practical implications. In this study, the ablations are especially important because the main results already suggest that performance is governed by model–training interaction rather than architecture alone.

5.1. CBAM Sub-Module Analysis

The first ablation examines a simple but necessary question: *if attention helps, which part of attention is most important*? To answer this, the contribution of the individual CBAM components is isolated on the MobileNetV3Small backbone under the `class_weight` strategy. Four controlled conditions are evaluated: the plain backbone without attention, channel attention only, spatial attention only, and full CBAM with sequential channel-spatial attention. This design allows the effect of each attention component to be interpreted directly rather than inferred indirectly from the full model.

Table 8. CBAM sub-module ablation on MobileNetV3Small under the class_weight strategy (5-fold stratified cross-validation, mean ± standard deviation).

Condition	Accuracy	F1 (weighted)	Params
No attention (baseline)	0.965 ± 0.022	0.965 ± 0.022	1,001,251
Channel attention only	0.953 ± 0.019	0.954 ± 0.019	1,042,723
Spatial attention only	0.956 ± 0.027	0.957 ± 0.027	1,001,349
Full CBAM	0.953 ± 0.019	0.954 ± 0.019	1,042,821

The result shown in Table 8 is unambiguous. All three attention variants underperform the plain baseline. The baseline achieves the highest weighted F1-score at 0.965±0.022, followed by spatial attention only at 0.957±0.027, while both channel attention only and full CBAM reach 0.954±0.019. The full CBAM configuration, therefore, reduces weighted F1 by about 1.1 percentage points relative to the baseline while also adding approximately 41,000 parameters. This is a small architectural increase, but it produces no compensating gain. More importantly, the pattern is systematic rather than random: the ranking is consistent enough to suggest that, under `class_weight` training on this dataset, the added attention capacity is not translated into better generalization.

This is an important negative result. It challenges the common expectation that attention modules are inherently beneficial once inserted into a CNN backbone. In the present setting, that assumption does not hold. The issue is not that CBAM is poorly designed. The issue is that the training regime does not provide enough data diversity to support the added representational flexibility. Under these conditions, the model gains more freedom than the supervision signal can reliably guide.

Among the three attention variants, spatial attention performs better than channel attention, although the margin is modest. This difference is meaningful. Spinal disorder classification depends primarily on geometric structure, especially curvature in scoliosis and vertebral

https://doi.org/10.31436/iiumej.v27i2.4341

displacement in spondylolisthesis. Spatial attention is more naturally aligned with this kind of signal because it emphasizes where discriminative evidence occurs in the image. Channel attention, by contrast, is more indirect in this context because it reweights feature channels without explicitly sharpening spatial localization. The result, therefore, suggests that, when attention is useful for this task, spatial selectivity is more relevant than channel recalibration.

At the same time, spatial attention alone still does not surpass the plain baseline. This shows that partial alignment with the task is not sufficient on its own. The broader conclusion is that, under the `class_weight` regime, the full CBAM module is not justified for MobileNetV3Small on this dataset. Channel and spatial attention do not combine to produce an additive benefit, and the plain backbone remains the strongest option.

This should not be interpreted as a rejection of attention mechanisms in general. Rather, it provides direct evidence that the value of attention is conditional. As shown later in Section 5.4, the same attention-based models behave differently under `augmentation_only`, where richer data diversity allows the added representational capacity to be regularized more effectively. The practical lesson is therefore precise: in spinal X-ray classification, attention should not be evaluated as a universal architectural upgrade, but as a component whose usefulness depends on the training regime that supports it.

5.2. Fine-Tuning Depth

This ablation examines a practical question in transfer learning: *how much of the pretrained backbone should be unfrozen before additional adaptation stops helping*? The question matters because too little fine-tuning leaves the model dependent on generic pretrained features, whereas too much fine-tuning risks unnecessary complexity and instability on a small medical dataset. To study this trade-off directly, five fine-tuning fractions were evaluated on CBAM_MobileNetV3Small under the `class_weight` strategy, ranging from a fully frozen backbone to unfreezing half of the network. This provides a controlled view of how increasing trainable capacity translates into measurable gains.

Table 9. Effect of backbone fine-tuning depth on CBAM_MobileNetV3Small under the class_weight strategy (5-fold stratified cross-validation, mean ± standard deviation).

Fraction Unfrozen	Trainable Params	Accuracy	F1 (weighted)	Train Time (s)
0% (frozen)	115,813	0.941 ± 0.018	0.942 ± 0.017	71.2
10%	409,909	0.950 ± 0.025	0.951 ± 0.024	107.3
20% (default)	704,581	0.956 ± 0.021	0.956 ± 0.020	111.1
30%	838,477	0.962 ± 0.022	0.962 ± 0.022	113.2
50%	904,109	0.962 ± 0.022	0.962 ± 0.022	116.0

The pattern is clear, as shown in Table 9. Performance improves steadily as the fine-tuning fraction increases from 0% to 30%, then levels off. Both the 30% and 50% settings produce identical mean accuracy, identical weighted F1-score, and identical variance across folds. This indicates that the useful domain-specific adaptation is already captured by unfreezing about 30% of the backbone. Beyond that point, additional trainable parameters do not yield further gain. They slightly increase optimization cost, but not learning quality.

Figure 8 makes this pattern easy to interpret. The fully frozen setting performs worst, with a weighted F1-score of 0.942±0.017, which is about 2 percentage points below the 30% plateau. This gap is large enough to show that feature extraction alone is insufficient for this task. Even with strong ImageNet pretraining, some degree of domain-specific adaptation is necessary to

https://doi.org/10.31436/iiumej.v27i2.4341

capture the geometric characteristics of spinal disorders. In this dataset, fine-tuning is therefore not optional. It is part of what makes the model competitive.

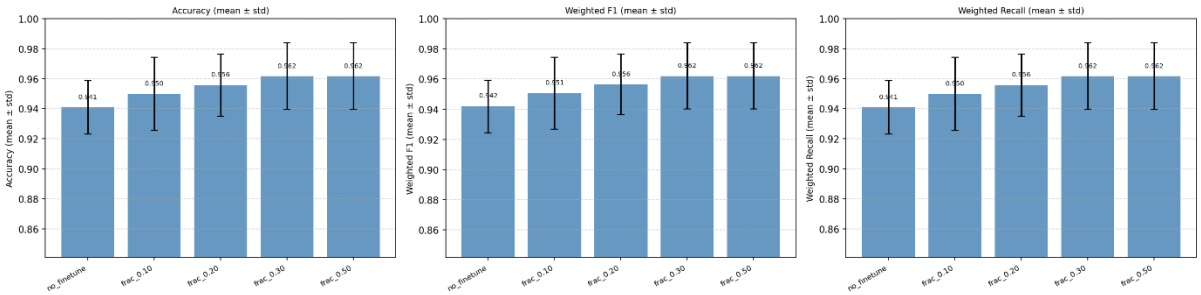


Figure 8. Weighted F1-score as a function of the backbone fine-tuning fraction. Performance plateaus at 30%, with the 30% and 50% conditions producing identical per-fold metrics.

At the same time, the results do not support the opposite extreme either. Unfreezing more than 30% of the backbone provides no measurable advantage. The 50% setting increases training time relative to 30%, but produces the same performance. The implication is practical: partial fine-tuning should continue only until performance saturates. Beyond that point, additional flexibility is unnecessary.

This result also clarifies the reviewer's concern regarding the difference between the benchmark protocol and the ablation optimum. In the main comparative experiments, 20% unfreezing was used as the *default benchmark setting* to maintain a uniform training protocol across all evaluated models. The purpose of that setting was fairness of comparison, not model-specific optimization. The ablation study shows that, for CBAM_MobileNetV3Small, a 30% fraction is the better task-specific choice, improving weighted F1 from 0.956 to 0.962 with only a negligible increase in training time. The main results should therefore be interpreted as results under a controlled common protocol, whereas this ablation identifies the more refined optimum for this specific model.

An equally important observation is what does not happen. Contrary to the common concern that unfreezing too many layers on a small dataset will inevitably lead to overfitting, no performance collapse is observed even at 50% unfreezing. This suggests that the training protocol, especially early stopping and adaptive learning rate scheduling, is effective in constraining the optimization process. In this setting, generalization is shaped not only by the parameter count but also by how the fine-tuning process is controlled.

Taken together, the fine-tuning ablation yields a simple design rule: a fully frozen backbone is too rigid, while deeper fine-tuning beyond 30% is unnecessary. For CBAM_MobileNetV3Small on this dataset, the useful adaptation window lies between. That is the point at which pretrained knowledge is preserved, but the model is given enough freedom to learn the subtle geometric signatures of spinal pathology.

5.3. Fusion versus Individual Branches

This ablation tests a central design question of the proposed framework: *does logit-level fusion between two CBAM-enhanced backbones yield a model stronger than either branch alone*? The motivation is straightforward. MobileNetV3Small and MNASNet differ in width, embedding dimension, and design objective, so they may learn complementary representations of spinal anatomy. If that complementarity is real, combining their predictions should improve robustness and, potentially, average performance. To isolate the effect of fusion, the two

individual branches are compared directly against the fused model using the `class_weight` strategy.

Table 10. Fusion versus individual CBAM branches under the class_weight strategy (5-fold stratified cross-validation, mean ± standard deviation).

Model	Accuracy	F1 (weighted)	Params (M)	Size (MB)
CBAM_MobileNetV3Small	0.944 ± 0.019	0.945 ± 0.019	1.040	3.980
CBAM_MNASNet	0.920 ± 0.042	0.922 ± 0.040	3.470	13.240
Fusion_MNv3_MNAS	0.941 ± 0.029	0.943 ± 0.028	4.510	17.220

The result is instructive, as shown in Table 10. Under `class_weight`, the fusion model does not outperform its strongest branch. Fusion_MNv3_MNAS achieves a weighted F1-score of 0.943±0.028, slightly below CBAM_MobileNetV3Small at 0.945±0.019. The gap is small and remains within one standard deviation, but the direction is clear. Fusion does not create a stronger classifier when one branch is consistently weaker than the other. Instead, uniform logit averaging pulls the final prediction toward the mean behavior of the two branches. In this case, the weaker CBAM_MNASNet branch reduces the advantage of the stronger CBAM_MobileNetV3Small branch rather than contributing enough complementary information to offset its own weakness.

At the same time, fusion provides something different from a higher mean score: it improves stability. CBAM_MNASNet shows substantial fold-to-fold variability, with a weighted F1 standard deviation of about 4.0% and worst-fold performance at 0.858. The fusion model compresses that variability and raises the worst-fold outcome to 0.901. This means that, even without improving the average score, fusion reduces the risk of particularly poor runs. That property is practically relevant. In clinical decision-support settings, a model that is slightly less optimal on average but more resistant to large performance drops may still be preferable, especially when reliability matters more than chasing a marginal gain in peak accuracy.

This result clarifies the role of fusion in the present framework. Under limited regularization, fusion primarily serves as a stabilizer rather than a performance amplifier. It can soften the effect of a weaker branch, but it cannot recover information that the branches themselves have not learned. For fusion to produce a genuine accuracy gain, both branches must contribute useful and sufficiently distinct representations.

That is exactly what appears under `augmentation_only`. In that regime, Fusion_MNv3_MNAS reaches a weighted F1-score of 0.976, clearly surpassing both individual branches. The contrast is important. It shows that fusion benefits from branch diversity only when both branches are trained under conditions that allow their representations to generalize well. Without sufficient regularization, the weaker branch contributes more noise than complementarity. With stronger augmentation, that same branch becomes more useful, and fusion begins to realize its intended benefit.

From a deployment perspective, the fusion model remains within practical limits despite its added complexity. With 4.51 million parameters and a model size of 17.22 MB, it is still substantially lighter than DenseNet121. Thus, the cost of fusion is real but controlled. The main conclusion of this ablation is therefore precise: *fusion is not guaranteed to improve mean performance under all training conditions, but it does improve stability and can become the best-performing strategy when the branches are sufficiently regularized*. In this study, the value of fusion is conditional rather than universal, consistent with the broader pattern observed in attention and class balancing.

https://doi.org/10.31436/iiumej.v27i2.4341

5.4. Balancing Strategy Interaction with Proposed Models

This ablation tests a question that is easy to overlook but central to the interpretation of the entire study: *Does the best model remain the best when the balancing strategy changes*? If the ranking changes, then architecture alone is not the real explanation of performance. To examine this directly, four proposed models were evaluated under both `class_weight` and `augmentation_only`. This allows the effect of the training regime to be separated from the effect of the architecture.

As shown in Table 11, the ranking is clearly not stable across strategies. Under class_weight, CBAM_EfficientNetB0 achieves the highest weighted F1-score at 0.958±0.019, while Fusion_MNv3_MNAS reaches only 0.948±0.022. Under augmentation_only, the order changes completely: Fusion_MNv3_MNAS becomes the strongest model at 0.974±0.026, whereas CBAM_EfficientNetB0 drops to 0.946±0.038. This inversion is the most important result of the ablation. It shows that conclusions about model superiority are conditional on the training regime. A model that appears suboptimal under one balancing strategy may become the best model under another.

Table 11. Performance of proposed models under class_weight and augmentation_only strategies (5-fold stratified cross-validation, mean ± standard deviation).

Model	Strategy	Accuracy	F1 (weighted)
CBAM_EfficientNetB0	class_weight	0.959 ± 0.019	0.958 ± 0.019
CBAM_EfficientNetB0	augmentation_only	0.947 ± 0.037	0.946 ± 0.038
CBAM_MobileNetV3Small	class_weight	0.956 ± 0.023	0.956 ± 0.023
CBAM_MobileNetV3Small	augmentation_only	0.947 ± 0.023	0.946 ± 0.023
CBAM_MNASNet	class_weight	0.902 ± 0.038	0.906 ± 0.036
CBAM_MNASNet	augmentation_only	0.926 ± 0.043	0.928 ± 0.041
Fusion_MNv3_MNAS	class_weight	0.947 ± 0.022	0.948 ± 0.022
Fusion_MNv3_MNAS	augmentation_only	0.973 ± 0.026	0.974 ± 0.026

The models separate into two distinct response patterns. CBAM_MNASNet and Fusion_MNv3_MNAS improve substantially under `augmentation_only`, gaining about 2.2 to 2.6 percentage points in weighted F1-score. By contrast, CBAM_EfficientNetB0 and CBAM_MobileNetV3Small decline slightly. This divergence is not random. The two stronger models under `class_weight` are already operating near the performance ceiling of this dataset, leaving limited room for further gains from augmentation. Additional input variability can then act more as a perturbation than as a benefit. In contrast, the weaker and higher-capacity models have more headroom and benefit more from the extra diversity introduced by augmentation.

This explains why `augmentation_only` works best for the proposed fusion architecture. Fusion_MNv3_MNAS has the largest representational capacity among the proposed models and therefore also the greatest need for regularization. Under `class_weight`, that capacity is not fully used. Under `augmentation_only`, the model is exposed to a broader distribution of anatomically plausible inputs, thereby reducing variance and enabling the two branches to learn more useful, complementary representations. The gain is therefore not caused by augmentation in isolation, but by the match between augmentation and model capacity. In this study, augmentation is the condition that allows fusion to become effective.

The same interaction also clarifies why `smote_features` is excluded from this comparison. SMOTE operates on flattened feature embeddings after global average pooling, whereas CBAM modifies intermediate feature maps in the (C, H, W) space before pooling. As

https://doi.org/10.31436/iiumej.v27i2.4341

a result, there is no single feature representation that both preserves the effect of CBAM and remains suitable for SMOTE interpolation. Applying SMOTE after CBAM would require retraining the full model end-to-end, defeating the purpose of the two-stage feature-space pipeline. Applying it directly to feature maps is not supported by the SMOTE formulation. The exclusion of SMOTE for CBAM and fusion models is therefore a structural constraint of the method, not an omitted experiment.

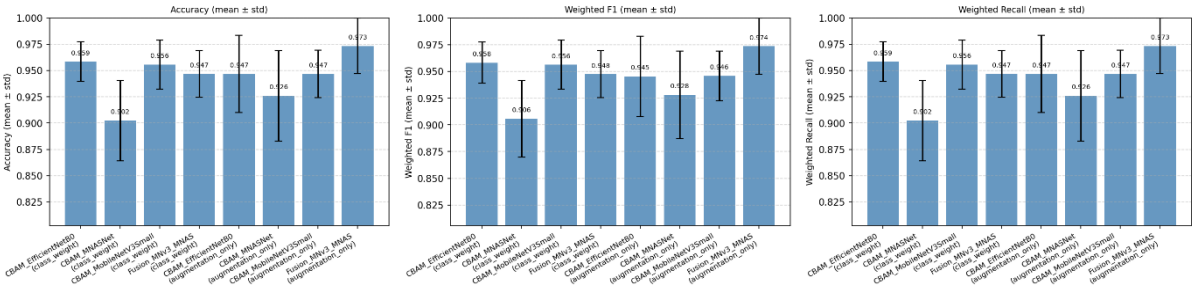


Figure 9. Weighted F1-score for each proposed model under class_weight and augmentation_only strategies. Fusion_MNv3_MNAS exhibits the largest strategy-dependent gain (+2.6 pp F1)

Figure 9 makes the strategy dependence visually explicit. The largest gain belongs to Fusion_MNv3_MNAS, while the two strongest single-branch models under `class_weight` show slight decreases under `augmentation_only`. This confirms that the contribution of CBAM and fusion is real but conditional. Evaluating these models only under `class_weight` would lead to the misleading conclusion that attention and fusion are ineffective. The more accurate conclusion is narrower and more useful: *on this dataset, CBAM and fusion become advantageous only when paired with sufficient data-driven regularization*. For the proposed models, `augmentation_only` is therefore the most appropriate training regime.

5.5. Summary of Ablation Findings

The ablation studies lead to two clear conclusions. *First*, attention is not universally beneficial in small-data settings. Under the class_weight regime, all CBAM variants underperform the plain backbone, with spatial attention proving the strongest variant. This is consistent with the geometric nature of the task: curvature and vertebral displacement favor spatial localization over channel reweighting. The fine-tuning ablation adds a complementary constraint: a fully frozen backbone is too rigid, while unfreezing beyond 30% yields no further gain. Useful domain adaptation is necessary, but bounded.

Second, and more importantly, model performance is governed by the interaction between architecture and training strategy rather than by architecture alone. Under `class_weight`, the proposed fusion model does not outperform its strongest branch. Under `augmentation_only`, the same model becomes the best-performing configuration in the study, with a weighted F1-score gain of about 2.6 percentage points. This directly explains why `augmentation_only` performs best: it provides the data diversity required to regularize the added capacity introduced by attention and fusion. The methodological implication is that, on small medical imaging datasets, architectural design and training strategy must be optimized jointly. Evaluating one without the other can mislead conclusions about which model is actually strongest.

6. CONCLUSIONS

This study presented an efficiency-aware deep learning framework for three-class classification of spinal disorders from X-ray images. The central message is that, on a dataset of this scale, the strongest model is not the largest, but the one whose architecture, regularization, and training strategy are matched to the task. MobileNetV3Small reached a weighted F1-score of 0.962 with only 1.0 million parameters and 61 million FLOPs, surpassing far heavier baselines such as DenseNet121 while remaining substantially smaller and faster.

CBAM and dual-branch fusion were not universally beneficial. Under `class_weight`, they did not surpass plain backbones, but under `augmentation_only`, `Fusion_MNv3_MNAS` achieved the best overall weighted F1-score of 0.976. Ablation analyses clarified the underlying mechanism: attention requires sufficient regularization to be useful, fusion improves stability before mean accuracy, and partial fine-tuning is necessary but saturates near 30% of the backbone. Throughout ROC, confusion-matrix, and efficiency analyses, Spondylolisthesis remained reliably separable, while the dominant residual difficulty lay at the Normal–Scoliosis boundary.

The principal contribution is therefore methodological rather than a single headline number: in spinal X-ray classification, lightweight models can be remarkably strong, architectural innovations are conditional rather than absolute, and reliable progress comes from co-optimizing model design with the training regime that allows it to generalize. Future work will extend this benchmark to larger, multi-center datasets and to additional spinal pathologies, to test how far the present conclusions generalize beyond the 338-subject setting examined here.

ACKNOWLEDGMENT

This research was supported by the SASMEC @ IIUM Research Grant (Project ID: SRG25-098-0098). The second author gratefully acknowledges the Kulliyyah of Engineering IIUM Engineering Merit Scholarship 2025 (KOEIEMS25). Ethical approval for this study, including the collection of additional scoliosis X-ray and ultrasound images, was granted by the IIUM Research Ethics Committee (IREC) under application number IREC-2025-3. The authors also wish to note that Figure 2 has been submitted for copyright registration to the Malaysian Intellectual Property Office through the IIUM Library. The authors used generative AI tools (ChatGPT and Claude) solely for language refinement and structural editing during manuscript preparation; all AI-assisted text was reviewed and verified by the authors, and no AI was used to generate research data, results, or substantive technical content.

REFERENCES

[1] M. Fraiwan, Z. Audat, L. Fraiwan, and T. Manasreh, "Using deep transfer learning to detect scoliosis and spondylolisthesis from x-ray images," *Plos One,* vol. 17, no. 5, p. e0267851, 2022.

[2] R. R. Maaliw III, "SCOLIONET: an automated scoliosis Cobb angle quantification using enhanced X-ray images and deep learning models," *Journal of Imaging,* vol. 9, no. 12, p. 265, 2023.

[3] M. I. Jamaludin, T. S. Gunawan, R. K. Karupiah, S. A. Zabidi, M. Kartiwi, and Z. Zakaria, "Optimizing U-Net Architecture with Feed-Forward Neural Networks for Precise Cobb Angle Prediction in Scoliosis Diagnosis," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI),* vol. 11, no. 3, pp. 883-895, 2023.

[4] N. A. Makhdoomi *et al.*, "Development of scoliotic spine severity detection using deep learning algorithms," in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, 2022: IEEE, pp. 0574-0579.

[5] H. Güneş, C. Hark, and A. E. Akkaya, "Comparison of deep learning models and optimization algorithms in the detection of scoliosis and spondylolisthesis from X-ray images," *Sakarya University Journal of Science,* vol. 28, no. 2, pp. 438-451, 2024.

[6] G. M. Trinh *et al.*, "Detection of lumbar spondylolisthesis from X-ray images using deep learning network," *Journal of Clinical Medicine,* vol. 11, no. 18, p. 5450, 2022.

[7] A. Vephasayanant, A. Jitpattanakul, P. Muneesawang, K. Wongpatikaseree, and N. Hnoohom, "YOLO-based image segmentation for the diagnostic of spondylolisthesis from lumbar spine X-ray images," *IEEE Access,* vol. 12, pp. 182242-182258, 2024.

[8] C. Xu *et al.*, "Deep Learning-Based Diagnosis of Lumbar Spondylolisthesis Using X-Ray Imaging," *Diagnostics,* vol. 15, no. 16, p. 2015, 2025.

[9] F. N. M. Zamri, T. S. Gunawan, S. H. Yusoff, A. A. Alzahrani, A. Bramantoro, and M. Kartiwi, "Enhanced small drone detection using optimized YOLOv8 with attention mechanisms," *IEEE Access,* vol. 12, pp. 90629-90643, 2024.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132-7141.

[11] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3-19.

[12] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," *arXiv preprint arXiv:1801.09927,* 2018.

[13] O. Oktay *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999,* 2018.

[14] G. Papanastasiou, N. Dikaios, J. Huang, C. Wang, and G. Yang, "Is attention all you need in medical image analysis? A review," *IEEE Journal of Biomedical and Health Informatics,* vol. 28, no. 3, pp. 1398-1411, 2023.

[15] P. Chen, Z. Zhou, H. Yu, K. Chen, and Y. Yang, "Computerized-Assisted Scoliosis Diagnosis Based on Faster R-CNN and ResNet for the Classification of Spine X-Ray Images," *Computational and Mathematical Methods in Medicine,* vol. 2022, no. 1, p. 3796202, 2022.

[16] Z. Al-Milaji and H. Yousif, "Lightweight deep learning model optimization for medical image analysis," *International Journal of Imaging Systems and Technology,* vol. 34, no. 5, p. e23173, 2024.

[17] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, 2019: PMLR, pp. 6105-6114.

[18] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510-4520.

[19] A. Howard et al., "Searching for mobilenetv3," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1314-1324.

[20] M. Tan *et al.*, "Mnasnet: Platform-aware neural architecture search for mobile," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2820-2828.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.

[22] J. Irvin *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, no. 01, pp. 590-597.

[23] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *Advances in neural information processing systems,* vol. 32, 2019.

[24] X. Liu *et al.*, "Advances in deep learning-based medical image analysis," *Health Data Science,* vol. 2021, p. 8786793, 2021.

[25] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine,* vol. 17, no. 1, p. 195, 2019.

[26] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine,* vol. 25, no. 1, pp. 44-56, 2019.

[27] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention," *arXiv preprint arXiv:1804.02391,* 2018.

[28] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Transactions on Medical Imaging,* vol. 37, no. 7, pp. 1597-1605, 2018.

[29] V. V. Valindria *et al.*, "Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI," in *2018 IEEE winter conference on applications of computer vision (WACV)*, 2018: IEEE, pp. 547-556.

[30] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested U-Net architecture for medical image segmentation," in *International workshop on deep learning in medical image analysis*, 2018: Springer, pp. 3-11.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[32] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data,* vol. 6, no. 1, pp. 1-48, 2019.

[33] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering,* vol. 21, no. 9, pp. 1263-1284, 2009.

[34] G. Douzas and F. Bacao, "Effective data generation for imbalanced learning using conditional generative adversarial networks," *Expert Systems with Applications,* vol. 91, pp. 464-471, 2018.