2026-08

Towards second opinion: anatomy-aware reasoning from explicit priors to adaptive gating for efficient medical image analysis

Bhuiyan, Siam Tahsin

Independent University, Bangladesh (IUB)

Towards SecondOpinion: Anatomy-Aware Reasoning from Explicit Priors to Adaptive Gating for Efficient Medical Image Analysis

August 2026

Prepared by:

Siam Tahsin Bhuiyan

ID: 2010764

Department of Computer Science and Engineering

Independent University, Bangladesh

Supervised by:

Md Rashedur Rahman, D. Eng.

Assistant Professor

Department of Computer Science and Engineering

Independent University, Bangladesh

A Final Year Design Project (FYDP) Report submitted for the Degree of Bachelor's of Computer Science & Engineering

Attestation

I am aware of the fact that plagiarism is strictly prohibited and it conflicts with my university's rules and regulations. I guarantee the authenticity of my work. I have used some copyrighted material and models of others in my project which has been properly cited following the international standards proper guidelines.

Author Name:

Siam Tahsin Bhuiyan

Signature:

. . . . . .

Evaluation Committee

Supervisor

Name: _____ Signature: _____

Internal Examiner 1

Name: _____ Signature: _____

Internal Examiner 2

Name: _____ Signature: _____

External Examiner

Name: _____ Signature: _____

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Md Rashedur Rahman (D. Eng.), Assistant Professor, Department of Computer Science and Engineering, Independent University, Bangladesh (IUB), for his continuous guidance, patience, and support throughout this research. His feedback shaped this report and the broader line of work it builds on.

I am also grateful to Dr. Saadia Binte Alam for her guidance and support throughout this research, and for her valuable input at every stage of this work.

I would also like to thank Dr. Syoji Kobashi for his mentorship during my time at the University of Hyogo, Japan under the one-year HUMAP research scholarship. That period was central to the anatomy-aware direction this report takes, and his guidance throughout has been invaluable.

I also thank my collaborators and co-authors at the Medical Imaging Research & Analysis Wing (MIRAW), Center for Computational & Data Sciences (CCDS), IUB, and the Advanced Medical Engineering Research Institute (AMERI), University of Hyogo, whose work on the datasets, prior architectures, and evaluation protocols this report builds on made this research possible.

I acknowledge the use of Claude (Anthropic) solely for editorial and language refinement in the preparation of this report. It was not used for developing the research methodology, hypotheses, data analysis, experimental results, or their interpretation.

Finally, I would like to thank my family and friends for their support throughout this work, and the Department of Computer Science and Engineering and CCDS at Independent University, Bangladesh, for the resources to carry it out.

List of Publications from this Research

1. **S. T. Bhuiyan**, R. Rahman, S. Wasi, R. Islam, S. Kobashi, A. Islam, and S. B. Alam, "SecondOpinion: Anatomy-aware gated reasoning for efficient medical image analysis," in Proc. 2nd MICCAI Workshop on Efficient Medical AI (EMA4MICCAI), MICCAI 2026, Strasburg, France, in press.

2. **S. T. Bhuiyan**, R. Rahman, S. Wasi, N. Yagi, S. Kobashi, A. Islam, and S. B. Alam, "Invisible yet detected: PelFANet with attention-guided anatomical fusion for pelvic fracture diagnosis," in Empowering Medical Image Computing and Research Through Early-Career Expertise (EMERGE 2025), MICCAI 2025, Daejeon, South Korea, Lecture Notes in Computer Science, vol. 16534, Springer, Cham, 2026, doi: 10.1007/978-3-032-24182-5_10.

3. **S. T. Bhuiyan**, R. Rahman, S. Wasi, H. Khatun, M. S. Chowdhury, R. Islam, S. K. Mazumder, N. Yagi, S. Kobashi, and S. B. Alam, "Beyond the pelvic ring: Benchmarking multi-bone segmentation for ROI-guided fracture detection," Scientific Reports, vol. 16, art. no. 24508, 2026, doi: 10.1038/s41598-026-47663-8.

4. **S. T. Bhuiyan**, R. Rahman, S. Wasi, H. Khatun, A. Islam, A. M. Rahman, S. B. Alam, and M. A. Amin, "PelviNeXt: A modality-agnostic hybrid network for pelvic imaging in women's health," in Proc. Joint Computer-Assisted Pelvic Imaging (CAPI) and Women's Health (WOMEN) Workshops, MICCAI 2026, Strasburg, France, in press.

5. **S. T. Bhuiyan**, R. Rahman, R. Islam, M. Haque, M. Islam, S. Kobashi, and S. B. Alam, "Lung disease diagnosis from CXR using convolutional neural network (CNN): A comparative study," Soft Computing, vol. 29, pp. 6333–6345, 2025, doi: 10.1007/s00500-025-10912-5.

6. **S. T. Bhuiyan**, Z. Zaman, R. Rahman, M. S. Choudhury, S. Wasi, A. Islam, S. B. Alam, and S. Kobashi, "CheXNet-CBAM: Enhancing large-scale medical pre-training with attention for lung disease diagnosis," in Proc. 2026 IEEE 56th Int. Symp. Multiple-Valued Logic (ISMVL), 2026, Miyagi, Japan, pp. 105–110, doi: 10.1109/ISMVL68998.2026.00027.

7. **S. T. Bhuiyan**, R. Rahman, S. B. Alam, H. Khatun, R. Islam, S. Wasi, M. A. R. Iftee, and A. Islam, "Patch-based deep ensemble learning for enhancing pelvic fracture detection," in Proc. 2025 28th Int. Conf. Computer and Information Technology (ICCIT), 2025, pp. 3022–3027, doi: 10.1109/ICCIT68739.2025.11491331.

8. **S. T. Bhuiyan**, R. Khatun, S. K. Mazumder, F. Israq, S. Wasi, R. Rahman, A. Islam, and S. B. Alam, "Preprocessing matters: Benchmarking image enhancement techniques for pelvic fracture detection," in Proc. 2025 IEEE Int. Women in Engineering (WIE) Conf. Electrical and Computer Engineering (WIECON-ECE), 2025, pp. 379–384, doi: 10.1109/WIECON-ECE69386.2025.11525920.

Abstract

In clinical practice, a radiologist does not spend the same amount of effort on every case. A straightforward presentation is read quickly, while an ambiguous one prompts more scrutiny: a second look, additional imaging, or a second opinion from a colleague. Most deep learning models built for medical image analysis do not work this way. Regardless of how easy or difficult a given case actually is, a model applies the same fixed amount of computation to it, whether that means a single forward pass through a classifier or, in anatomy-guided designs, a full second stream of anatomical reasoning run on every input without exception. This mismatch between how much a model works and how much a given case actually demands is the gap this research addresses. This report presents SecondOpinion, a dual-stream framework built around GateKeeper, a lightweight gating mechanism trained explicitly as a binary correctness classifier rather than relying on unsupervised confidence, which decides on a per-case basis whether a fast primary stream's prediction can be trusted or whether a second, anatomy-guided stream and a cross-attention fusion module should be invoked, much like a clinician calling in a second opinion only when a case warrants it. Training proceeds in two phases, first with GateKeeper frozen and then with it unfrozen under soft gating, with hard gating applied at inference.

SecondOpinion is tested on three tasks chosen to span a range of diagnostic difficulty: chest disease classification on a five-class CXR dataset of 21,865 images, and pelvic fracture detection on the private AMERI dataset, split into visible and invisible fracture cases. On CXR, it reaches 98.41% accuracy and 0.990 AUROC, within 1.2 points of an always-on version of the same model, while activating its second stream on only 9.23% of cases and using about half the FLOPs. On the pelvic tasks, it trails the always-on ceiling by 2.2 points of F1 on visible fractures and 1.2 points on invisible fractures, using 60% and 71% of always-on FLOPs. GateKeeper's activation rate rises from 9.23% on CXR, to 24.12% on visible fractures, to 45.71% on invisible fractures. This same order matches an independently measured AUROC gap between a single-stream baseline and the always-on model across the three tasks. The pattern holds both when comparing across domains, CXR versus pelvic fracture detection, and within a domain, visible versus invisible fractures.

These results support three conclusions. Conditional gating can match always-on performance at a fraction of the cost. A correctness-supervised gate's activation rate tracks real diagnostic difficulty rather than firing at some fixed or arbitrary rate. And this relationship holds across genuinely different diagnostic domains, not just one dataset. SecondOpinion is meant as a decision-support tool: it calls in extra anatomical reasoning only when a case seems to need it, the way a clinician spends more effort on a harder case.

Contents

List of Figures

List of Tables

Chapter-1: Introduction

Medical image analysis has become one of the most active application areas for deep learning, with automated systems now proposed, and in some cases deployed, across nearly every diagnostic imaging task. Much of this progress has come from architectures that incorporate anatomical knowledge directly into their design, rather than expecting a network to discover relevant anatomical structure entirely on its own. Yet as these architectures have grown more capable, they have also tended to grow less discriminating about when that added capability is actually needed. This research is concerned with that gap, and specifically with whether a diagnostic model can be taught to reason about when additional anatomical processing is worth its cost, rather than applying it indiscriminately to every case it encounters.

1.1 Background

1.1.1 Fixed Computation as a General Limitation in Medical AI

Deep learning has become a dominant approach across nearly every branch of medical image analysis, achieving strong and, in several reported settings, expert-comparable diagnostic performance [1]. This success spans a wide range of modalities, anatomical regions, and pathologies. Alongside this general success, a substantial body of work has shown that infusing anatomical knowledge directly into network design, rather than leaving the network to discover relevant structure in a fully unsupervised manner, further improves recognition performance. This benefit is most pronounced for pathological findings that are spatially sparse, low-contrast, or otherwise difficult to localize against a busy anatomical background [2]. A common architectural pattern for incorporating this kind of anatomical guidance is the dual-stream design, in which anatomical and pathological information are processed through separate representations before being combined, or fused, into a single prediction [3, 4]. The intuition behind this design is straightforward: one stream can specialize in recognizing pathology directly, while the other specializes in anatomical context, and the two are only reconciled at the point of final prediction.

This general pattern, anatomy-guided and often dual-stream deep learning, recurs across otherwise unrelated corners of medical AI, from pelvic fracture detection to chest disease classification to imaging domains entirely outside the scope of this research. A limitation recurs alongside it just as consistently, regardless of the specific task or anatomy involved. Both streams, or the full architectural pipeline more generally, are evaluated on every input

unconditionally. This is not a limitation specific to any one diagnostic task or imaging modality; it is a structural property of how dual-stream and anatomy-guided architectures are typically designed across the field. Whether the case at hand is an unambiguous finding or a genuinely difficult, borderline one, the model applies the same fixed amount of computation and follows the same fixed architectural pathway from input to prediction.

The practical consequence of this design choice is that a large fraction of cases in any given diagnostic task could already be resolved confidently using a single, faster stream, and yet the full dual-stream pipeline still runs regardless. This doubles inference cost across the entire dataset for the sake of a comparatively small and difficult subset of cases that genuinely require the additional anatomical reasoning capacity. The inefficiency here is not merely a matter of wasted compute cycles. In deployment settings where inference cost, latency, or hardware constraints matter, an architecture that pays full cost on every case regardless of need is a harder system to justify and scale than one that reserves its most expensive reasoning for the cases that actually demand it.

This mismatch stands in contrast to how diagnostic effort is allocated in actual clinical practice, across specialties and imaging tasks alike. A clinician does not apply the same depth of analysis to every case they encounter. Straightforward, unambiguous findings are resolved quickly, while ambiguous or borderline cases prompt more deliberate reasoning, additional imaging, or a second opinion from a colleague. This adaptive allocation of diagnostic effort is a general feature of clinical reasoning, not one confined to a particular organ system or disease category, and it is precisely the behavior that current automated systems fail to replicate. Current deep learning systems for medical image classification, across the range of tasks they are applied to, apply a fixed amount of computation to every input regardless of case difficulty, treating an obvious finding and a genuinely ambiguous one identically. Because this fixed-computation limitation is general rather than task-specific, addressing it convincingly requires evidence that any proposed solution generalizes across meaningfully different diagnostic settings, rather than being a narrow fix tailored to a single dataset or pathology. That requirement for cross-task evidence shapes the experimental design of this entire research, and it is the reason two structurally different diagnostic tasks are evaluated rather than one.

1.1.2 A Difficulty Gradient Across Diagnostic Tasks

To demonstrate that an adaptive, conditionally-gated approach to anatomical reasoning is a general solution to this problem, rather than a narrow one, this research evaluates the proposed framework across two deliberately different diagnostic tasks: multi-class chest disease classification from chest X-rays (CXR), and pelvic fracture detection from pelvic X-rays (PXR). These two tasks are not the subject of this research in themselves. They are chosen because they differ substantially in anatomy, pathology, and clinical context, while also spanning a meaningful gradient of diagnostic difficulty. This allows the central hypothesis of this research, that gating behavior can track case difficulty, to be tested under genuinely different conditions rather than a single, favorable one.

Chest X-rays remain the most widely used lung imaging modality in clinical practice, owing to their accessibility, low cost, and comparatively low radiation exposure relative to computed tomography [5]. Pneumonia alone affects a substantial share of the global

population and accounts for a large proportion of pediatric hospital admissions [6, 7]. Tuberculosis continues to be a leading cause of death from a single infectious disease worldwide, with an estimated 1.25 million deaths reported in 2023 [8], and lower respiratory infections more broadly remain a major contributor to global mortality [9]. What makes CXR classification a useful anchor point on the lower end of the difficulty gradient used in this research is not that it is an easy task in any absolute sense, but that many of its constituent findings, once a radiologist is looking at the right region of the image, are visually distinct enough to separate reliably. That said, conditions such as COVID-19 pneumonia, viral pneumonia, tuberculosis, and other lung opacities frequently present with overlapping radiological features, a phenomenon sometimes described as radiological mimicry [10], which increases the likelihood of misdiagnosis and means that even within CXR classification, case difficulty is far from uniform.

Pelvic fractures, by contrast, sit at the more difficult end of the spectrum used in this research. They are typically caused by high-energy trauma, and given the pelvis's anatomical complexity and its proximity to major blood vessels, nerves, and abdominal organs, such injuries can lead to severe complications including hemorrhage and multi-organ damage [11]. In-hospital mortality has been reported to range from 5% to 20%, depending on fracture severity and the presence of hemorrhagic shock [12, 13], and delays in treatment are themselves associated with worse outcomes [14, 15]. This clinical urgency raises the stakes of diagnostic accuracy considerably relative to many CXR findings, where a delayed or borderline diagnosis, while still serious, rarely carries the same immediate risk of hemorrhagic deterioration.

Diagnosis of pelvic fractures relies heavily on radiographic evaluation, and radiographic appearance alone can be an unreliable indicator of injury severity [16]. Subtle or complex fractures are frequently missed even by experienced radiologists [17, 18], and up to 20% of pelvic fractures have been reported as initially overlooked during first-pass evaluation in high-pressure trauma settings [19]. Within pelvic fracture detection itself, a further gradient exists. Some fractures are visibly apparent as cortical disruptions on the X-ray, while others, referred to as invisible fractures, show no visible disruption on the pelvic X-ray at all and are only later confirmed through 3D-CT imaging. An invisible fracture is, by definition, a case where the primary screening modality offers essentially no direct visual evidence of injury, which makes it a meaningfully harder case than even a subtle visible fracture. This three-tier structure, straightforward CXR findings, visible pelvic fractures, and invisible pelvic fractures, provides a natural and increasingly demanding difficulty gradient against which an adaptive gating framework can be evaluated, and it is this gradient, rather than either task in isolation, that the experimental design of this research is built around.

1.2 Motivation

A separate line of research outside anatomy-guided classification addresses computational efficiency directly, through conditional computation, in which the amount of computation applied to an input is adapted to that input rather than reduced or fixed uniformly across an entire dataset [20]. Early-exit networks are the most established instance of this idea. These architectures use a gating mechanism to decide, at one or more points during

inference, whether a sufficiently confident prediction has already been reached and further computation can therefore be skipped [21]. The value of this idea is that it allows a single network to service both easy and hard inputs appropriately, without requiring separate models of different sizes to be trained and selected between ahead of time.

Recent work in this space has examined how the gating mechanism itself should be trained, and has found that aligning training dynamics with the inference-time gating policy, rather than training the classifier and the gate under mismatched objectives, meaningfully improves the reliability of early-exit systems [22]. This finding directly motivates a specific design decision adopted in this research: to rely on explicit, ground-truth-supervised confidence signals for gating, rather than relying solely on implicit signals derived from task-loss gradients or unsupervised confidence heuristics. If a gate is only as trustworthy as its alignment with what it is actually used to decide, then training it directly against ground truth correctness, rather than against a proxy signal, is the more defensible choice.

Confidence-gated designs that route inputs down one of two paths have also been explored outside the early-exit literature, for instance in retrieval-augmented classification, where a dual-path framework is guided by confidence to select between prototype-based and learned representations [23]. This shows that the underlying idea, using a learned confidence signal to route an input down one of several computational paths, is not confined to early-exit architectures specifically. However, correctness-supervised gating specifically over anatomy-guided dual-stream architectures, of the kind used for pelvic fracture detection and CXR classification, has not been explored in prior work. Nor has this kind of gating been evaluated for whether it generalizes across diagnostic tasks of differing difficulty. This is the specific gap this research addresses.

This research is centered on a single underlying question: can a diagnostic model learn to allocate additional anatomical reasoning only when it is actually needed, rather than applying that reasoning uniformly to every case regardless of difficulty, and does this ability generalize across meaningfully different medical imaging tasks rather than being confined to one? Answering this question required moving beyond architectures that treat anatomical guidance as a fixed, always-on architectural component, and instead treating it as a conditional resource, one that is invoked selectively and only when a learned signal indicates that the primary, faster prediction cannot yet be trusted. This reframing needed to hold not just for one task, but consistently across tasks that differ in anatomy, pathology, and underlying diagnostic difficulty, which is why this research evaluates the resulting framework across CXR classification and pelvic fracture detection rather than a single diagnostic setting.

This motivation also continues a broader progression of prior work carried out in the course of this research. Earlier approaches in this research line relied on explicit anatomical priors, most commonly in the form of segmentation-guided regions of interest, to improve pelvic fracture detection accuracy by restricting or emphasizing the model's attention to clinically relevant anatomical structures. These approaches were effective within their specific task, but they are architecturally inflexible in an important sense. They apply the same anatomical processing pipeline to every case regardless of whether that particular case actually requires it, and this fixed design does not generalize easily across diagnostic tasks that differ substantially in their underlying difficulty profile. This observation raised the natural question that this research sets out to answer: rather than fixing anatomical guidance as a permanent, always-on part of the architecture for a single task, could a

model instead learn when anatomical guidance is worth invoking in the first place, using the model's own gating behavior as an implicit, task-agnostic signal of case difficulty? This progression, from explicit, unconditional, task-specific anatomical priors toward adaptive, learned, and demonstrably generalizable gating, forms the conceptual throughline that connects this research to the broader body of prior work it builds upon.

1.3 Research Questions

The motivation outlined above points toward a specific design possibility rather than a vague aspiration: a diagnostic model that allocates anatomical reasoning adaptively, the way a clinician allocates diagnostic effort, rather than uniformly. Turning that possibility into something a research project can actually test requires breaking it down into questions specific enough to be answered with evidence rather than argued for on intuition alone. This report is organized around three such questions, each addressed directly by the results reported in Chapter 6 and interpreted in Chapter 7, and each building on the one before it.

1. **RQ1:** Can anatomical guidance be applied conditionally, through a learned gating mechanism, and match the diagnostic performance of an unconditional, always-on dual-stream architecture at a substantially lower computational cost?
2. **RQ2:** Does a gating mechanism's activation rate, when trained explicitly for correctness prediction rather than relying on unsupervised confidence, correlate with the underlying diagnostic difficulty of the case or task it is applied to?
3. **RQ3:** Does this relationship between activation behavior and task difficulty hold consistently across diagnostically distinct domains, rather than being an artifact specific to a single task or dataset?

RQ1 is the most immediately practical of the three, and the one on which the other two depend. Every anatomy-guided dual-stream design traced in Chapter 3, including PelFANet, this research's own immediate architectural predecessor, evaluates both of its streams unconditionally, on every case, regardless of whether that case is one a single stream could already resolve confidently. This is a reasonable design if diagnostic performance is the only consideration, but it leaves an open question that performance alone cannot answer: how much of that always-on computation is actually necessary, and how much is being spent on cases that did not need it. RQ1 asks whether that unnecessary spending can be recovered, specifically by making the second stream's activation conditional on a learned decision rather than unconditional by architecture, without giving up a meaningful share of the diagnostic performance the always-on design achieves. Answering this question requires more than building a conditional architecture; it requires comparing that architecture directly against its own unconditional counterpart, isolating the effect of conditional gating from every other architectural choice, which is precisely the comparison Chapter 6 is structured around.

RQ2 follows directly from RQ1, but asks a more specific and, in some ways, more consequential question. A gating mechanism could in principle reduce computation

substantially while still answering RQ1 favorably in aggregate, and yet do so by activating close to arbitrarily, on a roughly fixed fraction of cases selected with little regard for which specific cases actually need the second stream's input. Such a mechanism would reduce cost without allocating that cost intelligently, which is a considerably weaker and less interesting claim than the one this research is actually interested in establishing. RQ2 asks whether GateKeeper's specific design choice, training it explicitly as a binary correctness classifier against a ground-truth-derived target, rather than relying on the primary stream's own unsupervised softmax confidence as most prior conditional-computation work does, produces activation behavior that tracks something real: whether the mechanism activates more often specifically on cases and tasks where the primary stream's unaided judgment is less trustworthy, rather than at some difficulty-blind baseline rate. This question sits at the center of what distinguishes SecondOpinion's contribution from a conditional-computation mechanism applied for its own sake.

RQ3 exists because a favorable answer to RQ2 on a single task, evaluated on a single dataset, would not be enough to establish that the underlying gating principle is doing something general. A gating mechanism could appear to track difficulty on one task purely by coincidence, for instance if that task's easy and hard cases happened to differ along some dimension the gate could exploit without that dimension generalizing to any other diagnostic domain. RQ3 asks whether the relationship RQ2 investigates holds across tasks that are genuinely different in kind, not merely different in difficulty within the same underlying domain. This is why this research evaluates SecondOpinion on two structurally distinct diagnostic tasks, multi-class chest disease classification and pelvic fracture detection, the latter further stratified into visible and invisible fracture cases specifically to introduce a within-domain difficulty gradient as well as a between-domain one. If GateKeeper's activation behavior tracks difficulty consistently across both the between-domain comparison, chest disease classification against pelvic fracture detection, and the within-domain comparison, visible against invisible fracture cases, RQ3 is answered with considerably more confidence than either comparison could provide on its own.

1.4 Objectives

To answer the three research questions above, this research pursues two objectives. The first is a design and implementation objective, concerned with building the system capable of generating evidence relevant to RQ1 and RQ2 in the first place; the second is an evaluation objective, concerned with generating and interpreting that evidence across a setting broad enough to speak to RQ3 as well.

1. **Design and implement SecondOpinion**, an adaptive dual-stream framework comprising a fast primary stream evaluated on every input, a second, anatomy-guided stream invoked only when needed, GateKeeper, a lightweight gating mechanism trained explicitly as a binary correctness classifier rather than relying on an unsupervised or heuristically thresholded confidence score, and a lightweight cross-attention fusion module invoked only when GateKeeper activates the second stream, allowing the primary stream's representation to selectively attend to anatomical context rather than fusing both streams unconditionally. Each of these four components is a necessary precondition for the questions above to be answerable at all: without

a genuinely conditional second stream, RQ1's comparison against an always-on baseline has nothing distinct to compare; without a correctness-supervised gate specifically, as opposed to an unsupervised-confidence one, RQ2's question about whether correctness supervision in particular produces difficulty-tracking behavior cannot be isolated from the alternative, more conventional gating strategies surveyed in Chapter 2; and without a fusion mechanism capable of combining the two streams effectively when the gate does activate, any shortfall in SecondOpinion's performance relative to the always-on baseline could be misattributed to gating when it in fact originated in a weak fusion design. This objective is addressed in full in Chapter 5, and its success is assessed directly through the ablation results in Chapter 6, where Stream A only and Stream B only are evaluated in isolation specifically to confirm that each component contributes independently before their conditional combination is assessed as a whole.

2. **Validate SecondOpinion across two diagnostic tasks of increasing and structurally distinct difficulty**, multi-class chest disease classification and pelvic fracture detection, with the latter further stratified into visible and invisible fracture cases, comparing it in each case against both an unconditional, always-on configuration of its own architecture and prior state-of-the-art baselines specific to that task. This objective addresses RQ1 directly, through the accuracy-efficiency comparison against the always-on configuration reported for every task in Chapter 6 and interpreted in Section 7.2, and addresses RQ2 and RQ3 jointly, by testing whether GateKeeper's activation rate tracks diagnostic difficulty consistently not only within a single task, where visible and invisible fracture cases provide a within-domain difficulty gradient, but across tasks that differ structurally in kind, where chest disease classification and pelvic fracture detection provide a between-domain comparison unavailable from either task alone. Deliberately including external, previously published baselines alongside SecondOpinion's own internal ablation configurations in this comparison, rather than evaluating SecondOpinion only against itself, further ensures that any claim of matching or exceeding prior performance is made against results already established in the literature this research builds on, rather than against a baseline constructed specifically to make SecondOpinion appear favorable.

1.5 Report Organization

The remainder of this report is organized as follows.

- **Chapter 2 (Literature Review)** reviews the broader literature this research builds on, covering deep learning in medical image analysis generally, CXR-based and PXR-based diagnostic approaches specifically, anatomy-aware and segmentation-guided methods, fusion-based learning, dual-stream architectures, and efficiency-aware deep learning, before closing with the specific research gap, correctness-supervised gating over anatomy-guided dual-stream architectures, that this research addresses.
- **Chapter 3 (Prior Approaches)** traces the progression of prior work that directly informed this research. It covers six successive pelvic fracture detection efforts in chronological order, from an initial architecture benchmark through preprocessing, patch-ensemble, multi-bone segmentation (Beyond the Pelvic Ring), fusion-ensemble,

and PelFANet studies, followed by two prior CXR classification efforts, an architecture comparison and CheXNet-CBAM. PelFANet is identified as this research's immediate architectural predecessor and the always-on baseline SecondOpinion departs from.

- **Chapter 4 (Dataset)** describes the two datasets used to train and evaluate SecondOpinion: the unified five-class CXR dataset, drawn from the COVID-19 Radiography Database and the TB Chest X-ray Dataset, and the private AMERI pelvic fracture dataset, collected under IRB approval at Steel Memorial Hirohata Hospital. For each, the chapter specifies segmentation mask provenance, ground-truth versus predicted masks, preprocessing, and the split into visible and invisible fracture subsets used for training and held-out generalization testing respectively.

- **Chapter 5 (Proposed Method)** presents SecondOpinion in full: the dual-stream EfficientNet-B0 architecture, the cross-attention fusion module, GateKeeper's design as a correctness-supervised binary gating mechanism and its hard-gate threshold, the two-phase training strategy (frozen GateKeeper, then unfrozen with soft gating), hard gating at inference, and the evaluation framework, including the classification metrics and activation-rate definition used throughout Chapter 6.

- **Chapter 6 (Result)** reports experimental results across all three evaluation settings, CXR classification, visible pelvic fracture detection, and invisible pelvic fracture detection. It presents internal ablation results (Stream A only, Stream B only, Always-On, SecondOpinion) and comparisons against external baselines for each task, an efficiency analysis of Params, FLOPs, and activation rate across the difficulty gradient, and a qualitative GradCAM analysis of two representative cases.

- **Chapter 7 (Discussion)** interprets the results reported in Chapter 6 against the three research questions posed above, examines the accuracy-efficiency and recall trade-offs observed on the CXR and pelvic tasks, discusses the coupling between GateKeeper's activation rate and the AUROC gap as evidence for correctness-supervised gating, situates SecondOpinion against PelFANet, early-exit, and confidence-gated dual-path literature, and states the limitations of the present evaluation.

- **Chapter 8 (Conclusion)** closes the report by answering the three research questions directly, summarizing the key findings, considering social, environmental, and ethical dimensions of the work, and outlining concrete directions for future work arising from Chapter 7's limitations.

Chapter-2: Literature Review

This chapter situates the proposed framework within the broader landscape of deep learning for medical image analysis. It begins with general trends that have shaped the field, then narrows through diagnostic approaches specific to chest and pelvic radiography, anatomy-aware and fusion-based architectures, and efficiency-oriented designs, before identifying the specific gap this research addresses.

2.1 Deep Learning in Medical Image Analysis

Deep learning has become the dominant approach for medical image analysis over the past decade, with convolutional neural networks in particular demonstrating strong performance across disease classification, segmentation, and detection tasks spanning a wide range of imaging modalities [24]. Much of this progress traces back to a shift away from classical, feature-engineered pipelines toward architectures that learn features and classification jointly, end-to-end, directly from data. In medical imaging specifically, this shift matters more than it might in some other domains, because the visual cues separating pathological from healthy tissue are frequently subtle, spatially distributed, or difficult for a human expert to articulate as an explicit rule in the first place. Hand-designing features for such cues is genuinely hard, and deep learning's ability to learn them automatically has been a large part of why the field moved so quickly toward CNN-based approaches once sufficient data and compute became available.

Transfer learning using architectures pretrained on large-scale datasets such as ImageNet [25] has further enabled effective learning in medical settings where labeled data are limited [1]. This matters because labeled medical data is expensive to produce. Annotation typically requires a trained clinician, and for some tasks, ground truth is only available after a delay, for instance once a suspected finding is confirmed by follow-up imaging or surgery. Pretraining on millions of natural images and then fine-tuning on a comparatively small medical dataset lets a model arrive already equipped with general-purpose visual features, so the fine-tuning stage only has to adapt those features to the specific diagnostic task rather than learn them from scratch. This general-purpose effectiveness, learned features plus transfer learning, has motivated the adoption of CNN-based approaches for the two diagnostic tasks central to this research, chest disease classification and pelvic fracture detection, both reviewed in detail below.

This general pattern, benchmarking CNN and, increasingly, transformer-based architectures against a shared classification task, recurs across medical imaging tasks far removed from

chest and pelvic radiography specifically. In liver disease detection from ultrasound images, a comparison of EfficientNet, MobileNet, and ConvNeXt variants on a benign, malignant, and normal classification task found MobileNetV2 to achieve the strongest overall performance, at 78% accuracy, 78% recall, and 0.89 AUROC, with an inference time of 28.47 ms, while EfficientNetB2 offered a comparably strong balance across metrics at 77% accuracy and 74.68 ms inference time; ConvNeXt models were competitive but not superior, managing an average accuracy of 74% at a higher inference cost [26]. A comparative evaluation of six CNN models for classifying pulmonary edema severity in chest X-rays, incorporating Grad-CAM visualization for interpretability, found CheXNet to perform best, at 91% accuracy and an AUC of 0.88, and reported that pretraining specifically on chest X-ray data improved performance over generic pretraining [27]. A comparison of five pretrained CNN architectures, ResNet50, DenseNet121, EfficientNetB3, InceptionV3, and MobileNetV2, for distinguishing benign lipomas from well-differentiated liposarcomas on T1-weighted MRI found MobileNetV2 to achieve the highest accuracy at 0.967, while EfficientNetB3 performed considerably worse at 0.635 accuracy, with a strong prediction bias toward malignancy reflected in an 89% false positive rate on benign lipomas [28]. A ten-model benchmark spanning both CNNs, ResNet-50, ResNet-101, DenseNet-121, DenseNet-201, and three EfficientNet variants, and Vision Transformers, ViT-16 and ViT-32, for brain tumor classification on the BraTS dataset found ViT-32 to achieve the highest performance overall, at 83.05% accuracy, 78% precision, 70% recall, and 79% F1-score, though at a greater computational expense than the CNN alternatives [29]. A bilateral ensemble framework using a ConvNeXt-XLarge backbone for multi-class retinal disease classification on the ODIR-5K dataset, evaluated across five categories, Normal, Diabetic Retinopathy, Cataract, Age-Related Macular Degeneration, and Myopia, found that combining bilateral feature fusion with color-corrected preprocessing achieved the highest overall accuracy, at 74.5%, relative to single-eye and non-color-corrected configurations [30]. A comparison of four architectures representing distinct design paradigms, DenseNet121 as a pure CNN, ConvNeXt as a modernized CNN with transformer-inspired design, ViT-B/16 as a pure Vision Transformer, and Swin-B as a hierarchical transformer, for breast ultrasound classification on the BUSI dataset found ConvNeXt to outperform DenseNet121 on every metric, at 83% accuracy and 91% AUROC, ViT-B/16 to achieve the highest AUROC at 94%, and Swin-B to attain the most balanced performance, at 86% accuracy, 85% recall, and 0.92 AUROC [31]. A study of augmentation's effect on four CNN models for classifying congenital heart disease from chest X-rays into four categories, normal, atrial septal defect, ventricular septal defect, and patent ductus arteriosus, found EfficientNetB1 with moderate augmentation to achieve the highest test accuracy at 69.51%, followed by DenseNet121 at 66.46%, with MobileNetV2 performing worst of the four [32]. A comparison of five CNN architectures, ResNet-50, ResNet-101, DenseNet-121, MobileNetV2, and NASNet-Mobile, for multiclass classification of reflective dental images across five categories, Dental Calculus, Dental Caries, Hypodontia, Mouth Ulcer, and Tooth Discoloration, on a Kaggle-sourced dataset of 5,048 images preprocessed with CLAHE and class-aware augmentation, found ResNet-101 to deliver the best overall performance after fine-tuning, at approximately 97% accuracy and a macro F1-score of approximately 94%, with dental caries and small, low-contrast lesions identified as common failure modes through visual inspection and saliency maps [33].

The same architecture-benchmarking pattern extends to segmentation tasks across an equally broad range of anatomical targets and imaging modalities. A morphology-aware preprocessing pipeline paired with a hybrid Segformer architecture, tested with ResNet18, EfficientNet-B3, MobileNet-V2, and DenseNet201 encoders on the Brain Tumor Progres-

sion Dataset, found EfficientNet-B3 to yield the best glioblastoma segmentation results, improving from an initial precision of 65.85%, recall of 61.68%, and Dice score of 63.70%, to a precision of 79.51% and a Dice score of 66.45% once the ground-truth masks were preprocessed to remove noise and reconnect broken regions [34]. An encoder-decoder pipeline built around a U-Net with a ResNet-18 backbone, incorporating Hounsfield-unit windowing tailored to lung tissue, for joint lung and pulmonary nodule segmentation on 888 cases from the LUNA16 thoracic CT dataset achieved mean intersection-over-union scores of 0.994 for the left lung, 0.960 for the right lung, and 0.741 for nodular structures, with an aggregate Dice similarity coefficient of 0.947 [35]. A comparison of eight CNN architectures, including variants of ResNet and EfficientNet, against two SegFormer models using a Mix-Transformer backbone, for stenosis segmentation in coronary X-ray angiography found the SegFormer-MiT-B3 architecture to achieve the highest performance, at a Dice Score Coefficient of 0.632 and an Intersection over Union of 0.462, exceeding every evaluated CNN [36]. A cross-modal benchmark of four U-Net encoder variants, ResNet18, EfficientNetB3, DenseNet121, and MobileNetV2, for liver segmentation, trained on an MRI dataset and externally validated on three separate CT and MRI datasets, found DenseNet121 to achieve the best Dice Similarity Coefficient on the primary MRI test set, at 0.928, while EfficientNetB3 generalized better under external validation, reaching Dice scores of 0.825 on CT and 0.852 on MRI [37]. A multi-strategy optimization pipeline, combining CLAHE, edge enhancement, denoising, stochastic augmentation, and boundary-aware loss, applied across five U-Net variants for segmenting dental structures in 361 annotated orthopantomograms found Vanilla U-Net to lead the comparison at 91.38% Dice and 84.13% IoU, ahead of UNet++ by 0.32 points of Dice and 0.54 points of IoU [38]. A two-stage U-Net pipeline for liver and hepatocellular carcinoma segmentation on the LiverHCCSeg MRI dataset, segmenting the liver first and then the tumor within a cropped region derived from the predicted liver mask, achieved a Dice Similarity Coefficient of 0.925 for liver segmentation, while tumor segmentation proved more difficult, improving from an initial Dice score of 0.312 to 0.473 after iterative curation of the train-test split [39].

Beyond CNN- and transformer-based classification and segmentation benchmarking specifically, recent work has also explored agentic, vision-language-model-based frameworks for medical image analysis. One such framework decomposes the medical imaging workflow into four coordinated agents, a Router that configures task- and specialization-aware prompts from the image, patient history, and metadata, a Retriever that uses exemplar memory and pass@k sampling to jointly generate free-text reports and bounding boxes, a Reflector that critiques each draft-box pair for clinical error modes including negation, laterality, unsupported claims, contradictions, missing findings, and localization errors, and a Repairer that iteratively revises both the narrative and spatial outputs while curating high-quality exemplars for future cases. Instantiated on chest X-ray analysis across multiple vision-language model backbones and evaluated on report generation and weakly supervised detection, this framework improves LLM-as-a-Judge scores by roughly 1.7 to 2.5 points and mAP50 by 2.5 to 3.5 absolute points over strong single-model baselines, without any gradient-based fine-tuning [40].

2.2 CXR-Based Diagnostic Approaches

Chest X-ray classification has been one of the most extensively studied applications of deep learning in medical imaging, helped considerably by the early availability of large public datasets. Chief among these is ChestX-ray14, which contains 112,120 frontal chest radiographs spanning 14 disease categories [41]. A dataset of this size was unusual for medical imaging when it was released, and it enabled a wave of subsequent work to benchmark increasingly capable architectures against a shared, standardized task rather than each working from its own small, incomparable dataset.

A range of established CNN backbones have been benchmarked on CXR classification after ImageNet pretraining, including ResNet [42], DenseNet [43], MobileNetV2 [44], and EfficientNet [45]. These architectures differ in how they are built, ResNet through residual connections that make very deep networks trainable, DenseNet through dense feature reuse across layers, MobileNetV2 through inverted residuals designed for efficient deployment, and EfficientNet through a principled scaling rule that balances depth, width, and resolution, but all have been shown capable of strong CXR classification performance once fine-tuned.

The reported numbers across this literature are strong, though they vary with dataset composition and class count. ResNet101 has been reported to achieve 96% accuracy on a three-class dataset of COVID-19, non-COVID, and normal cases, outperforming ResNet50, DenseNet121, and InceptionV3 on the same task [46]. DenseNet201 has performed comparably or better, with one study reporting 99.1% accuracy on a similarly structured three-class dataset [47]; this may reflect that DenseNet's dense connectivity is well suited to picking up the subtle, texture-level differences that separate COVID-19, other pneumonias, and normal lungs on a chest film, differences that are often more about overall texture than about a single, sharply localized lesion. MobileNetV2, despite being far more lightweight than either of the above, has still outperformed AlexNet, DenseNet121, and InceptionV3 on the considerably harder 14-class ChestX-ray14 benchmark [48], which suggests that on CXR tasks specifically, a smaller, more efficient architecture does not automatically mean a less accurate one. EfficientNetB3 achieved 95.28% accuracy on a four-class CXR dataset, a result its authors frame explicitly around the balance it strikes between accuracy and computational cost [49].

A particularly influential contribution in this space is CheXNet, a DenseNet121 model fine-tuned specifically on ChestX-ray14 rather than relying on generic ImageNet pretraining alone [1]. What makes CheXNet a useful reference point is less its architecture, which is a fairly conventional DenseNet variant, and more its framing: it was one of the earlier works to argue, with supporting evidence, that domain-specific pretraining rather than architectural novelty alone could close much of the gap between automated systems and expert radiologists on pneumonia detection. That argument has held up reasonably well as later work continued to find that pretraining choices matter as much as, and sometimes more than, backbone choice.

Taken together, this body of work establishes two things relevant to this research. First, that CNN architectures, particularly with domain-relevant pretraining, are capable of strong, often near-expert performance on multi-class CXR classification, which is why a CNN-based primary stream is a reasonable starting point for the proposed framework.

Second, that this performance has generally been achieved using architectures that apply one fixed processing pathway to every input, with no explicit mechanism for recognizing or responding to variation in per-case diagnostic difficulty. That second point becomes more consequential in the pelvic fracture literature reviewed next, where difficulty varies more sharply from case to case.

2.3 PXR-Based (Pelvic X-Ray) Diagnostic Approaches

Deep learning applied to pelvic and femur fracture detection has also produced strong results, though the underlying task differs from CXR classification in an important way. Rather than distinguishing between several disease categories based on relatively global image-level patterns, pelvic fracture detection often depends on identifying a comparatively localized, and sometimes visually subtle, structural discontinuity set against otherwise normal anatomy. Reported accuracies range from roughly 80% to 98%, depending heavily on the dataset and fracture type [50, 51]. Kassem et al. developed an explainable transfer learning framework for pelvis fracture detection and reported 98.5% accuracy on 876 pelvic radiographs [50]. The emphasis on explainability here is not incidental. Because a missed or incorrect pelvic fracture diagnosis carries direct clinical consequences, several works in this space have prioritized some form of interpretability or localization evidence alongside raw accuracy, rather than treating accuracy as sufficient on its own.

Tanzi et al. approached proximal femur fracture classification with a multistage pipeline, structuring the task hierarchically rather than as one flat multi-class problem [51]. Their headline number, 86% accuracy on a three-class task, is less interesting on its own than their finding that the system improved human specialist performance by 14% when used as a decision aid, which points toward a role for these systems as augmentation for clinical judgment rather than a replacement for it. Cheng et al. took a more direct approach, training a CNN on pelvic radiographs to detect trauma at a level comparable to practicing physicians, reporting an AUC of 0.98 and using Grad-CAM to localize the regions driving each prediction [52]. Zheng et al. extended similar methods to four distinct hip fracture types, a finer-grained task than simple detection, and reported high performance across all evaluated metrics [53]. More recently, PelviNeXt combined a dense convolutional feature extractor with hierarchical channel-spatial attention, multi-scale fusion, and self-attention refinement to set a new state-of-the-art on PXR150 across accuracy, recall, specificity, and AUROC [54], continuing this trend toward architectures that combine convolutional and attention-based components rather than relying on either family alone. Across these studies, the common thread is that deep learning adapts well to increasingly specific pelvic and hip fracture subtasks, not just to fracture detection in the broadest sense.

A particularly relevant strand of this literature, and one that bears directly on how the pelvic dataset in this research is stratified, concerns the distinction between fractures that are visibly apparent on a radiograph and those that are not. Rahman et al. proposed a two-step transfer learning approach built specifically around this distinction: first pretraining a network on digitally reconstructed radiograph (DRR) images synthesized from 3D-CT scans, then fine-tuning on real pelvic X-rays [55]. The idea behind this pretraining strategy was to expose the network to fracture patterns from the richer 3D-CT modality before it ever saw the comparatively information-poor 2D projection, under the assumption that

this would help the network recognize fracture-relevant patterns even when those patterns are only weakly expressed in the final 2D image. Their results support this to a meaningful degree: AUROCs of 0.933 for visible fractures and 0.801 for invisible fractures.

That gap between visible and invisible fracture performance is worth sitting with, because it does two things at once. It shows that automated pelvic fracture detection is feasible even without obvious cortical disruption, since 0.801 AUROC is well above chance. And it shows, quantitatively, that invisible fractures remain a measurably harder subtask than visible ones, even for a model specifically pretrained with that distinction in mind. This is not a small or artificial gap; it is close to 0.13 AUROC, a substantial difference in a metric that is often compressed near its ceiling for well-studied tasks. This measured difficulty gap is what motivates treating visible and invisible fractures as genuinely separate difficulty tiers in the pelvic evaluation later in this research, rather than lumping them together as one undifferentiated fracture detection task.

2.4 Anatomy-Aware Approaches

A substantial body of work improves diagnostic accuracy by first localizing anatomically relevant regions of interest, then performing classification on that localized region rather than on the full, unmodified image. The reasoning is intuitive. Much of a typical radiograph consists of background structures or anatomy that is not directly relevant to the pathology being diagnosed, and this irrelevant content can dilute or distract a classifier's learned representation away from the comparatively small region where diagnostically meaningful information actually lives. Infusing anatomical knowledge directly into network design this way, rather than expecting a network to discover relevant structure entirely unsupervised, has been shown to improve recognition performance, and this improvement is most pronounced for findings that are spatially sparse, low-contrast, or otherwise easy for a model to overlook amid surrounding visual noise [2].

This segmentation-guided approach has proven effective across imaging domains well beyond pelvic or chest imaging, which suggests it is a genuinely general architectural pattern rather than something narrowly tuned to one task. In colorectal cancer imaging, SG-TransUNet uses a segmentation-guided transformer U-Net to identify KRAS gene mutation status, using segmentation as an intermediate step that focuses the network's attention before the final classification decision [56]. The APEx framework goes further still, jointly modeling anatomical and pathological features with a query-based segmentation transformer, and reporting improved performance on both FDG-PET-CT and chest X-ray datasets, two modalities with quite different imaging physics [57].

Within pelvic imaging specifically, this literature has developed along two related tracks: segmentation quality on its own, and segmentation as an intermediate step feeding a downstream fracture classifier. Lee et al. did comprehensive work on the first track, evaluating U-Net, Attention U-Net, and Swin U-Net [58] on 940 pelvic X-ray images. Swin U-Net came out ahead, with a Dice coefficient of 96.32%, sensitivity of 96.77%, specificity of 98.50%, and accuracy of 98.03% [59]. This study is thorough on segmentation quality, but it stops there, without connecting those numbers to how well the resulting segmentation actually supports downstream fracture classification, leaving open exactly

the question a clinically-oriented reader would want answered.

A later study by Lee et al. picked up that thread, incorporating segmentation directly into a classification pipeline built around the Association for Osteosynthesis Foundation and Orthopedic Trauma Association (AO/OTA) fracture taxonomy [60]. Segmentation Dice scores landed between 0.96 and 0.97, broadly consistent with the earlier study, but downstream classification accuracy ranged more widely, from 0.69 to 0.88 across fracture types. That spread is informative on its own. Segmentation quality was consistently high and tightly clustered, yet classification accuracy varied considerably by fracture type, which suggests that accurate localization of the bone is necessary but not sufficient for reliable classification. Some fracture types are simply harder to distinguish correctly even once the model is looking at exactly the right region.

A limitation that recurs across this segmentation-guided literature, and one that matters directly for the motivation of this research, is that fracture diagnosis in pelvic radiographs can depend on information that lies outside the segmented bone boundary itself. Some fractures present as clear cortical disruptions that a segmentation-focused model should catch. Others show up through subtler, non-local signs, abnormal alignment, altered joint spacing, or asymmetric limb positioning, cues that can sit entirely outside the segmented bone region and are therefore at risk of being discarded when analysis is restricted strictly to the segmented structure [52]. This is not a small caveat on an otherwise complete approach. It points to a real tension between the benefit of anatomical focus and the risk of losing useful context by focusing too narrowly, and that tension is exactly what motivates the fusion-based and dual-stream approaches discussed next, both of which try to keep localized anatomical detail and broader image context available at the same time, rather than committing to one at the expense of the other.

2.5 Fusion-Based Learning in Medical Imaging

Fusion strategies integrate complementary information from multiple sources within a single pipeline, whether that means combining different imaging modalities, different views of the same anatomy, or, most relevant here, different representations of the same image, such as a raw image alongside its segmented counterpart. The case for fusion follows directly from the limitation raised at the end of the previous section. If neither a raw image alone nor a segmented representation alone reliably captures everything diagnostically relevant, combining both is a natural way to recover what either one, used alone, would leave out.

Fusion approaches are typically grouped into three levels: pixel-level, where raw inputs are combined before any feature extraction; feature-level, where learned representations from separate pathways are combined partway through the network; and decision-level, where independent predictions are combined only at the final output. Feature-level fusion has generally performed best across this literature, likely because it combines representations while they are still rich and specific, before the compression that happens when each pathway is reduced all the way down to a final decision [61].

Generative approaches have extended fusion beyond simple concatenation or averaging

of features. MedFusionGAN uses an unsupervised generative adversarial network to fuse cross-modality inputs such as CT and MRI, aiming explicitly to preserve both structural detail and soft-tissue contrast that simpler fusion methods tend to blur together [62]. MMIF-Net uses multi-scale hybrid attention to fuse anatomical and functional imaging within one network, weighting each modality's contribution differently at different spatial scales rather than combining them uniformly everywhere [63].

Despite this broader interest in fusion, strategies that specifically combine raw and segmented representations of the same image remain comparatively underexplored for fracture detection from X-rays. That gap matters, not as an abstract observation but concretely: since raw radiographs preserve global context that segmentation-only pipelines discard, a well-designed fusion of the two representations should plausibly outperform either one used alone, yet this specific hypothesis has not been tested directly and systematically for pelvic fracture detection. That absence is part of what motivates the architectural direction taken later in this research.

2.6 Dual-Stream Architectures in Medical Imaging

Dual-stream architectures process two separate representations of an input, or two related but distinct inputs, through parallel branches before combining their outputs into a single prediction. This is, in a sense, a concrete architectural commitment to the fusion strategies discussed above. Rather than fusion being an abstract technique applied at some unspecified point, a dual-stream design commits to two parallel pathways early, with fusion happening at a defined point once both streams have produced sufficiently developed representations.

A few recent examples illustrate the pattern across different domains. Bruno et al. proposed a dual-stage framework for breast ultrasound segmentation and classification, using separate stages for the two objectives rather than one shared pathway trying to do both [3]. Zafar et al. introduced a dual-stream contrastive latent learning GAN for brain image synthesis and tumor classification, using parallel streams to learn complementary latent representations that a single stream would struggle to disentangle on its own [4]. Despite applying to different anatomies and different tasks, both examples share the same underlying commitment: two parallel pathways, each contributing something distinct toward a final, fused prediction.

A pattern recurs across this dual-stream literature that matters more to this research than any individual result. Both streams are evaluated on every input unconditionally, whether or not the second stream is actually needed to resolve that particular case correctly. This is true of the dual-stream work reviewed directly here, and it is equally true of the segmentation-and-fusion work discussed in the two preceding sections. In every case, the computational cost of the second stream is paid on every input in the dataset, rather than reserved for the subset of cases that genuinely benefit from it. None of this work includes any mechanism, learned or otherwise, for deciding on a per-case basis whether the second stream is worth invoking at all. That uniform, unconditional invocation is the specific structural limitation the rest of this chapter, and this research as a whole, is built to address.

2.7 Efficiency-Aware Deep Learning in Medical Imaging

A separate line of research addresses computational efficiency not through architectural compactness, which was the concern of several lightweight CXR architectures reviewed earlier, but through conditional computation, where the amount of computation applied to an input adapts to that input rather than staying fixed across an entire dataset [20]. This is a meaningfully different strategy for achieving efficiency. Rather than shrinking the architecture uniformly for every input, conditional computation keeps the full computational capacity available for the inputs that need it, while cutting cost specifically for the ones that do not.

Early-exit networks are the most established version of this idea. These architectures place one or more decision points during inference, each governed by a gate that decides whether a sufficiently confident prediction has already been reached, allowing the network to stop early rather than run through its remaining layers [21]. The appeal is that one trained network can serve inputs of varying difficulty with varying amounts of computation, without needing separate models of different sizes trained and selected between ahead of time.

A question that matters directly to the gating mechanism used in this research is how the gate itself should be trained. Recent work found that aligning the gate's training dynamics with its actual inference-time decision policy, rather than training the classifier and the gate under separate or loosely coupled objectives, meaningfully improves how reliable the resulting early-exit decisions are [22]. The underlying principle here generalizes beyond early-exit specifically: a gate is only as trustworthy as the alignment between what it is trained to predict and what it is actually used to decide at inference time. This is exactly why the gating mechanism used in this research is trained explicitly as a correctness classifier, directly supervised by ground-truth labels, rather than relying on an implicit or heuristically derived confidence signal pulled indirectly from the primary classifier's output.

Confidence-gated dual-path designs, structurally similar to early-exit gating even outside that specific literature, have started appearing in adjacent corners of medical image classification too. Tang and Zhao proposed a teacher-guided dual-path multi-prototype retrieval-augmented framework for fine-grained medical image classification, where a confidence signal governs routing between a prototype-based retrieval path and a directly learned representation path [23]. This confirms that confidence-gated routing between two computational paths is a workable, increasingly explored idea in medical image classification more broadly, not something confined to early-exit architectures alone. What has not been done, in this work or the early-exit literature discussed above, is applying this kind of confidence-gated routing to gate an anatomy-guided dual-stream architecture specifically, or testing whether such gating behaves consistently across diagnostic tasks that differ substantially in difficulty.

2.8 Research Gap

The literature reviewed in this chapter points to a gap that only becomes visible once its separate threads are considered together. Anatomy-guided and dual-stream architectures, covered in Sections 2.4 and 2.6, consistently improve diagnostic performance by building explicit anatomical structure into the network, whether through segmentation-guided region localization or through parallel dual-stream processing. Fusion-based approaches, covered in Section 2.5, show that combining raw and anatomically-guided representations can preserve information that either one alone would discard, directly addressing the limitation identified in the segmentation-guided literature. But every architecture across these three sections runs its full pipeline, and therefore its full anatomical guidance component, unconditionally on every input, whether or not that particular case actually needs the extra reasoning to be resolved correctly.

The conditional computation literature reviewed in Section 2.7 developed independently of this anatomy-guided work. It has established that properly supervised gating mechanisms can reliably decide when additional computation is needed for a given input, improving efficiency without sacrificing accuracy. That literature is mature enough to offer clear design guidance, particularly around training the gate against a directly supervised, ground-truth signal rather than an implicit proxy. What it has not done is apply that guidance inside anatomy-guided medical image analysis specifically, and it has certainly not used gating to decide when an anatomy-guided second stream should be invoked in a dual-stream diagnostic architecture.

No prior work identified in this review combines correctness-supervised gating with an anatomy-guided dual-stream architecture. No prior work tests whether such a combination would generalize across diagnostic tasks that differ in difficulty, such as CXR classification and pelvic fracture detection with its visible and invisible subtypes. This is the specific gap this research addresses, and it is what motivates the design of SecondOpinion and its GateKeeper gating mechanism, presented in Chapter 5.

Chapter-3: Progression of Prior Approaches

The literature reviewed in Chapter 2 situates this research within the broader field. This chapter takes a narrower and more personal view. It traces the specific sequence of studies, conducted as part of the same research effort, that led directly to SecondOpinion. Each section below examines one study in that sequence: what it set out to do, what it found, and, most importantly, what limitation it left exposed. That exposed limitation is not incidental. It is the starting point for the study that follows. Read in order, these six studies on pelvic fracture detection, together with the CXR work discussed at the end of the chapter, tell the story of how this research arrived at its central idea.

3.1 Prior Approaches to Pelvic Fracture Detection

3.1.1 Architecture Benchmark

The starting point for this research line was a comparatively basic question: given the same dataset and the same task, which deep learning architecture actually performs best on pelvic fracture detection, and why? At the time, the pelvic fracture literature offered plenty of individual results but very little direct, controlled comparison across architectures. Kassem et al. had reported 94.7% accuracy using a ResNet50-based transfer learning model [50], a figure often cited as evidence that pretrained convolutional networks can recognize fracture-related patterns reliably. Krogue et al. took a different backbone entirely, using a DenseNet-based architecture to both detect hip fracture position and classify fracture type on pelvic X-rays, reporting 93.7% accuracy for binary detection and 90.8% for multi-class recognition [64]. Kitamura used DenseNet121 for a related but distinct task, jointly predicting patient position, surgical hardware presence, and fracture status, reporting an AUC of 0.75 for pelvic fractures specifically, alongside notably higher AUCs of 0.95 and 0.85 for proximal femoral and acetabular fractures respectively [65]. This spread in AUC across fracture subtypes within a single study is itself informative: it suggests that "pelvic fracture detection" is not one uniform task but a collection of subtasks of varying intrinsic difficulty, a theme that resurfaces later in this chapter.

CheXNet, originally developed for chest radiographs, was also adapted to pelvic imaging. Zhang et al. used it as a baseline for pelvic fracture detection, reporting an AUROC of 0.971, and then improved on that baseline with a point-annotation framework based on window

losses, reaching 0.983 [66]. Elsewhere in orthopedic radiology, newer architectural families were beginning to appear. ConvNeXt, a convolutional design that borrows structural ideas from transformer architectures, such as larger kernel sizes and layer normalization, had shown competitive results on general image benchmarks [67] and had been applied successfully to small, imbalanced orthopedic radiograph datasets [68]. Vision Transformers, which dispense with convolution entirely and rely solely on self-attention [69], had likewise begun appearing in femur fracture work, with one study reporting 83% accuracy using a Vision Transformer for sub-fracture detection [70], and a systematic review reporting figures as high as 92% accuracy and an AUC of 0.94 for Vision Transformer-based femoral fracture analysis [71].

Taken individually, each of these results reads as a point in favor of its respective architecture. Taken together, they are difficult to compare meaningfully, because each came from a different study, a different dataset, a different class balance, and often a different evaluation protocol entirely. A model reporting 98% accuracy in one study and a different model reporting 83% in another does not necessarily tell us anything about which architecture is intrinsically better suited to pelvic fracture detection; it may simply reflect differences in how each study was set up. This ambiguity was the specific gap this study addressed: rather than adding one more architecture to the pile, it held the dataset, the training protocol, and the evaluation procedure fixed, and varied only the architecture itself.

Five architectures were compared under this controlled setup, using the PXR150 dataset and identical five-fold cross-validation: ResNet50, representing the classic residual-connection CNN family; DenseNet121, representing dense feature reuse across layers; CheXNet, a DenseNet121 variant distinguished only by its medical-domain pretraining; ConvNeXt-Tiny, a modernized convolutional design; and ViT-B/16, a pure transformer architecture with no convolutional layers at all. Rather than treating this as a simple leaderboard, the comparison was structured around six specific paired contrasts, each isolating one design variable at a time, for instance ResNet50 against ConvNeXt-Tiny to isolate the effect of architectural modernization alone, or DenseNet121 against CheXNet to isolate the effect of domain-specific pretraining alone, holding backbone architecture constant.

ConvNeXt-Tiny came out ahead on every metric evaluated, reaching an accuracy of 0.8200 and an AUROC of 0.8280, with ViT-B/16 close behind at 0.8133 accuracy and 0.8200 AUROC. ResNet50 finished last on every metric, at 0.7733 accuracy and 0.7830 AUROC. Table 3.1 summarizes the full comparison.

Table 3.1: Comparative performance of five architectures on PXR150.

Architecture	Accuracy	Optimal Specificity	Optimal Recall	AUROC
ResNet50	0.7733	0.7400	0.7800	0.7830
DenseNet121	0.7800	0.7600	0.7900	0.7970
CheXNet	0.8000	0.7800	0.8200	0.8100
ViT-B/16	0.8133	0.7900	0.8500	0.8200
ConvNeXt-Tiny	**0.8200**	**0.8000**	**0.8600**	**0.8280**

CheXNet outperformed DenseNet121 despite sharing an identical backbone, 0.8100 AUROC against 0.7970, which confirms in a controlled setting what earlier, less comparable studies had only suggested: domain-specific pretraining measurably helps. But the more striking finding sits one row above that comparison. ConvNeXt-Tiny, with no domain-

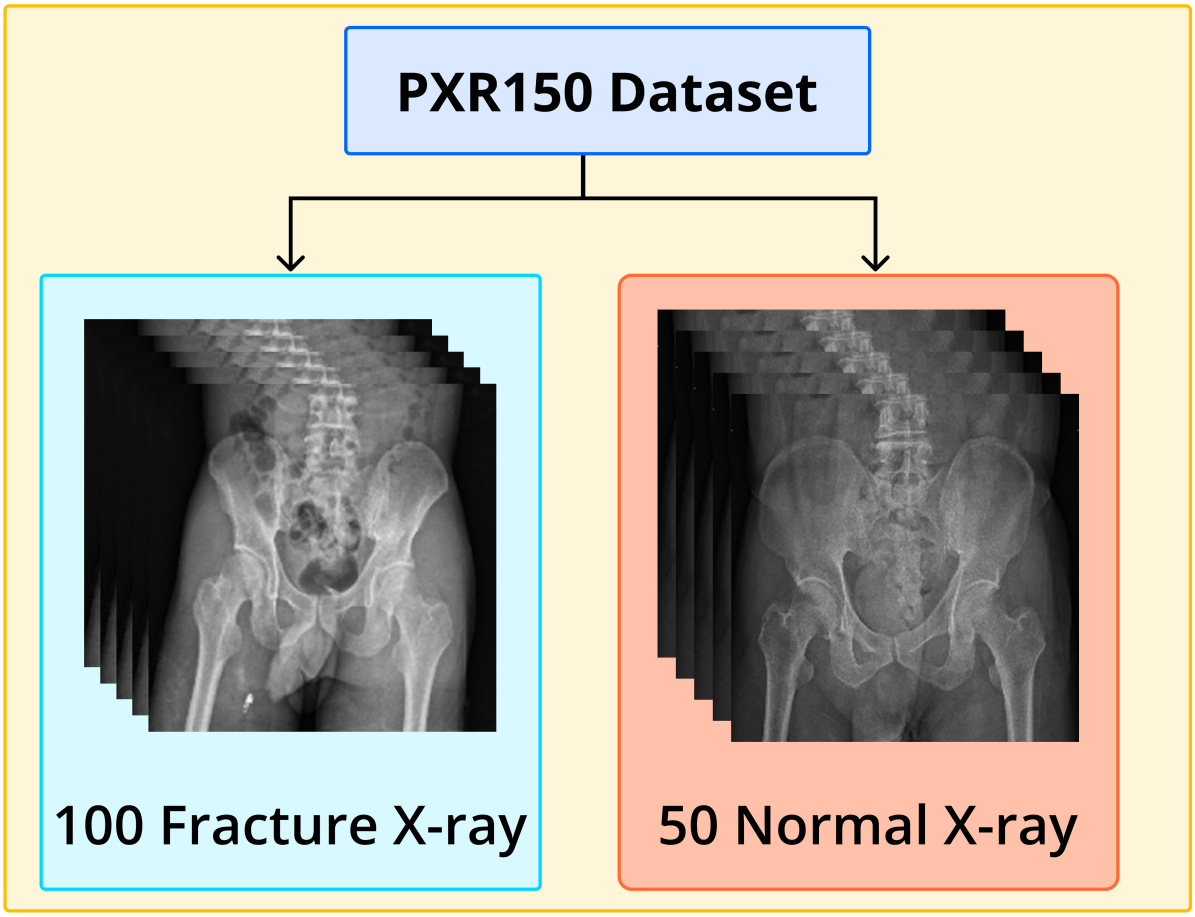

Figure 3.1: Overview of the PXR150 dataset composition used for architecture benchmarking.

specific pretraining at all, still outperformed CheXNet on every metric. Architectural modernization, in other words, was able to substitute for, and in this case exceed, the benefit of medical pretraining, at least on this task and this dataset. This is a genuinely useful finding for anyone choosing a backbone for pelvic fracture work, since it suggests that architecture choice and pretraining strategy are not strictly ordered in importance; a sufficiently well-designed architecture can close, or even overturn, a pretraining gap.

This result gave later work in this research line a defensible architectural starting point, and it settled a question that the existing literature had left ambiguous. But it also left an obvious limitation exposed, one the study's own authors acknowledged directly in their conclusion. Every one of the five architectures compared here was operating on exactly the same input: the full, raw pelvic radiograph, unmodified except for standard geometric augmentation such as rotation, shearing, and flipping. The study answered, convincingly, which architecture extracts features best from a raw image. It never asked whether the raw image itself, with its characteristically low contrast, uneven illumination, and often subtle fracture lines, was giving any of these five architectures a fair chance to begin with. If the input itself were improved before it ever reached the network, could performance rise across the board, independent of architecture choice? That question, whether preprocessing rather than architecture was the more fundamental lever, became the starting point for the next study in this line.

3.1.2 Preprocessing Matters

The Architecture Benchmark study had settled one question and immediately raised another. If ConvNeXt-Tiny and ViT-B/16 outperformed ResNet50 by a meaningful margin while all five architectures were still working from the same unmodified radiograph, how much of the remaining error was attributable to the architectures themselves, and how much was simply because none of them were ever given a genuinely clear image to work from in the first place? Pelvic radiographs are not an easy input to begin with. Beyond the anatomical subtlety of many fractures, the raw images themselves frequently suffer from low contrast, uneven illumination, and noise, all of which can obscure a fracture line that might otherwise be clearly visible. Poor visibility of bony landmarks and overlapping soft tissue can degrade a classifier's performance regardless of how sophisticated its architecture is. If a fracture line is barely distinguishable from surrounding tissue in the pixel values themselves, no amount of architectural sophistication can fully compensate for information that was never clearly present in the input.

The clinical stakes of this question are not abstract. Missed diagnoses in trauma settings are already common: up to 22% of pelvic fractures have been reported as missed during initial assessment, and 51% of patients have been found to be underdiagnosed by pelvic X-ray alone [19, 18]. Comparative work has further shown that plain pelvic radiographs correctly identify only around 48% of true positive fractures relative to CT, which remains the more reliable but far less accessible diagnostic modality in emergency settings [72, 73]. Given this gap between what a radiograph theoretically contains and what a clinician or a model can actually extract from it, improving the legibility of the image itself, before any classification takes place, is not a cosmetic step. It sits directly upstream of diagnostic accuracy.

Image enhancement had already shown promise in adjacent radiographic domains. In chest X-ray classification, adding preprocessing enhancements such as histogram equalization, combined with blurring and masking, raised a baseline CNN's accuracy from approximately 93% to above 97% on a normal versus pneumonia versus COVID classification task [74]. In neonatal chest radiographs, a Bayesian-optimized variant of CLAHE produced consistently higher accuracy and F1-scores than unprocessed images across several CNN architectures [75]. These results, however, came from chest imaging, not pelvic imaging, and to the authors' knowledge, no systematic study had yet applied a comparable preprocessing benchmark specifically to pelvic fracture detection. This was the gap this second study addressed directly, continuing from where the Architecture Benchmark study left off: rather than varying architecture, this study fixed the architecture and varied the preprocessing applied to the input instead [76].

The architecture was fixed to ResNet50, chosen for its proven stability and strong baseline performance in medical imaging rather than for being the top performer in the previous study, precisely so that any performance change observed could be attributed to preprocessing alone rather than to an interaction between architecture and preprocessing. Five preprocessing techniques were applied independently to the PXR150 dataset to produce five enhanced dataset variants, alongside the untouched original: Histogram Equalization (HE), which redistributes pixel intensities globally across the full dynamic range; Contrast-Limited Adaptive Histogram Equalization (CLAHE), which performs the same redistribution locally within small image tiles, limiting how much contrast can be amplified in any one region to avoid over-amplifying noise; Gamma Correction, a nonlinear

Figure 3.2: Visual comparison of the five preprocessing techniques applied to a representative pelvic X-ray, alongside the original image.

power-law transform that brightens or darkens midtones depending on the chosen exponent; Balance-Contrast Enhancement Technique (BCET), a parametric transformation that adjusts contrast while preserving the image's overall average brightness; and Image Complement, which simply inverts pixel intensities so that bright regions become dark and vice versa. Each of the six resulting dataset variants, the original plus five enhanced versions, was trained and evaluated independently using the same five-fold cross-validation protocol, allowing direct, isolated comparison of each technique's effect.

Table 3.2: Classification performance of ResNet50 on the original and five preprocessed variants of PXR150.

Variant	Accuracy	Optimal Specificity	Optimal Recall	AUROC
Original	0.7733	0.7500	0.7800	0.7790
HE	0.7800	0.7600	0.7800	0.7812
CLAHE	**0.8067**	0.8000	**0.8200**	0.8140
Gamma	**0.8067**	**0.8100**	0.8100	**0.8160**
Complement	0.7667	0.7400	0.7700	0.7690
BCET	0.8000	0.7900	0.8100	0.8090

CLAHE and Gamma Correction produced the strongest and most consistent improvements, each reaching 80.67% accuracy against the original's 77.33%, with AUROC improving from 0.7790 to 0.8140 and 0.8160 respectively. BCET also improved on the baseline, reaching 80.00% accuracy and 0.8090 AUROC. HE, despite being a well-established general-purpose contrast enhancement technique, produced only a marginal improvement, 78.00% accuracy and 0.7812 AUROC, only slightly above the original. Image Complement was the sole technique that underperformed the unmodified baseline, dropping to 76.67% accuracy and 0.7690 AUROC.

The pattern across these six results is not random, and it points to a specific mechanism rather than a general claim that "preprocessing helps." The two strongest techniques, CLAHE and Gamma Correction, share a common property: both adjust contrast locally or nonlinearly rather than applying a single global transformation uniformly across the entire image. CLAHE's tile-based local equalization is specifically designed to accentuate faint, localized detail, such as a subtle fracture line, without correspondingly amplifying noise elsewhere in the image. Gamma Correction's nonlinear midtone mapping serves a related purpose, making subtle structural differences in the mid-brightness range more visible without distorting the extremes of the intensity distribution. HE, by contrast, redistributes intensity globally across the entire image, and this global, one-size-fits-all adjustment appears too coarse to meaningfully help distinguish a localized fracture line from its immediate surroundings, which is consistent with its comparatively weak result. Image Complement's underperformance has an even more direct explanation: by inverting bright and dark regions, it renders bone unnaturally dark and soft tissue unnaturally bright, precisely the opposite of how a radiograph normally presents that anatomy, which plausibly disrupts rather than reinforces the visual cues the network had implicitly learned to associate with bone and fracture.

This finding directly answers the question that motivated the study: preprocessing does matter, but not uniformly, and the specific type of preprocessing matters more than the mere presence of preprocessing itself. Local, nonlinear contrast adjustment measurably improves fracture classification performance using a fixed architecture, closing roughly the same margin of AUROC, around 0.035 to 0.04, that separated ResNet50 from ConvNeXt-Tiny in the previous study, without any change to the network itself. This is a meaningful result on its own terms. It suggests that the field's tendency to focus primarily on architecture search, exemplified by the previous study in this chapter, may be leaving comparable performance gains on the table simply by not attending closely enough to how the input is prepared.

At the same time, this study, like the one before it, left a specific limitation clearly exposed, again acknowledged directly by its own authors. Every preprocessing technique evaluated here, whether global or local, whether linear or nonlinear, operates on the entire image uniformly. CLAHE enhances contrast within local tiles, but it has no concept of anatomy: it does not know which pixels belong to bone and which belong to background soft tissue, and it enhances both indiscriminately according to purely local pixel statistics. The improvement these techniques offer is therefore a general one, an image that is easier to see clearly in a generic sense, rather than an anatomically informed one, an image where the network's attention is actively directed toward the bone structures where fractures actually occur. This distinction, between making an image locally clearer everywhere versus making a network attend specifically to anatomically relevant regions, is precisely what the next study in this line set out to address, by asking whether restricting the network's input to the anatomical region of interest itself, rather than merely enhancing contrast across the whole image, could produce a more targeted improvement.

3.1.3 Patch Ensemble

The Preprocessing Matters study had closed off one direction of inquiry and, in doing so, opened another. Local, nonlinear contrast enhancement had been shown to help, but

it helped everywhere in the image indiscriminately, with no concept of anatomy at all. CLAHE does not know that a pelvis has a left half and a right half, or that fracture patterns are frequently asymmetric between the two sides. This observation motivated a different kind of intervention entirely, one that did not touch pixel intensities at all, but instead changed how the image was spatially divided before being fed into the network.

Pelvic fractures are not rare. They account for an estimated 3 to 8% of all fractures and affect roughly 20 per 100,000 individuals annually [77, 78], and in-hospital mortality has been reported in the range of 5 to 20%, driven primarily by exsanguinating hemorrhage in younger patients and multi-organ failure in older ones [12]. Despite this clinical burden, pelvic X-rays used as the initial diagnostic tool have been reported to have a detection rate of only around 78% in acute trauma patients [79, 80], and interpretation variability among readers, combined with growing radiologist workload, has been linked to diagnostic delays severe enough that some fractures reportedly remain undetected for up to two months.

By this point in the research line, deep learning's promise for pelvic fracture detection was already well established by the two preceding studies, and by related work such as Rahman et al.'s two-step transfer learning approach, which had achieved AUROCs of 0.9327 and 0.8014 for visible and invisible fractures respectively using synthesized 3D-CT-derived images [55]. Meta-learning approaches had also shown promise for femur fracture classification specifically [81]. Separately, ensemble learning, combining predictions from multiple models rather than relying on a single one, had repeatedly improved diagnostic accuracy across other areas of medical imaging: in liver mass classification from ultrasound, ensembles of 16 different CNNs consistently outperformed any individual model [82]; in diagnosing lumbosacral pathology, a stacking ensemble with feature selection reached 93.0% accuracy, exceeding transfer learning baselines [83]; and in hepatocellular carcinoma detection from serological indices, an ensemble of gradient-boosted models reached 96.62% accuracy [84]. These results, across genuinely different imaging modalities and clinical tasks, suggested that ensembling was a broadly reliable way to improve diagnostic performance, not one narrowly tied to a specific domain.

However, the conventional way of building an ensemble, training several different architectures on the same whole image and combining their predictions, has a specific weakness for pelvic fracture detection. Standard CNNs and standard ensembles both extract global features from the image as a whole. A model trained this way learns to summarize the entire pelvis into a single feature representation, and in doing so, it can dilute or wash out the kind of small, localized, and often asymmetric detail that a subtle fracture actually presents. Two conventional ensemble members disagreeing about global image statistics does not necessarily help if both are, in a sense, looking at the same blurred, whole-pelvis picture. This is the specific weakness Patch Ensemble was designed to address, and it represents a genuinely different strategy from both preceding studies in this chapter: rather than changing the architecture, and rather than changing the pixel values, this study changed how the image itself was spatially partitioned before any feature extraction took place [85].

Patch Ensemble partitions each pelvic X-ray vertically along its midline, producing two non-overlapping halves, left and right, each capturing one lateral side of the pelvis. Each half is passed independently through its own EfficientNetB0 backbone [45], producing two separate feature representations. These are flattened and concatenated into a single joint feature vector, which is then passed through a shared classification head, two dense layers

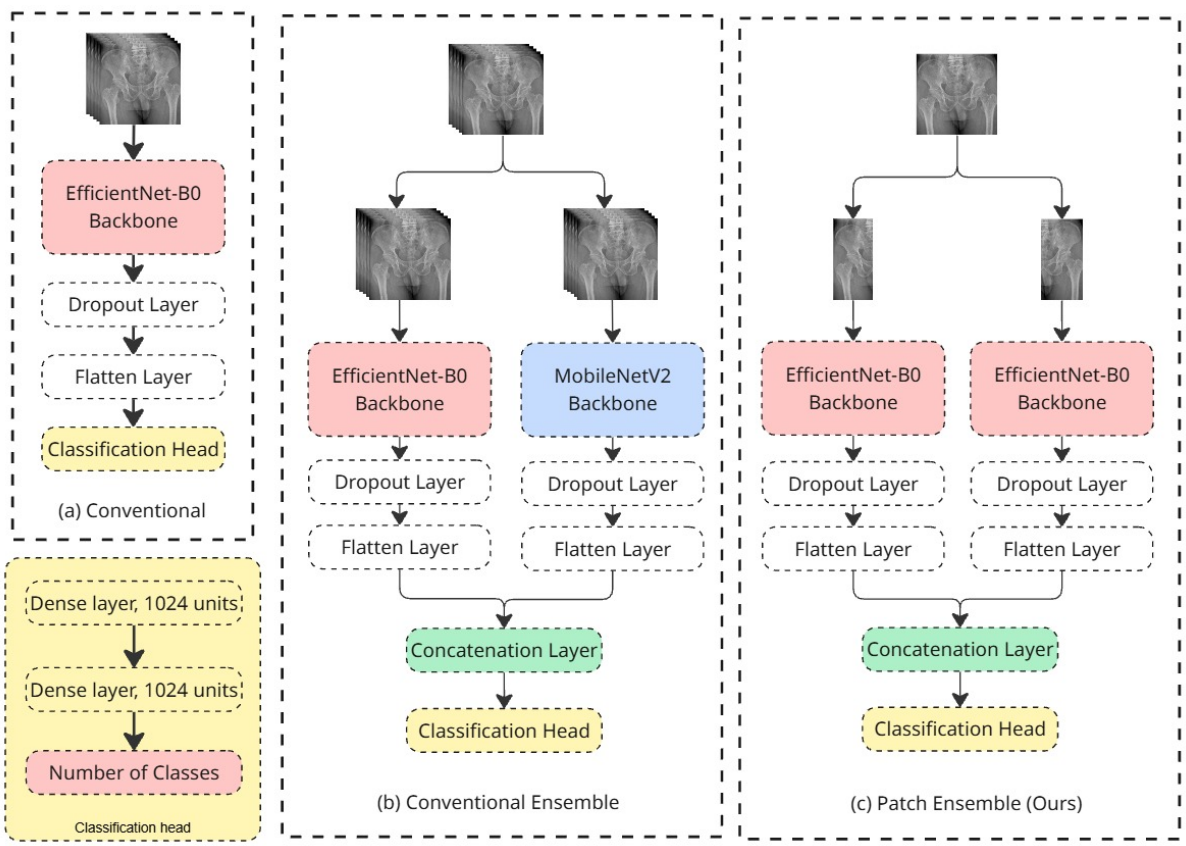

Figure 3.3: Comparison of the three evaluated methods of the Patch Ensemble Study.

of 1024 units each with ReLU activation, followed by a sigmoid output, to produce the final binary fracture prediction. Unlike a conventional ensemble, which achieves diversity through architectural variety, using different backbone types to extract different kinds of features from the same whole image, Patch Ensemble achieves diversity through spatial specialization: both branches share the identical architecture, but each is exposed only to one lateral half of the pelvis, forcing it to specialize in the localized, potentially asymmetric patterns present in that half specifically.

This design was evaluated against two baselines on PXR150, under the same five-fold cross-validation protocol used in the two preceding studies. The Conventional CNN baseline used a single EfficientNetB0 processing the whole, undivided image. The Conventional Ensemble baseline used two different backbones, EfficientNetB0 and MobileNetV2, both processing the same whole image independently, with their features concatenated before classification, representing the standard architectural-diversity approach to ensembling.

Table 3.3: Comparison of Conventional CNN, Conventional Ensemble, and Patch Ensemble on PXR150.

Method	Accuracy	Optimal Recall	Optimal Specificity	AUROC
Conventional CNN	0.77	0.81	0.74	0.77
Conventional Ensemble	0.80	0.82	0.76	0.79
Patch Ensemble (Ours)	**0.84**	**0.87**	**0.82**	**0.87**

Patch Ensemble outperformed both baselines across every metric, reaching an accuracy of 0.84, recall of 0.87, specificity of 0.82, and an average AUROC of 0.87, compared to

0.77 AUROC for the Conventional CNN and 0.79 for the Conventional Ensemble. The improvement in recall is particularly notable from a clinical standpoint: missing a fracture carries considerably more downstream risk than a false alarm, and Patch Ensemble's recall of 0.87 came without the specificity trade-off that both baselines exhibited, where a modest recall gain from Conventional CNN to Conventional Ensemble was accompanied by only a marginal specificity gain, 0.74 to 0.76. Patch Ensemble, by contrast, achieved its higher recall alongside the highest specificity of the three methods, 0.82, suggesting the spatial partitioning strategy was not simply trading one error type for another but genuinely improving discrimination in both directions.

The improvement from Conventional CNN to Conventional Ensemble, 0.77 to 0.79 AUROC, is consistent with ensembling's established reputation for producing modest, reliable gains through architectural diversity alone, as seen across the other medical imaging domains discussed earlier in this section. But the considerably larger jump from Conventional Ensemble to Patch Ensemble, 0.79 to 0.87, indicates that spatial, anatomically-motivated partitioning contributed far more than architectural diversity did on its own. This is a meaningful finding in the context of this chapter's progression: it is the first result in this research line to demonstrate that structuring the input around anatomical geometry, in however simple a form, produces a larger performance gain than either changing the backbone architecture or enhancing pixel-level contrast had achieved individually in the two preceding studies.

At the same time, Patch Ensemble's left-right split is a coarse and somewhat blunt form of anatomical structuring. It divides the image along a fixed vertical midline, which approximates bilateral symmetry but has no actual knowledge of where individual bones are located, and it cannot adapt to patient positioning variation or pelvic rotation that might shift true anatomical boundaries away from the image's geometric center. The study's own authors identified a further, quite specific limitation: because the two patches are processed and fused before classification, and because the underlying image is split rather than kept whole, Patch Ensemble is not compatible with Grad-CAM [86], the most common visual explainability method used elsewhere in this research line, since a Grad-CAM heatmap computed on a half-image patch cannot be straightforwardly reassembled into a coherent explanation over the original, undivided radiograph. In a clinical context, where a radiologist may want to see which region of the image drove a given prediction, this loss of interpretability is a real cost, not a minor implementation detail.

Both of these limitations, the coarseness of a fixed geometric split and the loss of interpretability that comes with it, point toward the same underlying need: a way of dividing the image that is anatomically precise rather than merely geometric, and that preserves a coherent, interpretable link back to the original radiograph. That need, for spatial structuring grounded in actual bone anatomy rather than a fixed midline, is exactly what the next study in this line set out to provide.

3.1.4 Beyond the Pelvic Ring

Patch Ensemble had shown that structuring an image around anatomical geometry helps, but its left-right split was a coarse approximation of anatomy, not a precise one, and it came at the cost of interpretability. What was needed was a form of anatomical structuring

that was both precise, grounded in the actual location of individual bones rather than a fixed midline, and interpretable, preserving a coherent link back to the original image that a technique like Grad-CAM could still meaningfully explain. Segmentation-guided classification offered a natural route to this, and by this point it had already been used successfully elsewhere in medical imaging: to localize kidney tumors in CT scans [87, 88], to improve breast cancer classification in histopathology [89], to support multi-class CXR classification [1], and to detect diabetic eye disease from retinal imaging [90]. Deep learning-based segmentation itself had also matured considerably as a general-purpose tool. nnU-Net offered a self-configuring framework for 2D and 3D segmentation that achieved strong performance across many tasks with minimal manual tuning [91], and TotalSegmentator extended this further, segmenting 104 distinct anatomical structures on whole-body CT with a Dice score of 0.943 [92]. Transformer-based segmentation architectures such as TransUNet [93] and foundation models such as MedSAM, trained on over 1.5 million image-mask pairs across ten modalities [94], had further pushed segmentation quality and generalizability.

Within pelvic imaging specifically, segmentation-guided approaches had already been explored, but with a specific and consistent limitation. Lee et al. developed an automatic pelvic segmentation method achieving Dice coefficients of 96.32% for pelvic segmentation, and although this study did not perform fracture classification directly, its analysis of prediction results on fractured images demonstrated the plausibility of using segmentation to guide downstream detection [59]. A subsequent study built directly on this, developing a segmentation-guided classification pipeline for the AO/OTA fracture classification system, reporting segmentation Dice scores of 0.96 to 0.97 and downstream classification accuracy between 0.69 and 0.88 across fracture types [60]. Both of these studies, however, shared the same underlying restriction: they treated the pelvis as a single region, or restricted analysis to the pelvic ring specifically, and neither considered the broader pelvic anatomy, including the proximal femurs, nor compared their segmentation-guided results directly against raw, unsegmented X-rays. This left it genuinely unclear how much of the improvement, if any, was actually attributable to segmentation itself, as opposed to other differences in how each pipeline was built.

This was the specific gap the Beyond the Pelvic Ring study addressed [95]. Rather than treating the pelvis as a single undifferentiated region, as every prior segmentation-guided study in this space had done, this study proposed a Multi-Bone Segmentation Framework that isolates nine individual pelvic bones, the left and right ilium, ischium, and pubis, the left and right proximal femurs, and the sacrum, as nine separate segmentation classes, before reconstructing them into a single anatomically complete representation of the pelvis.

The segmentation module uses a transformer-enhanced U-Net, combining a standard U-Net decoder [96] with a lightweight Mix Transformer (MiT-B0) encoder [97], pretrained on ImageNet [25], chosen specifically to balance segmentation accuracy against model complexity rather than relying on a larger, self-configuring framework such as nnU-Net. All nine bones were manually annotated on both the public PXR150 dataset and the private AMERI PXR dataset using a double-blind protocol, in which two independent annotators labeled each image separately, with the final ground truth obtained by taking the logical intersection of both annotations, reducing individual annotator bias. The resulting segmentation masks feed into a classification module built on an EfficientNetB3 backbone [45].

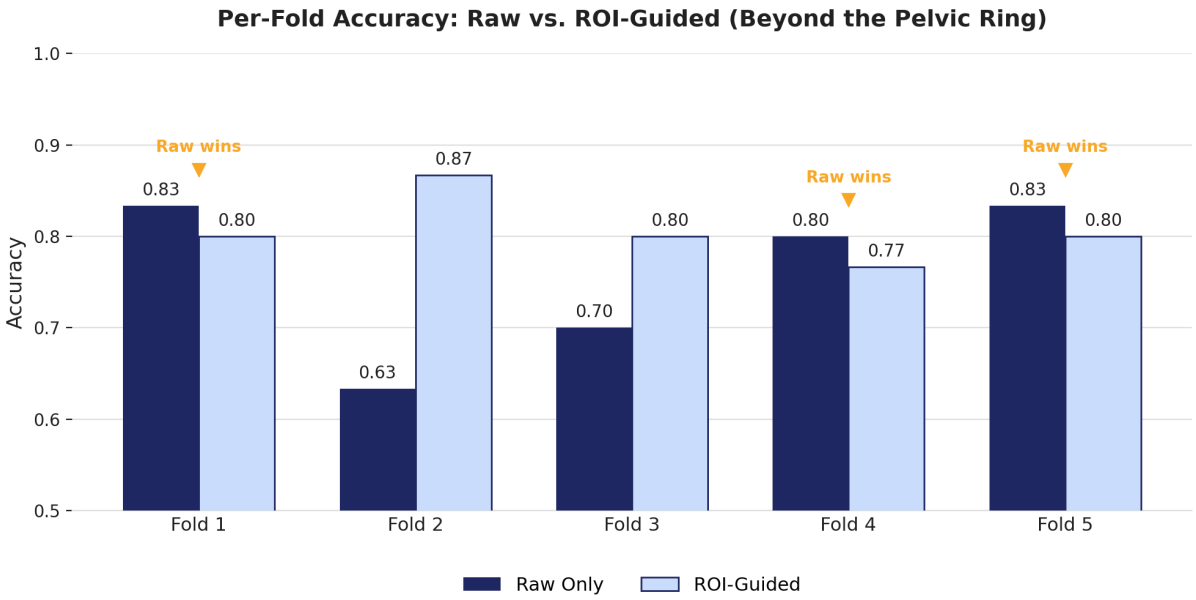

Figure 3.4: Per-fold accuracy comparison between the Raw and ROI-Guided method.

To isolate exactly how much benefit segmentation itself contributes, and at what level of anatomical granularity, this study benchmarked the proposed multi-bone method against three alternatives, all sharing an identical classification module so that only the input representation varied: a Conventional Method using the raw, unsegmented radiograph; a Conventional ROI-Guided Method using a single binary mask that isolates the whole pelvis as one undifferentiated foreground region, without distinguishing individual bones; and a Reference Segmentation Method using manually annotated masks rather than model-predicted ones, established specifically to serve as a performance upper bound that isolates the true value of anatomically precise input from any errors introduced by the segmentation model itself.

Table 3.4: Classification performance across four input representations on PXR150 and AMERI PXR (95% confidence intervals in parentheses).

Method (PXR150)	Accuracy	Opt. Recall	Opt. Specificity	AUROC
Conventional	76.00%	79.00%	76.00%	0.742
Conventional ROI-Guided	80.70%	82.00%	78.00%	0.802
Proposed Multi-Bone	**81.30%**	81.00%	**80.00%**	**0.822**
Reference Segmentation	82.70%	85.00%	80.00%	0.824
Method (AMERI PXR)	Accuracy	Opt. Recall	Opt. Specificity	AUROC
Conventional	79.50%	80.30%	76.70%	0.820
Conventional ROI-Guided	81.20%	**82.10%**	78.30%	0.836
Proposed Multi-Bone	**83.20%**	80.80%	**80.00%**	**0.838**
Reference Segmentation	85.40%	87.30%	85.00%	0.920

On PXR150, the Proposed Multi-Bone Method reached 81.30% accuracy and an AUROC of 0.822, ahead of the Conventional ROI-Guided Method's 80.70% and 0.802, and well ahead of the raw Conventional baseline's 76.00% and 0.742. The same ordering held on the external AMERI PXR dataset: 83.20% accuracy and 0.838 AUROC for the multi-bone method, against 81.20% and 0.836 for the ROI-guided method, and 79.50% and 0.820

for the raw baseline. In both cases, the multi-bone method also came within a narrow margin of the Reference Segmentation upper bound, particularly on AUROC, where the gap was only 0.002 on PXR150 (0.822 versus 0.824) and 0.082 on AMERI PXR (0.838 versus 0.920), indicating that automated multi-bone segmentation was capturing most of the achievable benefit of anatomically precise input without requiring labor-intensive manual annotation.

The five-fold-averaged comparison in Table 3.4 shows ROI-Guided ahead of the raw baseline on PXR150, 80.70% against 76.00% accuracy. The per-fold breakdown behind that average, shown in Figure 3.4, tells a less tidy story. Raw actually outperforms ROI-Guided on three of the five individual folds, Fold 1 (0.83 versus 0.80), Fold 4 (0.80 versus 0.77), and Fold 5 (0.83 versus 0.80), and it is only Fold 2, where ROI-Guided reaches 0.87 against Raw's comparatively weak 0.63, that pulls the five-fold average in ROI-Guided's favor. Fold 3 also favors ROI-Guided, 0.80 against 0.70, but by a narrower margin. In other words, the headline result that segmentation-guided input outperforms raw input on average is being driven disproportionately by a single fold, Fold 2, where the raw baseline happened to perform unusually poorly, rather than by a consistent, fold-by-fold advantage for ROI-Guided across the board. This does not overturn the study's conclusion, since the proposed multi-bone method outperformed both the raw baseline and the ROI-guided method on the five-fold average, which remains the appropriate way to summarize overall performance. But it is a useful caution against reading too much into any single averaged comparison from a dataset as small as PXR150's 150 images split five ways, and it reinforces a point already raised in Chapter 4 regarding the AMERI dataset: with cross-validation folds this small, a single fold's idiosyncrasies can meaningfully shift an averaged result, and fold-level variance deserves at least as much attention as the average itself.

This result is worth sitting with for a moment, because it directly and cleanly confirms the reasoning that motivated the study: distinguishing individual bones rather than treating the pelvis as one undifferentiated region does yield a real, measurable classification improvement, on both an internal dataset and an external one the model was never trained on. It also, notably, resolves the ambiguity left by the two prior segmentation-guided pelvic studies discussed earlier, since this study is the first to directly compare segmentation-guided classification against a matched raw-image baseline using an identical classification module, isolating segmentation's actual contribution rather than leaving it entangled with other pipeline differences.

At the same time, the results contain a specific and important disparity that complicates a simple "finer segmentation is always better" reading, and it appears specifically in the segmentation quality metrics rather than the downstream classification metrics. On PXR150, the proposed multi-bone segmentation outperformed the conventional ROI-guided segmentation on both measures of raw segmentation quality, reaching an IoU of 0.9498 and a Dice score of 0.9741, against 0.9252 and 0.9608 for the coarser ROI-guided mask. But this ordering reversed on the external AMERI PXR dataset, where the segmentation models, trained only on PXR150, were evaluated without any retraining. There, the simpler, coarser ROI-guided segmentation generalized better, reaching an IoU of 0.9027 and a Dice score of 0.9483, while the multi-bone segmentation's quality dropped more sharply to an IoU of 0.8504 and a Dice score of 0.9183. The more anatomically detailed nine-class segmentation task, which requires the model to correctly distinguish nine adjacent, often overlapping bone boundaries rather than a single coarse foreground region, appears to be inherently more sensitive to the kind of distribution shift that occurs when moving from

one dataset's imaging protocol and patient population to another's. A coarser, binary segmentation task has fewer boundaries to get wrong and is correspondingly more forgiving of exactly this kind of domain shift.

What makes this finding genuinely interesting, rather than simply a limitation to note in passing, is that this segmentation-quality reversal did not carry through to downstream classification performance. Despite generalizing less well as a segmentation task in isolation, the multi-bone method still outperformed the ROI-guided method on classification accuracy and AUROC on AMERI PXR, exactly as it had on PXR150. This suggests that the classification module was able to extract sufficient anatomical benefit from the multi-bone segmentation even when that segmentation was measurably less precise on unseen data, and that the relationship between segmentation quality and downstream diagnostic utility is not as tightly coupled as a simple upstream-quality-determines-downstream-performance assumption would predict. It is a nuance worth carrying forward: a segmentation method's value for a diagnostic pipeline cannot be judged purely on its own segmentation metrics, but must be evaluated through its actual effect on the downstream task it is meant to support.

The more fundamental limitation this study leaves exposed, however, is a different one, and one its own authors identify directly. However precisely the nine bones are segmented, the classification module in every one of the four benchmarked methods here still receives only the segmented, cropped bone regions as its final input, never the raw image alongside it. The improvement over the raw baseline demonstrates that anatomical focus helps. But it says nothing about whether the non-local, non-bone information that a raw radiograph preserves, soft tissue alignment, joint spacing, overall pelvic symmetry, might be adding something segmentation-guided classification is, by construction, unable to see. That open question, whether combining anatomical focus with the retained global context that segmentation necessarily discards could push performance further still, is precisely what the next study in this line was designed to test.

3.1.5 Fusion Ensemble

Beyond the Pelvic Ring had answered one question cleanly: distinguishing individual bones through segmentation improves fracture classification over a raw baseline, even under external validation. But it left a different, more fundamental question open. Every method benchmarked in that study, including the proposed multi-bone approach, still received only the segmented, cropped bone regions as its final classification input. None of them ever saw the raw radiograph alongside the segmentation. This was not an oversight so much as an unexamined assumption running through the entire segmentation-guided literature: that isolating anatomically relevant regions and discarding everything else is a strictly beneficial simplification, never a loss.

Pelvic fracture diagnosis in clinical practice does not always work this way. While some fractures present as unambiguous cortical disruptions, precisely the kind of local, bone-boundary signal segmentation is well suited to isolate, others manifest through more indirect, non-local cues: abnormal pelvic alignment, altered joint spacing, or asymmetric limb positioning [52]. These cues are not confined to the bone itself. Limb rotation, joint dislocation, and pubic symphysis widening can all suggest a fracture even where

no visible cortical break is present. A segmentation pipeline that crops the image down to bone regions alone necessarily discards exactly this kind of surrounding contextual information, since by construction it retains only what falls inside the segmented mask. Raw pelvic radiographs, by contrast, preserve this full anatomical context, including soft tissue and overall alignment, but at the cost of also retaining background noise and irrelevant structures that segmentation is specifically designed to filter out. Framed this way, segmentation and raw imaging were not competing options where one simply outperforms the other, they were each giving up something the other retained.

By this point, the broader value of deep learning across varied diagnostic tasks was well established within this same research group's own work, spanning CXR classification [98], fundus-based ocular disease detection [99], tuberculosis detection in histopathology [100], and lumbar spine degeneration classification from MRI [101], alongside the broader literature on deep learning in medical image analysis generally [24] and the value of transfer learning specifically in data-limited medical settings [102]. Fusion, combining information from multiple sources within a single pipeline, was an established strategy for reconciling exactly this kind of complementary-but-incomplete information, with feature-level fusion generally outperforming pixel-level or decision-level fusion in prior surveys [61], and generative fusion approaches such as MedFusionGAN [62] and multi-scale attention approaches such as MMIF-Net [63] demonstrating fusion's value across other cross-modal medical imaging settings. But to the authors' knowledge, no prior work had systematically compared pixel-level and feature-level fusion strategies specifically for combining raw and segmented representations of the same pelvic radiograph. This was the gap the Fusion Ensemble study addressed, picking up directly from where Beyond the Pelvic Ring left off: rather than choosing between raw and segmented input, this study asked whether combining both, through a properly designed fusion mechanism, could outperform either one used alone.

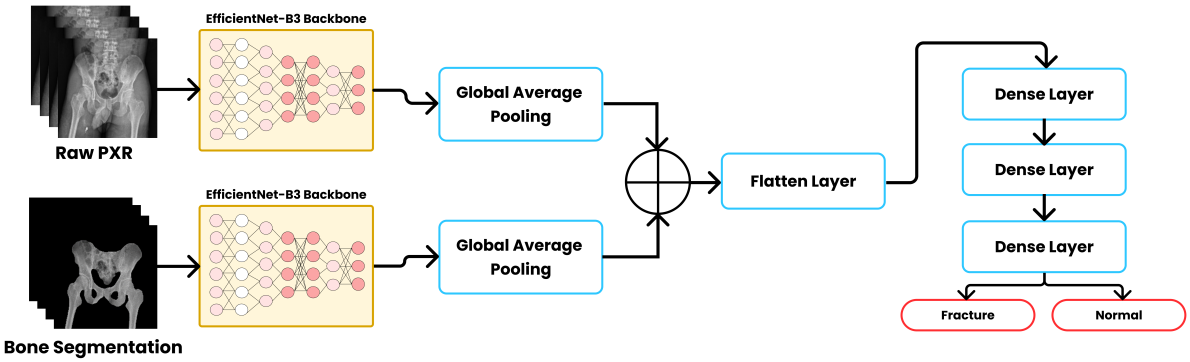

Figure 3.5: Overview of the Fusion Ensemble framework.

The proposed Fusion Ensemble uses a dual-branch architecture in which the raw pelvic X-ray and its corresponding bone-segmented counterpart, produced by a single-class segmentation model isolating the pelvis as one region rather than nine individual bones, are each processed through separate EfficientNetB3 backbones. The high-level feature representations from both branches are aggregated using global average pooling and then concatenated into a single unified feature vector, before being passed through two fully connected layers and a sigmoid output for final binary classification. This is a feature-level fusion strategy specifically: rather than combining raw pixel values before feature extraction, or combining two independent predictions after classification, the two branches are allowed to learn modality-specific representations first, and only then are those

representations combined, preserving the distinct characteristics each input contributes while still enabling the two to interact before the final decision.

To establish exactly where feature-level fusion sits relative to simpler alternatives, this study benchmarked it against four other approaches, all using an identical EfficientNetB3-based classification pipeline so that only the input representation varied: a Conventional Method using the raw image alone; a Segmentation-Guided Method using only the segmented image; an Average Fusion approach, combining raw and segmented images through simple pixel-wise averaging into a single grayscale image; and an RGB Fusion approach, encoding the raw image in one color channel and the segmented image duplicated across the remaining two, allowing both to be processed by a standard three-channel convolutional backbone without pixel-level blending.

Table 3.5: Classification performance across five input/fusion strategies on PXR150 (95% confidence intervals in parentheses).

Method	Accuracy	Opt. Spec.	Opt. Recall	Opt. F1	AUROC
Raw PXR	78.00%	78.00%	74.00%	76.00%	0.7850
Segmented PXR	80.00%	79.00%	76.00%	77.50%	0.8050
Average Fusion	81.33%	77.80%	**80.00%**	78.90%	0.8140
RGB Fusion	82.00%	80.00%	78.00%	79.00%	0.8204
Fusion Ensemble	**84.00%**	**84.00%**	**80.00%**	**82.00%**	**0.8340**

The ordering across these five methods is consistent and, importantly, monotonic: performance improves step by step as the input representation moves from raw-only, to segmented-only, to increasingly sophisticated fusion. The raw baseline reached 78.00% accuracy and 0.7850 AUROC; the segmented-only baseline improved on this, 80.00% accuracy and 0.8050 AUROC, confirming the same anatomical-focus benefit already established in the preceding study. Both pixel-level fusion strategies improved further still, Average Fusion reaching 81.33% accuracy and RGB Fusion reaching 82.00% accuracy and 0.8204 AUROC, showing that even a fairly naive combination of raw and segmented pixel values captures some complementary benefit neither input provides alone. The proposed feature-level Fusion Ensemble outperformed all of these, reaching 84.00% accuracy, 84.00% specificity, and an AUROC of 0.8340, the best result across every metric reported.

This progression directly validates the reasoning that motivated the study: raw and segmented representations are not substitutes for each other but complements, and the more effectively their information is combined, the better the resulting classification. It is worth noting specifically why feature-level fusion outperformed the two pixel-level approaches, since the gap is informative rather than incidental. Average Fusion blends raw and segmented pixel values directly, which by construction has no way to account for the fact that different regions of the two images carry different relative importance; a fracture-relevant pixel in the segmented image and a background pixel in the raw image are averaged with equal weight regardless of their diagnostic relevance. RGB Fusion avoids blending pixel values directly by keeping the two inputs in separate color channels, but the network is still required to implicitly learn how to weigh and interpret the two channels' relative contributions during feature extraction, with no explicit mechanism guiding that process. Feature-level fusion sidesteps both limitations by extracting a full, independently learned representation from each input before any combination takes place, allowing each branch to develop a representation suited specifically to its own input, raw or segmented,

before the two are brought together.

This result closes a gap Beyond the Pelvic Ring had left open, and it does so in the direction that study's own limitation implied it should: combining anatomical focus with retained global context outperforms either one in isolation. But the way this study achieves that combination is, on reflection, a fairly simple one. The fusion mechanism here amounts to concatenating two independently pooled feature vectors, with no explicit interaction between the raw and segmented representations during feature extraction itself, and no mechanism for the network to select or weight which of the two inputs is more informative for a given case. Every input, easy or difficult, is processed through the full dual-branch pipeline identically, and the segmentation used here treats the pelvis as a single region rather than distinguishing individual bones, reverting to the coarser segmentation granularity that Beyond the Pelvic Ring had already shown to underperform the nine-bone approach. Two open threads remain, then, heading into the next study: whether a more interactive form of fusion, one where the two streams can exchange and refine information rather than simply being concatenated at the end, could push performance further, and whether combining this kind of fusion with anatomically finer, multi-bone segmentation rather than a single coarse mask could recover the additional benefit Beyond the Pelvic Ring had already demonstrated. Both of these threads are picked up directly in the next and final study in this progression.

3.1.6 PelFANet

The Fusion Ensemble study had closed one gap and left two threads open. The first was architectural: its fusion mechanism combined raw and segmented representations only after independent pooling, through simple concatenation, with no explicit interaction between the two streams during feature extraction, and no mechanism allowing the network to weigh one input more heavily than the other for a given case. The second was anatomical: its segmentation treated the pelvis as a single undifferentiated region, reverting to the coarser masking strategy that Beyond the Pelvic Ring had already shown to underperform a nine-bone segmentation. PelFANet, the final study in this progression [103], was designed to address both threads at once, and, in doing so, to confront a harder and more clinically consequential version of the pelvic fracture detection problem than any of the five preceding studies had directly targeted: fractures that are not visible on the pelvic X-ray at all.

An invisible fracture, in the sense used throughout this research line, is a case confirmed by 3D-CT imaging where the pelvic X-ray itself shows no obvious sign of injury. Every study so far in this chapter, from the Architecture Benchmark through the Fusion Ensemble, had evaluated exclusively on visible fractures, cases where the fracture, however subtle, is at least in principle present in the 2D radiograph a model is asked to interpret. Invisible fractures represent a fundamentally different and harder challenge, since by definition the primary diagnostic signal a model would normally rely on is largely absent from its input. Prior work outside this specific research line had already shown these cases to be genuinely difficult for existing deep learning methods, with Rahman et al.'s two-step transfer learning approach reaching a strong 0.933 AUROC on visible fractures but a markedly lower 0.801 on invisible ones [55], a gap of roughly 0.13 AUROC that stands as the clearest prior quantitative evidence that invisible fractures are not simply a harder version of the same task, but a qualitatively more difficult one.

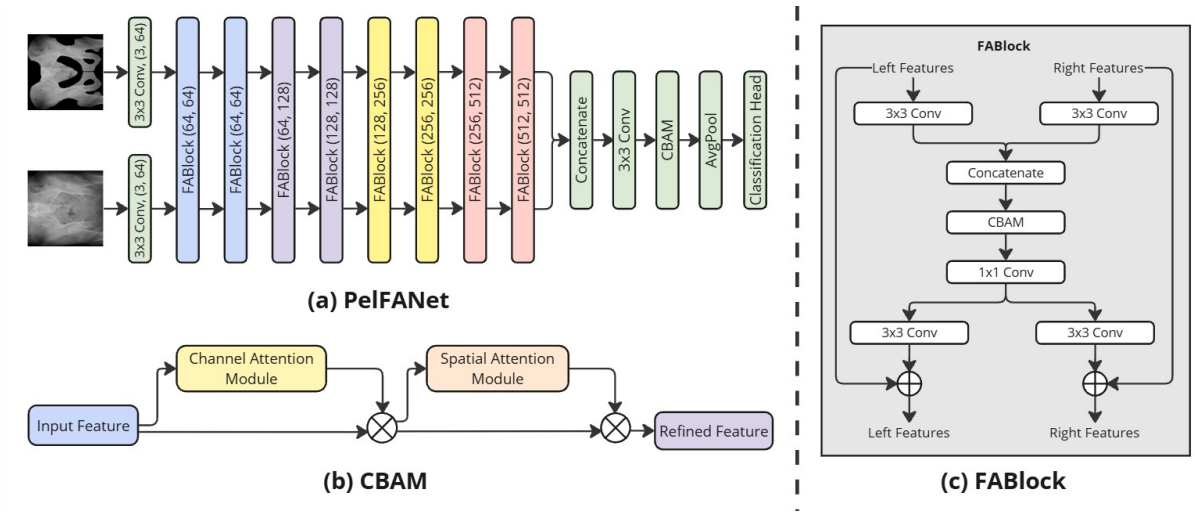

Figure 3.6: The PelFANet architecture.

PelFANet addresses the fusion limitation directly through its Fused Attention Blocks (FABlocks). Rather than extracting independent representations from the raw and segmented streams and combining them only once, at the very end, as the Fusion Ensemble had done, PelFANet interleaves fusion throughout the network. At each of eight stacked FABlocks, feature maps from the raw and segmented branches are independently convolved, concatenated, and passed through a Convolutional Block Attention Module (CBAM) [104], which applies channel and spatial attention sequentially to highlight the most informative parts of the combined representation. This CBAM-refined output is then projected and split back into two streams via residual connections, allowing each branch to absorb attention-refined information from the other before continuing to the next block. This repeated, iterative fusion and redistribution is a qualitatively different mechanism from the single terminal concatenation used in the Fusion Ensemble: rather than the two streams working independently until a single combination point at the end, they exchange and refine information continuously throughout the network, with attention determining, at each stage, which combined features are worth propagating forward.

The segmentation used to produce PelFANet's second input stream follows the same single-class approach as the Fusion Ensemble, using a U-Net with a Mix Transformer B0 encoder [96, 97] to isolate the full pelvic region as one class, rather than the nine-bone segmentation used in Beyond the Pelvic Ring. This means PelFANet resolves the fusion-mechanism thread from the Fusion Ensemble's limitations directly, through FABlocks and iterative CBAM-based attention, but does not resolve the anatomical-granularity thread; segmentation remains coarse, single-class, rather than bone-specific. This is a deliberate scope choice rather than an oversight, since combining fine-grained multi-bone segmentation with iterative attention fusion simultaneously would introduce two architectural changes at once, making it harder to isolate which change was driving any observed improvement.

PelFANet was pretrained on the COVID-QU-Ex chest X-ray dataset [105], a large, paired dataset of chest radiographs and lung masks, chosen specifically because its scale and paired mask structure exposed the network to a wide range of anatomical structures and radiographic variation before fine-tuning on the considerably smaller pelvic dataset, reducing the risk of over-specialization to a limited training set. The model was then fine-tuned and evaluated on the AMERI dataset, using its Visible Fracture (VIS) subset of 228 images for training and evaluation, and, critically, using its separately curated Invisible

Fracture (INVIS) subset of 35 images, confirmed via 3D-CT but showing no visible sign of fracture on the X-ray itself, purely for evaluation. PelFANet was never trained on a single invisible fracture case; its INVIS performance reflects generalization from visible-fracture training alone.

Table 3.6: PelFANet performance on Visible (VIS) and Invisible (INVIS) fracture subsets of AMERI, with fold variance in parentheses.

Subset	Accuracy	Precision	Recall	Specificity	F1	AUC
VIS	88.68%	92.49%	92.21%	78.33%	84.71%	0.9334
	(0.11%)	(0.09%)	(0.17%)	(1.27%)	(0.46%)	(0.10)
INVIS	82.29%	88.36%	84.35%	78.33%	81.23%	0.8688
	(0.01%)	(0.07%)	(0.07%)	(0.44%)	(0.06%)	(0.04)

On the VIS subset, PelFANet reached 88.68% accuracy and an AUC of 0.9334, with a notably high recall of 92.21%, at the cost of a comparatively more modest specificity of 78.33%, indicating a reasonable but non-trivial rate of false positives alongside strong fracture sensitivity. On the INVIS subset, despite never having seen a single invisible fracture case during training, PelFANet reached 82.29% accuracy and an AUC of 0.8688, a result that, while lower than VIS performance as expected given the qualitatively harder nature of the task, still represents strong generalization to a genuinely out-of-distribution evaluation set.

Table 3.7: Comparison with prior methods on AMERI VIS and INVIS subsets.

Method	VIS AUC	VIS F1	INVIS AUC	INVIS F1
ImageNet	0.8961	80.00%	0.7549	72.70%
DRR20	0.9290	85.20%	0.8002	78.60%
ImageNet+DRR20	0.9280	83.90%	0.7140	72.10%
ImageNet+DRR20_Full	0.9151	83.30%	0.6896	77.50%
PelFANet (Ours)	**0.9334**	84.71%	**0.8688**	**81.23%**

This comparison against prior pretraining-based approaches, all using ResNet-based classifiers with various combinations of ImageNet and Digitally Reconstructed Radiograph (DRR) pretraining, is where PelFANet's contribution becomes clearest. On VIS, PelFANet's 0.9334 AUC edges out the previous best, DRR20's 0.9290, by a relatively narrow margin. On INVIS, the margin is far larger: PelFANet's 0.8688 AUC against DRR20's 0.8002, an improvement of roughly 0.069 AUC, and PelFANet's F1 score of 81.23% against DRR20's 78.60%. This pattern, a modest improvement on the easier subset and a substantially larger improvement on the harder one, is the clearest evidence across this entire chapter that architectural improvements targeting anatomical reasoning and cross-stream fusion pay off disproportionately on the hardest cases. It is not simply that PelFANet is a marginally better classifier across the board; it is specifically better at the kind of case where fracture evidence is weakest and most indirect, which is exactly the case where a model most needs to draw on retained global context and refined anatomical attention rather than relying on a locally obvious signal that, for invisible fractures, does not exist.

This VIS-to-INVIS pattern is, in a real sense, the single most important empirical result this chapter has produced, because it previews the central mechanism this research is built around. PelFANet demonstrates that a well-designed architecture can generalize from

visible to invisible fractures, and that the size of the benefit such an architecture provides scales with how difficult the underlying case is. But PelFANet, like every dual-stream architecture reviewed throughout this chapter and in Chapter 2, still applies its full architecture, both streams, all eight FABlocks, identically to every input, whether that input is an unambiguous visible fracture or a genuinely difficult invisible one. There is no mechanism within PelFANet, or within any of the five studies preceding it in this chapter, for the network itself to recognize that a given case is easy and could be resolved with less computation, or that a given case is hard and warrants the full anatomical reasoning capacity the architecture provides. Every case pays the same full computational cost, regardless of whether the visible-fracture-level confidence the primary stream could provide alone would already have sufficed.

This is precisely the limitation SecondOpinion is designed to address. The progression traced across this chapter, from comparing architectures on a raw image, to enhancing that image's contrast, to spatially partitioning it, to anatomically segmenting it, to fusing segmented and raw representations, to fusing them iteratively with attention, has arrived at architectures capable of strong performance even on the hardest, most indirect cases. What none of them provide is a way to invoke that capability selectively, only when a case actually demands it. GateKeeper, introduced in Chapter 5, is designed to supply exactly that missing piece: a learned, correctness-supervised gate that decides, on a per-case basis, whether the kind of anatomy-guided reasoning PelFANet and its predecessors developed is worth invoking at all.

3.2 Prior Approaches to CXR Classification

3.2.1 Architecture Comparison for CXR Classification

The CXR side of this research line opens with the same kind of question that opened the pelvic side: given a fixed dataset and a fixed evaluation protocol, which backbone architecture is actually best suited to multi-disease chest radiograph classification, and does domain-specific pretraining matter as much as architectural choice? At the time this study was conducted, CXR classification work tended to report strong results on narrow, often binary tasks, normal versus COVID-19, or normal versus pneumonia, rather than the more clinically realistic setting where several overlapping conditions must be told apart in the same pass. Differentiating pneumonia from tuberculosis, for instance, is difficult precisely because the two conditions share overlapping radiological features [1], and a model evaluated only on cleaner, two-class splits gives little evidence of how it would behave once that ambiguity is actually present in the label space.

To address this, a unified five-class dataset was constructed [98] by combining two publicly available sources: the COVID-19 Radiography Database, contributing 10,192 normal, 6,012 lung opacity, 1,345 viral pneumonia, and 3,616 COVID-19 images, and the Tuberculosis Chest X-ray Database, contributing 700 tuberculosis-positive images, for a combined total of 21,865 images across five classes. All images were resized to 224×224 pixels, and a five-fold cross-validation protocol was used throughout, with each fold serving as the held-out test set exactly once, so that reported performance reflects an average over the full dataset rather than a single favorable split.

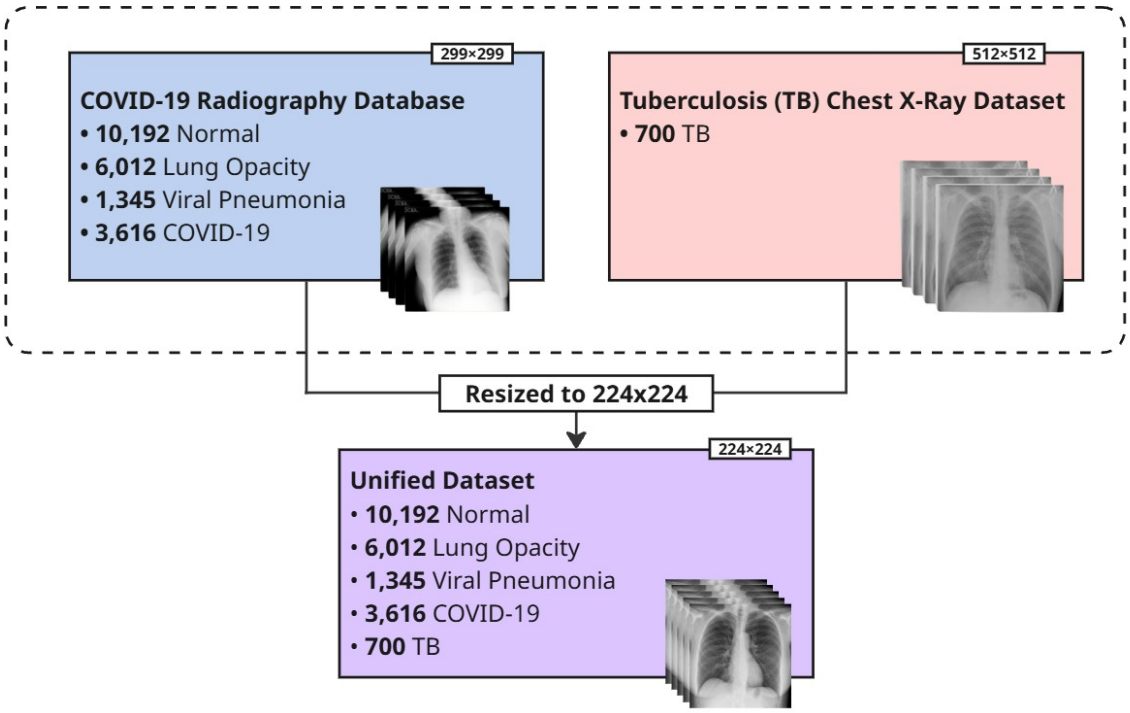

Figure 3.7: Overview of the unified five-class CXR dataset construction.

Six architectures were compared under this shared protocol: ResNet101, representing the classic residual-connection family [42]; DenseNet201 and DenseNet121, representing dense feature reuse at two depths [43]; CheXNet, a DenseNet121 variant distinguished only by additional pretraining on the 112,120-image ChestX-ray14 dataset [1, 41], isolating the effect of domain-specific pretraining while holding backbone architecture constant; MobileNetV2, a lightweight architecture designed for resource-constrained deployment [44]; and EfficientNetB3, representing a principled, compound-scaled convolutional design [45]. All six models were first trained with newly appended fully-connected classification layers while the backbone remained frozen, and then fine-tuned end-to-end for a second phase, isolating how much of each architecture's performance came from generic ImageNet-level features versus features adapted specifically to CXR data.

Table 3.8: Comparative performance of six architectures on the unified five-class CXR dataset, after fine-tuning.

Architecture	Accuracy	Precision	Recall	Specificity	F1 Score	AUROC
MobileNetV2	85.79%	89.08%	82.92%	90.56%	0.859	0.963
DenseNet201	91.05%	94.77%	91.64%	93.67%	0.932	0.982
DenseNet121	91.42%	93.84%	92.05%	92.96%	0.929	0.983
ResNet101	92.75%	95.45%	92.80%	94.92%	0.941	0.983
EfficientNetB3	93.35%	**95.99%**	**93.81%**	95.11%	**0.949**	0.987
CheXNet	**94.12%**	94.76%	93.05%	**95.81%**	0.939	**0.988**

CheXNet reached the highest accuracy, specificity, and AUROC of the six, at 94.12%, 95.81%, and 0.988 respectively, while EfficientNetB3 finished narrowly ahead on precision, recall, and F1 score, at 95.99%, 93.81%, and 0.949. MobileNetV2 finished last on every metric by a wide margin, with an accuracy of 85.79% and an AUROC of 0.963, roughly 0.02

to 0.03 AUROC behind every other architecture evaluated, evidence that its lightweight design, while attractive for deployment, trades away real discriminative capacity on a task with this much class overlap.

The comparison between CheXNet and DenseNet121 is the most informative single contrast in this table, for the same reason the ConvNeXt-Tiny-versus-CheXNet contrast mattered in the pelvic architecture benchmark: the two share an identical backbone, differing only in that CheXNet received additional pretraining on ChestX-ray14 before fine-tuning on the unified dataset. CheXNet's AUROC of 0.988 against DenseNet121's 0.983, and its specificity of 95.81% against DenseNet121's 92.96%, isolate the benefit of domain-specific pretraining from any confound of architectural depth or capacity. Unlike the pelvic architecture benchmark, where a modernized architecture without domain pretraining, ConvNeXt-Tiny, was able to overturn CheXNet's pretraining advantage entirely, no such reversal occurs here: even EfficientNetB3, the architecturally newest and most efficient design in this comparison, does not exceed CheXNet's AUROC. This suggests that for CXR classification specifically, at least under this five-class formulation, the benefit of pretraining on a large, disease-labeled chest radiograph corpus is not easily substituted for by architectural modernization alone, in contrast with what the pelvic domain, where no comparably large, disease-labeled pretraining corpus exists, was found to allow.

This result gave the CXR side of this research line a defensible backbone choice, CheXNet, on the same reasoning basis established for ConvNeXt-Tiny on the pelvic side. But the comparison, exactly like its pelvic counterpart, was conducted entirely at the level of backbone selection. Every one of the six architectures processed the same 224×224 input through a single, unmodified forward pass, with no mechanism for emphasizing diagnostically relevant regions of the image over irrelevant background, and no mechanism for varying computational effort by case difficulty. CheXNet's own error patterns, examined qualitatively through Grad-CAM in the study that produced Table 3.8, showed that even the best-performing backbone did not always attend to disease-relevant regions consistently, misclassified cases in particular showed activation concentrated on anatomically irrelevant areas of the radiograph. This raised a natural next question, structurally identical to the one that motivated the Preprocessing Matters study on the pelvic side: if the backbone itself is fixed, could a mechanism explicitly designed to guide the network's attention toward diagnostically relevant regions recover the performance CheXNet was otherwise leaving on the table? That question is addressed directly by the next study in this progression.

3.2.2 CheXNet-CBAM

The architecture comparison had identified CheXNet as the strongest available backbone for this dataset, and had done so specifically by isolating the benefit of domain-specific pretraining from architectural depth or capacity. But the comparison, like its pelvic counterpart, was conducted entirely at the level of which backbone extracts features best from an unmodified input; it said nothing about whether that backbone was actually attending to the right regions of the image once features had been extracted. Qualitative inspection of CheXNet's misclassified cases had suggested it was not always doing so consistently, with activation sometimes concentrated on anatomically irrelevant portions of the radiograph rather than the disease-relevant lung regions a correct diagnosis would

depend on. This raised the natural next question: if the backbone itself is fixed, could an explicit attention mechanism, one designed specifically to guide the network toward diagnostically relevant regions rather than leaving that association to be learned implicitly through the classification loss alone, recover the performance CheXNet was otherwise leaving on the table?

This question had already been answered affirmatively elsewhere in medical imaging. The Convolutional Block Attention Module (CBAM), which applies channel attention and spatial attention sequentially to refine an intermediate feature map, had proven effective at improving both performance and interpretability across a range of CNN backbones [104], and had already played a structural role earlier in this same research line, refining the fused representation inside PelFANet's FABlocks on the pelvic side of this chapter. What had not yet been established was whether inserting CBAM directly into CheXNet, the strongest backbone identified for CXR classification specifically, on the exact same five-class, 21,865-image dataset used in the architecture comparison, would produce a comparable gain. This was the gap the CheXNet-CBAM study addressed [106], continuing directly from where the architecture comparison left off: rather than varying the backbone again, this study fixed the backbone to the established winner and inserted an explicit attention mechanism into it.

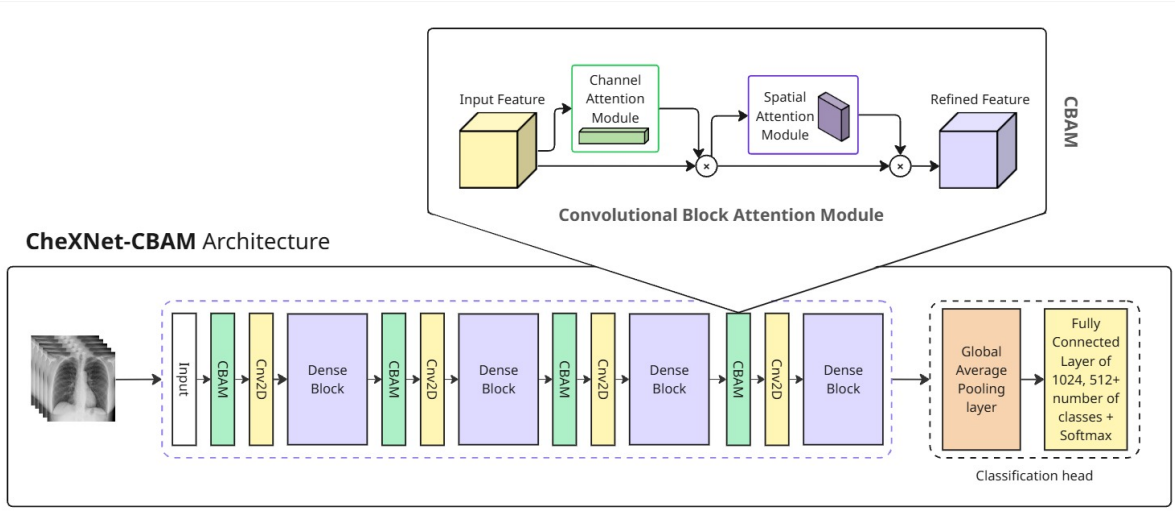

Figure 3.8: The CheXNet-CBAM architecture.

CheXNet-CBAM integrates CBAM modules between the convolutional and dense blocks of the underlying DenseNet121 backbone, rather than appending attention only at a single point in the network. Given an intermediate feature map F, channel attention is applied first, aggregating spatial information through both global average pooling and max pooling and passing the result through a shared multi-layer perceptron to produce channel-wise attention weights, before spatial attention is applied to the channel-refined output, aggregating information across the channel axis and applying a convolution to produce a spatial attention map highlighting the regions that contribute most to the prediction. Because CBAM modules are inserted at multiple points throughout the pipeline, rather than once at the end, attention refinement occurs at multiple hierarchical levels of feature abstraction, from early, texture-level features to late, disease-specific ones. The final feature maps are aggregated through global average pooling and passed through a classification head of two fully connected layers before a softmax output over the five classes. As with CheXNet itself, the model was pretrained on ImageNet before

further pretraining on ChestX-ray14, preserving the domain-specific pretraining advantage the architecture comparison had already confirmed, while adding attention as the one additional variable under test.

Table 3.9: Comparative performance of CheXNet-CBAM against the six architectures from the architecture comparison, on the same unified five-class CXR dataset.

Architecture	Accuracy	Precision	Recall	Specificity	F1 Score	AUROC
MobileNetV2	85.79%	89.08%	82.92%	90.56%	0.859	0.963
DenseNet201	91.05%	94.77%	91.64%	93.67%	0.932	0.982
DenseNet121	91.42%	93.84%	92.05%	92.96%	0.929	0.983
ResNet101	92.75%	95.45%	92.80%	94.92%	0.941	0.983
EfficientNetB3	93.35%	95.99%	93.81%	95.11%	0.949	0.987
CheXNet	94.12%	94.76%	93.05%	95.81%	0.939	0.988
CheXNet-CBAM (Ours)	**96.21%**	**96.14%**	**94.67%**	**97.55%**	**0.954**	**0.9894**

CheXNet-CBAM outperformed every architecture from the preceding comparison on every metric, reaching 96.21% accuracy, 96.14% precision, 94.67% recall, 97.55% specificity, an F1 score of 0.954, and an AUROC of 0.9894. Relative to the plain CheXNet baseline, this represents a 2.09 percentage point improvement in accuracy and a specificity gain from 95.81% to 97.55%, meaningful in a clinical context because higher specificity translates directly into fewer healthy patients flagged for unnecessary follow-up. More notably, CheXNet-CBAM also overtook EfficientNetB3, the architecture that had led the comparison on precision, recall, and F1 score, on all three of those metrics as well, improving over CheXNet specifically by 1.38, 1.62, and 0.015 points respectively. This closes what had been the one gap in CheXNet's otherwise dominant position in the architecture comparison: CheXNet led on accuracy, specificity, and AUROC but trailed EfficientNetB3 on precision, recall, and F1. Attention did not just add a uniform improvement on top of CheXNet's existing strengths; it specifically corrected the metrics where CheXNet had been weakest, consistent with the qualitative observation that motivated this study, that CheXNet's errors were disproportionately traceable to attending to the wrong regions of the image rather than to any deficiency in the underlying feature representation itself.

Grad-CAM visualizations of CheXNet-CBAM's predictions support this interpretation directly. Across the five classes, the model's attention concentrated on class-appropriate regions rather than diffuse or off-target areas: lower-lung regions with dense opacities for COVID-19, subtler mid-lung activation consistent with patchy, non-specific infiltrates for lung opacity, upper-lobe concentration for tuberculosis, the anatomical region where tuberculosis most commonly manifests, and bilateral activation for viral pneumonia, consistent with the diffuse involvement pattern viral infections typically produce. This class-appropriate localization is precisely what plain CheXNet's Grad-CAM maps had lacked, and it is the direct, visually interpretable evidence that CBAM's channel and spatial attention are doing more than simply adding parameters; they are steering the network toward the same kind of region a radiologist would examine first.

This result closes the specific gap this study set out to address: attention, inserted directly into the strongest available backbone rather than into a weaker one, produces a meaningful and metric-consistent improvement over the architecture comparison's best result. But the way CheXNet-CBAM achieves this improvement is, in a structural sense, no different from every architecture that preceded it in this chapter, on either the pelvic or the CXR side. CBAM refines which regions of a given image the network attends to, but it does

so on every input, every time, with no mechanism for recognizing that a large fraction of cases, an unambiguous COVID-19 presentation, a clearly normal chest film, could already be resolved correctly by a faster, less attention-intensive pathway, while reserving CBAM's more expensive, iterative refinement for the harder, more ambiguous cases where it is actually needed. The full CheXNet-CBAM pipeline runs unconditionally on every image in the dataset, exactly as PelFANet's full dual-stream architecture runs unconditionally on every pelvic radiograph.

Taken together, the two arcs traced across this chapter, from architecture selection through preprocessing, spatial partitioning, and anatomical segmentation on the pelvic side, and from architecture selection through attention refinement on the CXR side, converge on the same structural limitation from two independent diagnostic tasks of substantially different difficulty. Both arcs produce architectures capable of strong, often state-of-the-art performance, and both arcs do so by making the model's processing of every single case progressively more sophisticated, never more selective. Each successive study in this chapter, whether adding a segmentation stage, a spatial partitioning scheme, or an attention module, made the same underlying bet: that more anatomical or attentional machinery, applied uniformly, would outperform less. That bet paid off, repeatedly, in the results traced above. But it was never tested against the alternative this research is built around, that the machinery itself could be applied selectively rather than uniformly, and that doing so might recover most of the same performance at a fraction of the cost.

Neither CheXNet-CBAM nor PelFANet, nor any of the studies that led to them, provides a mechanism for the network itself to recognize when that sophistication is unnecessary. Every gain documented in this chapter was purchased by adding a new fixed stage to the pipeline and running it on every case without exception, never by teaching the model to tell, on a case-by-case basis, when a given stage's extra reasoning was worth invoking at all. This is true even where the underlying task made the answer to that question obvious in hindsight: an unambiguous COVID-19 presentation and a genuinely invisible pelvic fracture are, on the evidence of this chapter's own results, very different kinds of cases, yet every architecture surveyed here processes both through the same fixed depth of computation. The progression traced across this chapter is therefore a progression in architectural capability, not in architectural judgment; each study made the model more capable of handling a hard case, but none made it capable of recognizing which cases were hard in the first place.

This is the gap SecondOpinion and its GateKeeper gating mechanism, introduced in Chapter 5, are designed to close, not by proposing a better fixed pipeline for either task individually, but by making the invocation of the kind of anatomy-guided reasoning both arcs of this chapter developed conditional on a learned, correctness-supervised signal of case difficulty. Rather than replacing PelFANet's dual-stream design or CheXNet-CBAM's attention refinement outright, SecondOpinion asks a question neither could answer: whether the additional reasoning each of them always applies could instead be applied only when a case actually calls for it, closing this chapter's shared structural limitation from both directions with a single underlying mechanism rather than two separate, task-specific fixes.

Chapter-4: Dataset

This chapter describes the two datasets used to train and evaluate SecondOpinion: a unified, five-class chest X-ray dataset, and a private pelvic fracture dataset stratified into visible and invisible cases. Both datasets were introduced in earlier chapters in the context of the studies that first used them, the CXR dataset in Chapter 3's architecture comparison and CheXNet-CBAM study, and the pelvic dataset's constituent training data in Chapter 3's segmentation, fusion, and PelFANet studies. This chapter consolidates and restates both in full, and specifies the particular preprocessing, mask, and evaluation choices made for SecondOpinion, which in places depart from how these same datasets were used earlier in this research line.

The two datasets were chosen for the same reason articulated in Chapter 1: they sit at different points along a difficulty gradient, multi-class CXR classification at the comparatively easier end, and pelvic fracture detection, further stratified into visible and invisible cases, at the harder end, allowing GateKeeper's activation behavior to be evaluated against a genuine spread of diagnostic difficulty rather than a single task. Because SecondOpinion's second stream is anatomy-guided, both datasets additionally require a segmentation mask alongside each raw image, and each dataset section below specifies where that mask comes from, whether it is available as verified ground truth, or whether it had to be produced by a trained segmentation model, since this distinction matters directly to how confidently the second stream's anatomical input can be trusted for a given case.

4.1 CXR Dataset

The chest X-ray portion of this research uses the same unified five-class dataset introduced in Chapter 3, combining two publicly available sources: the COVID-19 Radiography Database [107, 108] and the Tuberculosis (TB) Chest X-ray Dataset [109]. This section restates the dataset in full for completeness and specifies the preprocessing and evaluation protocol used for SecondOpinion specifically.

4.1.1 Sources and Composition

The COVID-19 Radiography Database contributes four of the five classes: 10,192 Normal, 6,012 Lung Opacity, 1,345 Viral Pneumonia, and 3,616 COVID-19 images, all originally provided at 299×299 resolution. The Tuberculosis Chest X-ray Dataset contributes the

fifth class, 700 Tuberculosis-positive images, originally provided at 512 × 512 resolution. Combining both sources produces a unified dataset of 21,865 images across five classes, summarized in Table 4.1 and illustrated in Figure 4.1.

Table 4.1: Composition of the unified five-class CXR dataset.

Class	Source	Image Count
Normal	COVID-19 Radiography Database	10,192
Lung Opacity	COVID-19 Radiography Database	6,012
Viral Pneumonia	COVID-19 Radiography Database	1,345
COVID-19	COVID-19 Radiography Database	3,616
Tuberculosis	TB Chest X-ray Dataset	700
Total		**21,865**

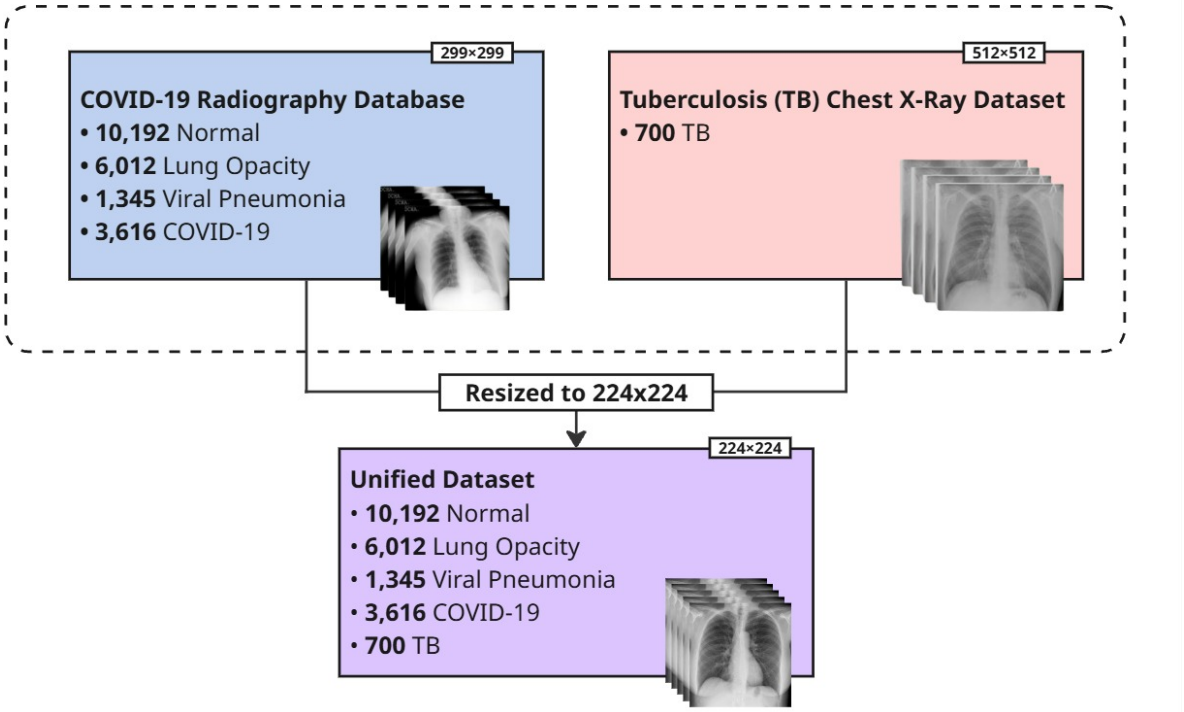

Figure 4.1: Construction of the unified five-class CXR dataset.

4.1.2 Lung Segmentation Masks

SecondOpinion's second, anatomy-guided stream requires a lung segmentation mask alongside each raw radiograph. For four of the five classes, Normal, Lung Opacity, Viral Pneumonia, and COVID-19, ground-truth lung masks are provided directly by the COVID-19 Radiography Database, having been produced and verified as part of its original release rather than generated separately for this research. Both prior studies in this research line that used this dataset, the architecture comparison and CheXNet-CBAM reviewed in Chapter 3, explicitly excluded these provided masks and worked from the raw radiographs alone [98, 106]; this work is the first point in this research line at which they are used.

The Tuberculosis Chest X-ray Dataset does not ship corresponding ground-truth masks for its 700 images. Consequently, the Tuberculosis class is the one exception to ground-truth mask availability in this unified dataset: rather than a verified ground-truth mask, every Tuberculosis case is paired with a fully predicted lung mask, generated by a segmentation model trained on the classes for which ground truth is available. The training and application of this segmentation model, and the broader mechanics of how predicted versus ground-truth masks feed into the second stream, are described in Chapter 5.

4.1.3 Preprocessing and Evaluation Protocol

Images from both source datasets, originally provided at 299×299 and 512×512 resolution respectively, were resized to a uniform 224×224 resolution to ensure consistency across the unified dataset and compatibility with the input requirements of the primary and second-stream backbones. Corresponding masks, whether ground-truth or predicted, were resized identically to preserve pixel-level alignment with their source image. No further preprocessing or data augmentation was applied to the CXR portion of this study; the resized images were used directly, without contrast enhancement, cropping, or synthetic augmentation of any kind.

The dataset was evaluated using five-fold cross-validation, with each fold serving as the held-out test set exactly once, so that reported CXR performance reflects an average over the full dataset rather than a single train-test split. This departs from the fixed 80:10:10 train-validation-test split used in the CheXNet-CBAM study reviewed in Chapter 3 [106], and was adopted here specifically to obtain a more robust estimate of GateKeeper's activation rate and the primary stream's per-case correctness, both of which are more sensitive to test-set composition than a single top-line accuracy figure would suggest.

4.2 AMERI Pelvic Fracture Dataset

The pelvic side of this research uses the AMERI PXR dataset, a private cohort of pelvic radiographs collected at Steel Memorial Hirohata Hospital, Japan, and used throughout the pelvic progression traced in Chapter 3, from the earliest architecture and preprocessing studies through PelFANet. Unlike the CXR dataset, which draws entirely on publicly available sources, AMERI is a clinically sourced, radiologist-confirmed dataset assembled specifically for pelvic fracture research within this same broader research effort [55], and its two constituent subsets, visible and invisible fracture cases, are what allow this research to evaluate SecondOpinion against the harder end of the difficulty gradient introduced in Chapter 1. The subsections below describe how the underlying data were acquired, how bone-level segmentation ground truth was annotated, and how the dataset is stratified into the visible and invisible fracture subsets used respectively for training and for held-out generalization testing.

4.2.1 Data Acquisition

The AMERI PXR dataset was collected at Steel Memorial Hirohata Hospital, Japan, between April 2013 and August 2019, under IRB approval (IRB No. 2019-1-52) [55]. The source cohort comprised 478 subjects, with a mean age of 64.22 ± 19.08 years and an age range of 20 to 93 years; 268 subjects were male and 209 were female, summarized in Table 4.2. Three-dimensional CT scans were acquired from 473 of these subjects, of whom 201 had confirmed pelvic fractures, using multidetector-row CT (MDCT) scanners operated at a tube voltage of 120 kVp with automatic tube current modulation (auto mAs). Alongside the CT data, a total of 481 pelvic X-ray (PXR) images were obtained from 315 subjects, of whom 199 had fractures confirmed across 365 fracture-positive images. In every case, the presence or absence of a fracture in both the 3D-CT and PXR modalities was confirmed by expert radiologists and physicians at Steel Memorial Hirohata Hospital, rather than inferred from imaging alone.

Table 4.2: Demographic data of the AMERI PXR source cohort.

Variable	AMERI PXR Source Cohort ($n = 478$)
Age (mean ± SD)	64.22 ± 19.08 years
Age Range	20–93 years
Male	268 (56.1%)
Female	209 (43.9%)
Subjects with 3D-CT	473
Subjects with confirmed pelvic fracture (CT)	201
PXR images obtained	481
Subjects contributing PXR images	315
Fracture-positive PXR images	365
Subjects with fracture-positive PXR	199

This source cohort is substantially larger than the subset actually used to train and evaluate SecondOpinion. Section 4.2.2 describes the segmentation annotation protocol applied to these images, and Sections 4.2.3 and 4.2.4 describe the exclusion criteria that reduce the cohort to the 228-image VIS subset and the separately curated 35-image INVIS subset.

4.2.2 Annotation Protocol

To support SecondOpinion's anatomy-guided second stream, every pelvic radiograph in the AMERI PXR dataset was manually annotated with bone-level segmentation masks. Nine distinct bones in the pelvic region, the left and right ilium, ischium, and pubis, the left and right proximal femurs, and the sacrum, were annotated separately as individual classes, following the same nine-bone segmentation scheme established in Beyond the Pelvic Ring [95]. Annotation was performed using the Medical Imaging Interaction Toolkit (MITK), with each bone's mask saved as a separate PNG file, where a pixel value of 1 denotes the target bone and 0 denotes background.

To ensure high-quality annotations and reduce individual annotator bias, a double-blind protocol was employed throughout. Two independent annotators produced separate masks

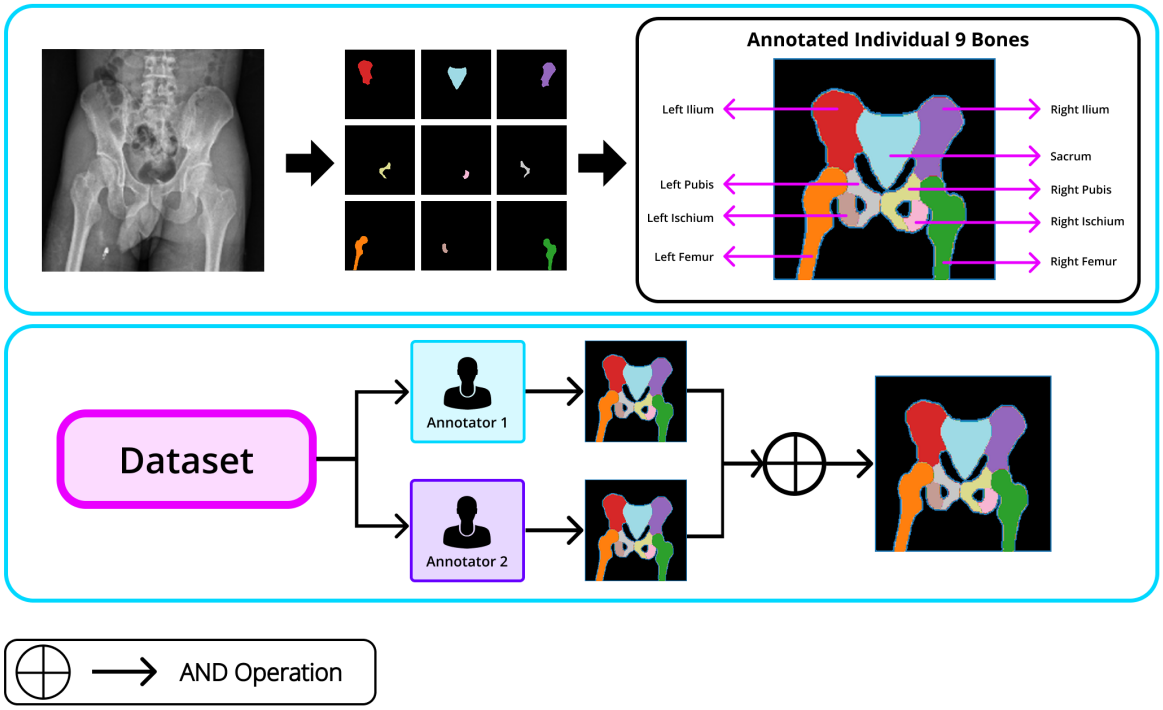

Figure 4.2: Annotation workflow for generating bone-level segmentation masks on the AMERI PXR dataset.

for each image, working without visibility into each other's annotations. The final ground truth for each bone was then obtained by applying a logical AND operation between the two annotators' masks, retaining only the regions both annotators agreed upon. This agreement-based construction discards any region one annotator marked but the other did not, favoring precision over recall in the ground-truth masks themselves, and produces segmentation labels that are consistent and reliable enough to serve as training and evaluation targets for the segmentation model described in Chapter 5. Figure 4.2 illustrates this workflow end to end, including the double-blind annotation stage, the AND operation used to reconcile the two annotators' masks, and the resulting breakdown into the nine individual bone classes.

This annotation protocol was applied uniformly across both the Visible Fracture (VIS) and Invisible Fracture (INVIS) subsets described in the following two sections, since both are drawn from the same source cohort and require the same nine-bone anatomical guidance for SecondOpinion's second stream.

4.2.3 Visible Fracture Cases

From the 481 PXR images and 315 subjects making up the AMERI source cohort described in Section 4.2.1, images exhibiting surgical implants or showing only partial pelvic regions were excluded, since both conditions compromise the reliability of the nine-bone segmentation annotation described in Section 4.2.2 and would introduce anatomical structure not representative of a standard pelvic radiograph. After applying these exclusion criteria, 228 PXR images remained, comprising 168 fracture-positive cases and 60 normal cases. This 228-image cohort constitutes the Visible Fracture (VIS) subset used throughout this

research, and is the same VIS subset used to train and evaluate PelFANet in Chapter 3 [103].

Every fracture case in the VIS subset is one in which the fracture, however subtle, is at least in principle visible on the two-dimensional PXR itself, the primary diagnostic signal a radiologist or an automated model would ordinarily rely on. This is also what makes VIS the appropriate subset for training: a model can only learn to recognize a visual pattern from cases in which that pattern is actually present in the input it is given.

Consistent with the CXR evaluation protocol described in Section 4.1, the VIS subset was evaluated using five-fold cross-validation, with each fold serving as the held-out test set exactly once. This ensures that reported VIS performance, like reported CXR performance, reflects an average over the full subset rather than a single favorable split, and keeps the evaluation protocol consistent across both diagnostic tasks examined in this research. The VIS subset serves as the sole source of training data for SecondOpinion's pelvic fracture stream; the INVIS subset, described next, is reserved exclusively for evaluation and is never used for training.

4.2.4 Invisible Fracture Cases

Alongside the 228-image Visible Fracture (VIS) subset described in Section 4.2.3, the AMERI PXR dataset includes a separately curated Invisible Fracture (INVIS) subset of 35 images. These images are drawn from the same source cohort, collected under the same acquisition protocol, and annotated under the same double-blind, nine-bone segmentation workflow described in Section 4.2.2. The imaging protocol, hospital, and annotation process are identical across both subsets. What separates an invisible fracture from a visible one is a property of the fracture itself: no fracture is apparent on the two-dimensional PXR, yet its presence is confirmed by the corresponding 3D-CT scan of the same patient.

Figure 4.3 shows why. The zoomed region of the PXR shows no discernible cortical disruption and nothing that would prompt a radiologist reading the X-ray alone to flag a fracture at that location. The corresponding 3D-CT scan of the same patient tells a different story: across axial, coronal, and sagittal views, the fracture is unambiguous, with the annotated region in the axial slice showing a clear discontinuity in the sacral bone that the PXR gives no indication of. This is not a fracture a more careful reading of the X-ray would eventually catch. The information needed to detect it is simply absent from the two-dimensional projection; PXR imaging compresses a three-dimensional anatomical volume into a single view, and a fracture line that happens to align with, or sit behind, overlapping bone density in that projection can be rendered effectively invisible, even though the underlying injury is unambiguous in three dimensions.

This is a qualitatively different kind of difficulty, not simply a harder version of the same task. A visible fracture, however subtle, still leaves some trace in the input a PXR-only model actually receives, and a sufficiently well-designed architecture can learn to recognize that trace. An invisible fracture leaves no such trace. Whatever a model predicts on these cases has to come from something other than a directly observable pattern in the image at hand, whether a weaker indirect cue such as asymmetry, altered joint spacing,

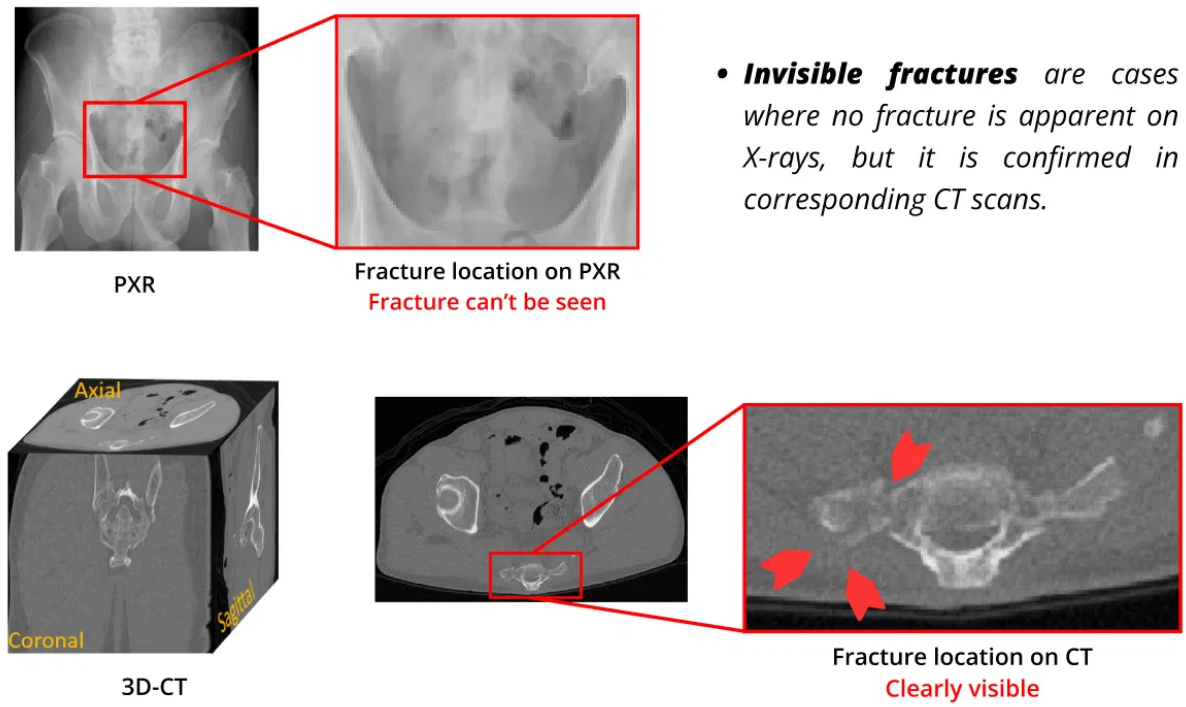

Figure 4.3: An invisible fracture case from the AMERI PXR dataset.

or abnormal alignment elsewhere in the pelvis, or generalization from patterns learned on visible fractures that happen to transfer to the harder case. This is why the difficulty gradient introduced in Chapter 1 places pelvic fracture detection at its harder end, with INVIS specifically, rather than VIS, as its upper bound. Rahman et al.'s two-step transfer learning approach, discussed in Chapter 2, reported a substantial AUROC gap between visible and invisible fracture performance, 0.933 against 0.801 [55]. PelFANet, reviewed in Chapter 3, showed a comparable pattern, 0.9334 AUROC on VIS against 0.8688 on INVIS [103], despite never having been trained on a single invisible fracture case.

This is why the 35 INVIS images are reserved exclusively for evaluation and never enter any training split, for either the primary stream or SecondOpinion's second, anatomy-guided stream. This is a deliberate choice, not a limitation imposed by the subset's small size. Because an invisible fracture is defined by the absence of a directly learnable visual pattern, training on INVIS cases would risk teaching a model to memorize specific instances rather than genuinely generalize from what it learned on VIS. Evaluating on INVIS is therefore a direct test of generalization under exactly the conditions where fixed, always-on anatomical reasoning would be expected to matter most, and where GateKeeper's activation rate, described in Chapter 5, would be expected to rise most sharply relative to VIS and the CXR dataset described in Section 4.1.

4.2.5 Dataset Summary

Table 4.3 summarizes the two AMERI subsets used in this research. Together, VIS and INVIS total 263 images drawn from the same 478-subject source cohort, with VIS supplying both the training data and one evaluation split for SecondOpinion's pelvic fracture stream, and INVIS supplying a held-out evaluation set the model never trains on.

Table 4.3: Summary of the AMERI VIS and INVIS subsets.

Subset	Fracture	Normal	Total
VIS	168	60	228
INVIS	35	0	35

Table 4.4 places both AMERI subsets alongside the CXR dataset from Section 4.1. The three together span the difficulty gradient introduced in Chapter 1: multi-class CXR classification at one end, where the diagnostic signal for most cases is directly present in the image; VIS pelvic fracture detection in the middle, where the signal is present but often subtle; and INVIS at the far end, where the signal that a PXR-only model would ordinarily rely on is absent from the input entirely.

Table 4.4: Comparison of the CXR and AMERI datasets used in this research.

Property	CXR	AMERI VIS	AMERI INVIS
Task	5-class disease classification	Binary fracture detection	Binary fracture detection
Images	21,865	228	35
Classes/Fracture cases	5 classes	168 fracture, 60 normal	35 fracture
Mask source	Ground truth (4 classes), predicted (TB)	Ground truth, double-blind	Ground truth, double-blind
Split protocol	5-fold CV	5-fold CV	Evaluation only

Both datasets require the same underlying anatomical guidance signal for SecondOpinion's second stream, a segmentation mask paired with each raw image, but differ in how reliably that signal is available. For CXR, ground truth is available for four of five classes, with only the Tuberculosis class relying on a predicted mask. For AMERI, every image in both VIS and INVIS carries a double-blind, nine-bone ground-truth mask, since the segmentation annotation described in Section 4.2.2 was applied uniformly across the full curated cohort before the VIS/INVIS split was made. This distinction matters directly to Chapter 5: it determines which cases SecondOpinion's second stream can draw on a fully reliable anatomical prior for, and which cases depend on a segmentation model's own prediction instead.

Chapter-5: Proposed Method

The preceding chapters motivate this research from two directions that ultimately converge here. Chapter 2 traced a research gap at the intersection of anatomy-guided learning and conditional computation: dual-stream architectures improve diagnostic performance by incorporating anatomical structure explicitly, but every prior design in this line, and every dual-stream design surveyed more broadly, evaluates both streams on every input regardless of whether the case actually warrants the additional computation. Chapter 3 traced this work's own research lineage arriving at the same conclusion from the opposite direction, moving from a single architecture benchmark, through preprocessing and patch-based refinements, to explicit anatomical guidance in Beyond the Pelvic Ring and PelFANet [103], each study improving on the last but each also paying the same fixed, always-on computational cost regardless of how confidently a case could already be resolved. Chapter 4 described the two datasets, the unified CXR dataset and the AMERI pelvic fracture dataset, that this chapter's proposed method is evaluated against, with particular attention to the segmentation masks that make anatomy-guided processing possible for every case in both datasets.

This chapter presents SecondOpinion [110], a framework that resolves the tension between these two lines of work directly. Rather than treating anatomical guidance as a fixed architectural commitment applied uniformly to every input, SecondOpinion treats it as a conditional resource, invoked only for the subset of cases that a fast, general-purpose stream cannot already resolve with confidence. The clinical analogy that motivates this design is explicit and deliberate: a clinician does not routinely order a second, more deliberate reading of every case that crosses their desk. Straightforward cases are resolved on first inspection; only genuinely ambiguous ones prompt a second opinion. SecondOpinion is an attempt to give a diagnostic model the same adaptive discipline, formalized through a learned gating mechanism rather than a fixed rule or heuristic threshold.

The remainder of this chapter is organized as follows. Section 5.1 gives an overview of the full pipeline and the design rationale connecting its parts. Section 5.2 describes the dual-stream architecture in full, including the EfficientNet-B0 backbones shared by both streams, the segmentation pipeline that constructs Stream B's anatomy-guided input, and the cross-attention fusion mechanism that combines the two streams when both are active. Section 5.3 introduces GateKeeper, the learned gating mechanism that decides, on a per-case basis, whether Stream B is needed. Section 5.4 details the two-phase training procedure used to train the full system. Section 5.5 describes the evaluation framework used to assess SecondOpinion in Chapter 6, covering both the classification metrics and the activation-rate accounting specific to a conditionally computed system, with each metric stated formally.

5.1 Overview

SecondOpinion is built around a single governing principle: every input is processed by a fast primary stream, and a second, anatomically informed stream is invoked only when a learned signal indicates that the primary stream's prediction cannot be trusted on its own. This principle translates into three structural components, each addressed in its own section below. A dual-stream architecture, comprising Stream A on the raw image and Stream B on an anatomy-guided representation, together with the cross-attention fusion mechanism that combines them, is described in Section 5.2. GateKeeper, the gating mechanism that decides, per sample, whether Stream B should be activated, is described in Section 5.3. The two-phase training procedure that makes it possible to train all of these components jointly despite the discrete, non-differentiable nature of the gating decision itself is described in Section 5.4.

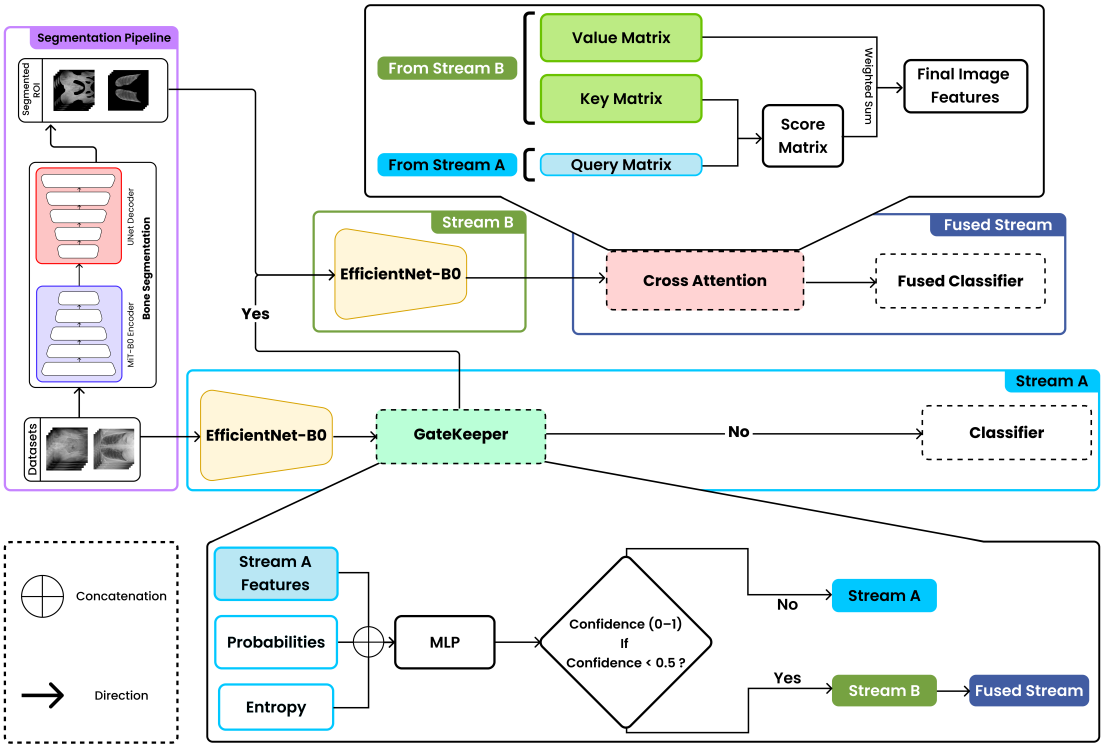

Figure 5.1: SecondOpinion full pipeline.

Figure 5.1 illustrates the full pipeline. Reading it left to right traces the path any input takes through the system: the raw image enters Stream A directly, while a parallel segmentation pipeline produces the anatomy-guided representation that would be used by Stream B if activated. Stream A's pooled features, together with its output probabilities and their entropy, are consumed by GateKeeper, whose output determines the branch the sample takes. On the majority of inputs, GateKeeper's confidence estimate is high enough that the sample proceeds directly to classification from Stream A alone, and Stream B is never evaluated. On the remaining inputs, Stream B is evaluated, its features are fused with Stream A's through cross-attention, and classification proceeds from the fused representation instead. GateKeeper's own internal decision process, summarized here only at the level of what it consumes and what it outputs, is described in full architectural detail in Section 5.3, with its own dedicated diagram.

This design deliberately departs from the architecture used throughout Chapter 3's progression of prior approaches in one specific respect. Every anatomy-guided design examined there, from the segmentation-guided ROI approach through the fusion ensemble to PelFANet itself, treats anatomical guidance as an unconditional architectural commitment: the anatomical pathway runs on every sample, whether or not that sample's diagnosis is actually ambiguous enough to benefit from it. PelFANet in particular, the immediate architectural predecessor to the pelvic side of this research, processes the raw pelvic radiograph and its segmented bone representation through two always-on convolutional branches, exchanging and refining features between them at every one of its eight stacked Fused Attention Blocks regardless of case difficulty [103]. SecondOpinion instead treats the decision of whether to consult anatomical guidance as something the model itself should learn to make, conditioned on the specific case in front of it, rather than something fixed in advance by the architecture. The remaining sections of this chapter describe how each component of the pipeline is built to support that decision.

5.2 Dual-Stream Architecture

5.2.1 EfficientNet-B0 Streams

SecondOpinion's two streams share a common backbone family but differ entirely in what they are given to look at, in what they are asked to accomplish, and, as later sections describe, in when they run at all. This subsection describes each stream's construction in turn, beginning with the primary stream that processes every input, followed by the anatomy-guided stream and the segmentation pipeline that feeds it, which together carry the bulk of this chapter's architectural detail.

5.2.1.1 Stream A: Primary Pathway

Stream A is the pathway every input passes through, and its design goal is accordingly different from Stream B's: rather than maximizing diagnostic accuracy on its own, Stream A must be fast, generally reliable across the full range of cases in a given task, and must produce internal signals, features, output probabilities, and their associated uncertainty, that GateKeeper can use to judge whether its own prediction should be trusted.

Stream A takes as input the raw image x_a, resized to 224×224 resolution with three channels, matching the preprocessing protocol described in Chapter 4 for both the CXR and pelvic datasets. This raw input is passed through an EfficientNet-B0 backbone, initialized from ImageNet-pretrained weights rather than trained from scratch. EfficientNet-B0 was chosen for this role for reasons directly connected to the architecture benchmarking study reviewed in Section 3.1: rather than reopening the architecture selection question from scratch, this research carries forward the same reasoning established there and in the CXR architecture comparison study reviewed in Section 3.2, that a compact, well-regularized convolutional backbone offers a favorable balance of accuracy and computational cost for this class of task, without the added complexity or data hunger of a transformer-based backbone. EfficientNet-B0 additionally offers a specific practical advantage for

the conditional-computation setting SecondOpinion is built around: its low parameter count and FLOP footprint mean that, on the large fraction of samples where Stream B is never activated, the total cost of processing that sample stays close to the cost of a single lightweight backbone rather than a much larger unconditional model.

The backbone produces two outputs from the raw input. Global average pooling over the final convolutional feature map yields a pooled feature vector $h_a \in \mathbb{R}^{1280}$, a fixed-length summary of the image that discards spatial resolution but retains the information most relevant to classification. This pooled vector is passed through a linear classification head to produce logits $z_a \in \mathbb{R}^C$, where C is the number of classes for the task at hand, five for the CXR dataset and two for the pelvic fracture datasets. Both h_a and z_a serve two purposes in the larger pipeline. First, z_a is Stream A's own prediction, used directly as the final output for any sample GateKeeper does not flag for further scrutiny. Second, h_a and the softmax probabilities derived from z_a are the primary inputs GateKeeper uses to decide whether that prediction can be trusted, described fully in Section 5.3.

It is worth being explicit about what Stream A does not have access to. Stream A never sees the segmentation mask or any anatomy-guided representation; it operates on the raw image alone, exactly as a single-stream baseline would. This is a deliberate design choice rather than an oversight. Keeping Stream A anatomy-blind ensures that the gap between Stream A's performance and the fused model's performance on any given case is attributable specifically to the value anatomical guidance adds for that case, rather than being confounded by Stream A already having partial access to anatomical information through some shared representation. This separation also directly supports the ablation structure used in Chapter 6, where Stream A evaluated in isolation serves as a lower-bound baseline against which the value of Stream B and the fusion mechanism can be measured.

5.2.1.2 Stream B: Anatomy-Guided Pathway

Where Stream A is designed to be fast and generally reliable, Stream B is designed to be diagnostically thorough on the specific cases where general-purpose reasoning from the raw image alone is judged insufficient. Its defining characteristic, inherited from the anatomy-guided design philosophy traced across Chapter 3, is that it does not reason over the raw image directly, but over a representation that has been explicitly informed by the underlying anatomical structure of the case.

Segmentation Pipeline. The anatomy-guided representation x_b that Stream B consumes is derived from a segmentation pipeline that sits upstream of the classification model entirely, trained and applied separately from the rest of SecondOpinion. Rather than designing this pipeline from scratch, this research carries forward the segmentation architecture and training protocol established by PelFANet, the immediate predecessor to SecondOpinion on the pelvic side of this research line [103], applying the same approach to both the pelvic and CXR datasets used in this research.

The segmentation model itself is a U-Net architecture with a Mix Transformer B0 (MiT-B0) encoder [96, 97], a hybrid design that combines the U-Net decoder's established strength at pixel-level spatial reconstruction, via a symmetric decoder with skip connections that

recover fine spatial detail lost during downsampling, with a transformer-based encoder capable of modeling long-range dependencies across the image, capturing global anatomical context that a purely convolutional encoder can miss. The MiT-B0 encoder processes the input through a hierarchical sequence of four stages, each operating at progressively coarser spatial resolution and producing its own feature map, with self-attention computed within each stage rather than globally across the full-resolution image, keeping the encoder computationally lightweight despite its transformer-based design. This staged, pyramid-like structure is precisely what allows the encoder to be paired naturally with a U-Net decoder: each of the decoder's upsampling stages consumes the skip connection from the encoder stage of matching resolution, so that fine boundary detail recovered from early, high-resolution encoder stages is combined with the coarser, more semantically rich context accumulated in later, low-resolution stages, giving the decoder access to both precise anatomical boundaries and a broader sense of where the anatomical structure of interest is likely to be located overall. The encoder specifically is the lightweight MiT-B0 variant, pretrained on ImageNet before being fine-tuned for the segmentation task, following the same encoder choice used throughout this research line's segmentation stages. Implementation follows the same modular segmentation library convention used in PelFANet, which provides ready-made support for pairing a U-Net decoder with a range of transformer-based encoders including MiT-B0 [103].

Consistent with PelFANet's own segmentation design, and in deliberate contrast to the nine-bone multi-class segmentation explored separately in the Beyond the Pelvic Ring study reviewed in Section 3.1, the segmentation model used here treats its target anatomical region, the full pelvic bone structure for the pelvic dataset, or the lung fields for the CXR dataset, as a single foreground class against background, rather than decomposing it into multiple anatomically distinct sub-classes. Mechanically, this means the segmentation model's final layer produces a single output channel per pixel, passed through a sigmoid activation to yield a per-pixel probability that the pixel belongs to the anatomical region of interest, rather than the multi-channel softmax output a nine-class segmentation model would require to assign each pixel to one of several competing bone classes plus background. The model is optimized with Dice loss, chosen for its direct correspondence to the overlap-based metrics, Dice Similarity Coefficient and Intersection over Union, used to evaluate segmentation quality, and for its established robustness to the foreground-background class imbalance typical of anatomical segmentation, where the anatomical region of interest occupies a comparatively small fraction of the total image area.

This single-class design was carried forward from PelFANet specifically because SecondOpinion's central architectural contribution lies in the conditional gating mechanism described in Sections 5.3 and 5.4, not in the granularity of anatomical segmentation itself; holding the segmentation strategy fixed to an established, previously validated design isolates the effect of conditional computation from any confound introduced by simultaneously changing how anatomical structure is represented. A nine-bone segmentation, of the kind explored in Beyond the Pelvic Ring, would in principle provide Stream B with more anatomically granular input, but adopting it here alongside a newly introduced gating mechanism would make it impossible to attribute any resulting change in performance cleanly to either change individually, precisely the same methodological concern that motivated PelFANet's own choice to hold segmentation granularity fixed while varying its fusion mechanism, discussed further in Section 3.1.

Once trained, the segmentation model is used to produce a predicted mask for every

image in both datasets, and it is this predicted mask, rather than any ground-truth annotation, that Stream B's input is constructed from at both training and inference time. This is an important and deliberate distinction from how anatomical information is used elsewhere in this research line and worth stating precisely: even for images where a verified, expert-annotated ground-truth mask exists, as it does for four of the five CXR classes and for the full AMERI pelvic dataset as described in Chapter 4, Stream B is never given that ground-truth mask directly. It is trained and evaluated exclusively on the segmentation model's own predictions, so that its behavior at inference time, where no ground truth is ever available for a genuinely new case, matches its behavior during training exactly. Ground-truth masks, where available, are used only to train the upstream segmentation model itself, never to construct Stream B's input directly. This choice matters for how SecondOpinion's reported performance should be interpreted: because Stream B never sees a shortcut in the form of ground-truth anatomical structure, any diagnostic benefit it provides reflects the value of anatomical guidance as it would actually be available in a deployed system, rather than an inflated benefit that would disappear once ground-truth masks were no longer available at inference time, as they never are for a genuinely new patient.

For the Tuberculosis class specifically, which as described in Chapter 4 lacks ground-truth masks entirely even for training the segmentation model, the predicted mask used to construct x_b comes from a segmentation model trained exclusively on the four CXR classes for which ground truth does exist, applied to Tuberculosis images purely at inference time within the segmentation stage. This is a form of generalization in its own right: the segmentation model must transfer its learned notion of lung field boundaries to a class of image it never saw during its own training, and any error this introduces propagates forward into the quality of Stream B's anatomical input for that class specifically. This is structurally analogous to the cross-dataset generalization PelFANet's own segmentation model demonstrated when applied outside its training distribution, and the same underlying expectation, that a well-trained one-class segmentation model can transfer its notion of anatomical boundary to related but unseen data, motivates carrying it forward into this new setting.

Backbone and Feature Extraction. Once constructed, x_b is passed through a second EfficientNet-B0 backbone, matching Stream A's architecture exactly but with independently initialized and independently trained weights. No parameters are shared between the two streams at any point. This independence is not an incidental implementation detail but a direct extension of the clinical analogy underpinning the whole framework: a genuine second opinion is formed by a separate line of reasoning over the case, not by re-running the same reasoning process with a slightly different input. Sharing early-layer weights between the two streams was considered and deliberately avoided for this reason, since doing so would risk Stream B inheriting the same failure modes and blind spots that led Stream A's prediction to be flagged as unreliable in the first place, undermining the independence the second opinion is meant to provide.

Stream B's backbone produces two outputs, mirroring Stream A's but retaining more spatial detail. Global average pooling again yields a pooled feature vector $h_b \in \mathbb{R}^{1280}$, used identically to h_a as an input to downstream fusion. In addition, and unlike Stream A, Stream B retains its final spatial feature map without pooling, $S_b \in \mathbb{R}^{1280\times7\times7}$, a 7×7 grid of 1280-dimensional feature vectors corresponding to spatial regions of the input. This

spatial map is retained specifically because it is required by the cross-attention fusion mechanism described in Section 5.2.2, which needs Stream B's features to remain spatially resolved rather than collapsed into a single pooled vector, so that Stream A can attend selectively to different anatomical regions rather than to an undifferentiated summary of the whole image.

Stream B is, by construction, always more computationally expensive to evaluate than Stream A, since its input has already passed through an upstream segmentation model before ever reaching its own backbone. This is precisely the cost that GateKeeper exists to make conditional rather than universal, and Section 5.3 describes how that conditional decision is made.

5.2.2 Cross-Attention Fusion

When GateKeeper activates Stream B for a given sample, Stream A's prediction is no longer used directly. Instead, the two streams' features are combined through a dedicated fusion module before a final prediction is produced. The design question at this stage is not merely how to combine two feature vectors, but how to combine them in a way that reflects the asymmetric relationship between the two streams established by the framework's central analogy: Stream A is the primary reasoner seeking a second opinion, and Stream B is the source of that second opinion, not an equal and interchangeable partner in a symmetric combination.

Motivation for Cross-Attention over Simple Combination. A simpler alternative, averaging or concatenating h_a and h_b directly before a shared classification head, was available and was considered against the cross-attention approach ultimately adopted. Such a symmetric combination treats every spatial region of Stream B's anatomical representation as equally relevant to every prediction, discarding the specific reason Stream B was invoked in the first place: GateKeeper flagged this particular sample as one where Stream A's reasoning needs correction, which implies that some regions of the anatomical representation are likely to be more informative than others for correcting that specific error. A fixed combination rule cannot express this selectivity, since it applies the same weighting to every spatial region regardless of what that region actually contains for the specific case at hand. Cross-attention can, by construction, since it allows the query, drawn from Stream A, to determine which regions of the key and value, drawn from Stream B, receive weight in the final fused representation, on a sample-by-sample basis rather than uniformly across the dataset.

This design also departs, deliberately, from the fusion strategy used in PelFANet, the closest architectural predecessor on the pelvic side of this research line. PelFANet fuses its two streams iteratively, at every one of eight stacked FABlocks distributed throughout the network, using a CBAM-based attention mechanism that refines a concatenated feature map before splitting it back into two separately continuing branches [103]. This iterative, symmetric fusion strategy is well suited to a design in which both streams are always active and the network is free to exchange information between them repeatedly throughout the forward pass. SecondOpinion's fusion, by contrast, happens once, late, and only when Stream B has actually been evaluated, and it is explicitly asymmetric: Stream A

queries Stream B, not the reverse. This is a direct architectural consequence of conditional computation rather than an independent design preference, since a fusion mechanism distributed throughout the network, in the style of PelFANet's FABlocks, would presuppose that both streams are always running, which is exactly the assumption SecondOpinion is built to avoid.

Fusion Mechanism. The fusion module operates on Stream A's pooled feature vector h_a, treated as query, and Stream B's retained spatial feature map $S_b \in \mathbb{R}^{1280\times7\times7}$, flattened into $N = 49$ spatial tokens and treated as key and value. Note that Stream A itself does not retain a spatial feature map analogous to S_b in the base configuration described in Section 5.2.1; to serve as a query source for cross-attention, Stream A's own final feature map S_a, prior to global pooling, is likewise retained and flattened into the same 49-token layout, so that query and key-value pairs share a compatible spatial resolution. Each of the three streams, query, key, and value, is projected through its own learned linear projection into a shared dimensionality d, chosen to be smaller than the backbone's native 1280-dimensional feature width specifically to keep the attention computation itself lightweight, consistent with the broader efficiency goals of the framework, rather than projecting the full 1280-dimensional representation directly into the attention mechanism, which would multiply the cost of the query-key similarity computation without a corresponding gain in expressiveness given the comparatively small, 49-token spatial resolution being attended over.

$$Q = S_a W_Q, \qquad K = S_b W_K, \qquad V = S_b W_V \tag{5.1}$$

Attention is computed as standard scaled dot-product attention over a single head,

$$\tilde{F} = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V \tag{5.2}$$

producing a fused spatial output $\tilde{F}$ with the same 49-token structure as the inputs. A single attention head, rather than a multi-head arrangement, was chosen deliberately to keep the fusion module lightweight, consistent with the broader efficiency goals of the framework: the fusion mechanism is only ever invoked on the minority of cases GateKeeper flags, but keeping its own cost low ensures that even the more expensive, Stream-B-activated path remains substantially cheaper than an unconditional dual-stream model, of the kind PelFANet represents, evaluated on every sample would be.

The attended output $\tilde{F}$ is mean-pooled across its 49 spatial tokens and layer-normalized, producing a single fused vector $\tilde{f} \in \mathbb{R}^d$,

$$\tilde{f} = \mathrm{LayerNorm}\left(\frac{1}{N}\sum_{i=1}^{N}\tilde{F}_i\right) \tag{5.3}$$

Reducing $\tilde{F}$ to a single pooled vector at this stage, rather than retaining its spatial structure into the classification head, is a deliberate simplification rather than an oversight:

because the attention weights inside Equation 5.2 have already determined which spatial regions of Stream B contribute most to the fused representation, the pooling step at Equation 5.3 discards positional structure only after the mechanism responsible for spatial selectivity has already done its work, so nothing that the fusion module was specifically designed to capture is lost by collapsing to a pooled vector at this final stage.

This pooled vector $\tilde{f}$ is concatenated with the original pooled features from both streams, h_a and h_b, and passed through a final classification head to produce the fused logits,

$$z_{\text{fused}} = \text{Head}\big([\tilde{f}\,;\, h_a\,;\, h_b]\big) \tag{5.4}$$

the prediction used for any sample GateKeeper routes through the fused pathway. Concatenating the raw pooled features alongside the attended output, rather than relying on the attended output alone, preserves direct access to each stream's own summary representation even after fusion, so that the final classification head is not forced to reconstruct information from the attention output that was already available more directly in h_a and h_b.

5.3 GateKeeper

GateKeeper is the component that gives SecondOpinion its adaptive character. Its role is narrow and specific: for each input, estimate the probability that Stream A's prediction, made without any anatomical guidance, is already correct. If that estimated probability is high, there is no diagnostic value in consulting Stream B, and the sample proceeds through the pipeline at the cost of a single lightweight backbone. If that estimated probability is low, Stream A's prediction cannot be trusted on its own, and the sample is routed to Stream B and subsequent fusion.

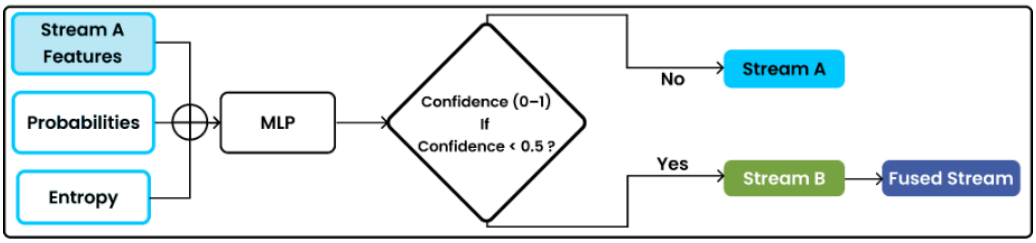

Figure 5.2: GateKeeper's internal architecture.

5.3.1 Architecture

Figure 5.2 shows GateKeeper's internal structure in isolation from the rest of the pipeline, expanding on the compact representation of GateKeeper shown within Figure 5.1.

Design Rationale: Correctness Supervision over Unsupervised Confidence. The most immediately obvious way to build a gate of this kind would be to use Stream

A's own softmax output directly as a confidence signal, activating Stream B whenever the maximum softmax probability falls below some threshold. This is, in essence, the strategy used by early-exit networks, the most established line of work on conditional computation reviewed in Chapter 2, where a gating decision is typically derived from the network's own unsupervised confidence rather than from an explicitly supervised correctness signal. This research deliberately departs from that convention. Softmax confidence is well known to be an unreliable proxy for correctness in modern deep networks: models are frequently overconfident on inputs they get wrong, and the raw probability a network assigns to its own top prediction does not necessarily track the actual likelihood that prediction is correct. Relying on this signal directly would risk a gate that fails to activate Stream B precisely on the cases where Stream A is confidently wrong, which are exactly the cases a second opinion would be most valuable for.

GateKeeper instead is trained explicitly as a binary correctness classifier, with a supervision target constructed directly from ground truth rather than derived implicitly from the training dynamics of the classification loss. This design choice is motivated directly by recent findings in the early-exit literature, specifically evidence that explicitly aligning a gating mechanism's training objective with its inference-time decision policy produces more reliable gating behavior than gates trained only as a byproduct of an unrelated task loss. GateKeeper's training target is defined precisely for this reason as whether Stream A's prediction actually matches the ground-truth label for that sample, described formally in Section 5.4.2, so that the quantity GateKeeper is optimized to predict is exactly the quantity its gating decision is meant to reflect. A consequence of this design is that α is, by construction, a calibrated estimate in a specific and useful sense: because its training target $\hat{g}$ is a genuine binary indicator of correctness rather than a soft or heuristic proxy for it, and because it is trained with binary cross-entropy, the loss function whose minimizer is exactly the true conditional probability of the target given the input, α is trained toward being interpretable directly as $P(\text{Stream A is correct} \mid x)$, rather than merely as a score that happens to correlate with correctness. This is what licenses treating 0.5 as a principled decision boundary in Section 5.3.2, rather than an arbitrary cutoff chosen for convenience.

Formulation. GateKeeper takes three signals from Stream A as input, each capturing a different aspect of how trustworthy Stream A's prediction on a given sample appears to be. The first is Stream A's pooled feature vector $h_a \in \mathbb{R}^{1280}$ itself, giving GateKeeper access to the same internal representation Stream A used to arrive at its prediction, which may contain signal about case difficulty beyond what the output probabilities alone reveal, for instance features associated with image quality, artifact presence, or atypical anatomical presentation that never surface explicitly in the softmax output but may nonetheless be encoded somewhere in the pooled representation. The second is the softmax probability vector $p_a = \text{softmax}(z_a)$, the class-wise probabilities Stream A assigns to the sample, a comparatively low-dimensional signal, five-dimensional for the CXR task and two-dimensional for the pelvic tasks, but one that speaks directly to Stream A's own stated confidence in each candidate class. The third is the Shannon entropy of that probability distribution,

$$H(p_a) = -\sum_{i=1}^{C} p_{a,i} \log(p_{a,i} + \epsilon) \tag{5.5}$$

where ϵ is a small constant added for numerical stability when a class probability approaches zero. Entropy provides GateKeeper with a single scalar summary of how spread out Stream A's prediction is across classes, distinct from the raw probability vector itself: a low-entropy distribution indicates Stream A is concentrating its belief on one class, while a high-entropy distribution indicates genuine uncertainty spread across several plausible classes, a distinction the raw probability vector expresses only implicitly and that becomes considerably more informative as a single explicit scalar than as something the small MLP downstream would otherwise have to learn to derive from the raw probability vector on its own.

These three signals, the pooled feature vector, the probability vector, and the scalar entropy, are concatenated into a single input vector and passed through a small multilayer perceptron, denoted $\phi(\cdot)$, followed by a sigmoid activation, producing GateKeeper's output

$$\alpha = \sigma\big(\phi([h_a\,;\,p_a\,;\,H(p_a)])\big) \in (0,1) \tag{5.6}$$

α is interpreted directly as GateKeeper's estimate of the probability that Stream A's prediction is correct for the given sample. GateKeeper is deliberately kept small relative to the two backbones it sits alongside, since its role is to interpret a compact, already-extracted summary of Stream A's behavior rather than to perform any additional feature extraction of its own from raw pixel data; this asymmetry in scale is itself part of the efficiency argument for the framework, since GateKeeper's own cost is negligible next to the cost of evaluating either backbone, and is incurred on every sample regardless of Stream B's activation. The dominant contribution to GateKeeper's input dimensionality comes from h_a itself, at 1280 dimensions, against which p_a and $H(p_a)$ contribute a comparatively small number of additional dimensions; despite this imbalance, both are retained explicitly alongside h_a rather than being left for $\phi(\cdot)$ to rederive from h_a alone, since p_a and $H(p_a)$ are computed downstream of h_a through Stream A's own classification head and are not trivially recoverable from h_a by a small MLP without effectively re-learning that head's function internally.

5.3.2 Hard Gate Threshold

GateKeeper activates Stream B according to the decision rule

$$\text{Activate Stream B} \iff \alpha < 0.5 \tag{5.7}$$

This threshold is not a hyperparameter tuned separately for each task or dataset, but the natural decision boundary implied by the binary cross-entropy objective GateKeeper is trained with, described in Section 5.4.2. Because α is trained to approximate a genuine posterior probability of correctness rather than an arbitrary heuristic score, as argued in Section 5.3.1, 0.5 is the threshold at which GateKeeper's own training objective already defines the switch from predicting Stream A correct to predicting Stream A incorrect, and no additional tuning is required to select it.

This is a meaningful departure from how activation thresholds are typically set in

conditional-computation systems, where a confidence cutoff is often swept over a validation set and chosen post hoc to hit a particular accuracy-efficiency trade-off. Such post hoc tuning would be a reasonable strategy if GateKeeper's output were an arbitrary score with no inherent interpretation, but it becomes an unnecessary complication once that output is trained to mean something specific, a probability of correctness, since the point at which "probably correct" becomes "probably incorrect" is fixed by the definition of the quantity itself rather than by whatever trade-off happens to look best on a validation curve. A tuned threshold would also introduce an additional axis of variation between the CXR and pelvic experiments reported in Chapter 6, since a threshold tuned separately per task could no longer be interpreted as evidence that GateKeeper's activation behavior tracks a shared, task-general notion of correctness; fixing the threshold at its principled value of 0.5 across every task keeps that comparison uncontaminated.

It is worth being explicit about what GateKeeper is not optimized to do directly. Its training objective, detailed fully in Section 5.4.2, targets correctness prediction, not computational efficiency. No explicit penalty or reward term for activation rate appears anywhere in its loss. Any low activation rate that SecondOpinion exhibits at inference is not the result of GateKeeper being pushed toward sparsity directly, but falls out as a natural consequence of a well-calibrated correctness classifier applied to a task where Stream A, despite lacking anatomical guidance, is already correct on the majority of cases. Efficiency, in other words, is a consequence of accurate gating rather than a design target pursued independently of it. Section 5.5.2 returns to this point when defining how activation rate is measured and interpreted for the results reported in Chapter 6.

5.4 Two-Phase Training Strategy

SecondOpinion's three trainable components, Stream A, Stream B, and GateKeeper, cannot be trained independently of one another in a naive sequential fashion, since GateKeeper's supervision target is itself defined in terms of Stream A's correctness, which is only meaningful once Stream A has already learned something non-trivial about the task. Training instead proceeds in two phases, moving from a phase in which both streams learn the underlying task without any gating decision affecting either of their gradients, to a phase in which GateKeeper is introduced and all three components are trained jointly.

5.4.1 Phase 1: Frozen GateKeeper

In the first phase, GateKeeper remains frozen and plays no role in either stream's training. The reasoning behind this is direct: at the very start of training, neither stream has learned enough about the task for the notion of "Stream A's correctness" to be a stable or meaningful target. Training GateKeeper against a moving, largely uninformative target this early would risk it converging to a degenerate solution, for instance always predicting high confidence simply because Stream A's early-training predictions are close to random and therefore only weakly separable into correct and incorrect cases in any consistent way, a failure mode that, once GateKeeper's own parameters had settled into it, could persist well after Stream A's predictions had become genuinely informative, since a degenerate gate has little training signal available to pull it back out of a trivial solution.

Instead, both streams are run on every sample unconditionally during this phase, and both are trained directly against the ground-truth label, with no gating or fusion-dependent routing involved yet. The Phase 1 objective is

$$\mathcal{L}_{\text{phase1}} = \mathcal{L}_{\text{CE}}(z_a, y) + \mathcal{L}_{\text{CE}}(z_{\text{fused}}, y) \tag{5.8}$$

where $\mathcal{L}_{\text{CE}}$ denotes standard cross-entropy loss and y is the ground-truth label. Note that z_{fused} here is computed with Stream B and the fusion module active on every sample, unconditionally, since GateKeeper is not yet making any routing decision during this phase; the conditional routing behavior that defines SecondOpinion at inference time is a Phase 2 and inference-time property, not a Phase 1 one. Training both branches, Stream A alone and the fused pathway, against the same ground truth during this phase ensures that by the time Phase 2 begins, both Stream A's predictions and the fused predictions are already meaningful enough for GateKeeper to be trained against a genuinely informative correctness signal derived from them, avoiding exactly the degenerate cold-start scenario described above.

5.4.2 Phase 2: Unfrozen GateKeeper with Soft Gating

In the second phase, GateKeeper is unfrozen and trained jointly alongside both streams. GateKeeper's supervision target is constructed directly from Stream A's own predictions at this point in training: for each sample, the binary target

$$\hat{g} = \mathbb{1}[\arg\max(z_a) = y] \tag{5.9}$$

is 1 if Stream A's predicted class matches the ground-truth label and 0 otherwise. This is the correctness-supervision strategy motivated in Section 5.3.1: rather than an unsupervised confidence score, GateKeeper is trained against a target that is defined, unambiguously and per-sample, in terms of whether Stream A actually got the case right.

A direct complication arises from introducing GateKeeper's binary decision into an otherwise differentiable training pipeline: a hard threshold on α, routing each sample discretely to either Stream A alone or the fused pathway, is not differentiable, and would block gradient flow back through GateKeeper's own parameters during training, since the gradient of a step function is zero almost everywhere. Phase 2 resolves this by using a soft-gated combination of the two prediction pathways during training, rather than the hard routing decision used at inference,

$$z_{\text{final}} = \alpha z_a + (1 - \alpha) z_{\text{fused}} \tag{5.10}$$

This soft combination is differentiable with respect to α, allowing gradients from the final prediction's loss to flow back into GateKeeper's parameters and shape α directly, something a hard, non-differentiable routing decision could not support during training.

5.4.2.1 Loss Formulation

The full Phase 2 objective combines four unweighted terms,

$$\mathcal{L}_{\text{phase2}} = \mathcal{L}_{\text{CE}}(z_{\text{final}}, y) + \mathcal{L}_{\text{CE}}(z_a, y) + \mathcal{L}_{\text{CE}}(z_{\text{fused}}, y) + \text{BCE}(\alpha, \hat{g}) \tag{5.11}$$

The first term supervises the soft-gated final prediction directly, the quantity that most closely mirrors what the system actually outputs. The second and third terms are auxiliary losses retained from Phase 1, continuing to supervise Stream A and the fused pathway directly against ground truth even as GateKeeper's gating begins to influence the final combined prediction. These two auxiliary terms serve a specific and necessary purpose, which is worth tracing through mechanically. Consider what happens to the gradient reaching Stream A's parameters through z_{final} alone, without the auxiliary term $\mathcal{L}_{\text{CE}}(z_a, y)$: by Equation 5.10, $\partial z_{\text{final}}/\partial z_a = \alpha$, so as $\alpha \to 0$, meaning GateKeeper has come to believe Stream A is almost certainly wrong on some region of the input space, the gradient of $\mathcal{L}_{\text{CE}}(z_{\text{final}}, y)$ with respect to z_a shrinks toward zero in proportion. Stream A would then receive vanishingly little training signal precisely on the cases GateKeeper has flagged as difficult, which are exactly the cases where Stream A most needs continued refinement rather than abandonment. The symmetric argument applies to z_{fused} as $\alpha \to 1$: on cases GateKeeper judges Stream A already handles well, the fused pathway's gradient through z_{final} alone would shrink toward zero, leaving the fusion module and Stream B undertrained specifically on the easier majority of cases where they are rarely invoked at inference in the first place, and therefore have comparatively few other opportunities to receive gradient signal at all. The auxiliary terms $\mathcal{L}_{\text{CE}}(z_a, y)$ and $\mathcal{L}_{\text{CE}}(z_{\text{fused}}, y)$ route a gradient to each branch that does not pass through α at all, and is therefore unaffected by whichever direction α happens to be saturating toward, ensuring both Stream A and the fused pathway continue learning throughout Phase 2 regardless of how GateKeeper's activation pattern evolves over the course of training. The fourth term is GateKeeper's own binary cross-entropy loss against the correctness target $\hat{g}$ defined in Equation 5.9, the term that actually shapes GateKeeper's gating behavior.

All four terms are combined without any relative weighting, an explicit design choice made to avoid introducing additional tunable hyperparameters into an already multi-component training objective. No separate efficiency-penalty term, of the kind used in some conditional-computation architectures to explicitly discourage over-activation of the more expensive pathway, appears anywhere in this objective. This omission is deliberate and follows directly from the correctness-supervision argument made in Section 5.3.1: because $\hat{g}$ is defined by actual correctness, a case where Stream A is already correct is by definition a case where the target $\hat{g}$ is 1, meaning GateKeeper's own loss already pushes α toward 1 and therefore toward not activating Stream B, on precisely the cases where activating it would add cost without adding diagnostic value. Low activation on the easy majority of cases is therefore already an implication of the correctness-supervision objective itself, not something that needs to be separately enforced through an additional penalty term.

5.4.3 Inference: Hard Gating

The soft combination described by Equation 5.10 is a training-time device only, needed specifically to keep GateKeeper's gradient path differentiable during Phase 2. At inference time, this soft combination is replaced entirely by the hard gating decision from Equation 5.7, which gives SecondOpinion its conditional-computation property. For each new sample, Stream A and GateKeeper are always evaluated; Stream B and the fusion module are evaluated only if $\alpha < 0.5$, and the final prediction is

$$z_{\text{predicted}} = \begin{cases} z_a & \text{if } \alpha \geq 0.5 \\ z_{\text{fused}} & \text{if } \alpha < 0.5 \end{cases} \tag{5.12}$$

This is the behavior illustrated in Figure 5.1 and the behavior all inference-time results reported in Chapter 6 reflect: the soft, differentiable combination used during Phase 2 training exists solely to make GateKeeper trainable, and plays no role in how SecondOpinion actually behaves once deployed.

Both phases use the AdamW optimizer with cosine annealing learning rate scheduling, a batch size of 16, and a learning rate of 1×10^{-4} held constant across both phases and both tasks. The number of training epochs per phase differs by task, reflecting the difference in dataset scale between the CXR and pelvic fracture datasets described in Chapter 4: the CXR dataset, at 21,865 images, is trained for 25 epochs in Phase 1 and 35 epochs in Phase 2, while the substantially smaller pelvic fracture dataset, at 228 images in its Visible Fracture subset, is trained for 15 epochs in Phase 1 and 25 epochs in Phase 2. This inverse relationship, fewer epochs allotted to the smaller dataset rather than more, follows the standard reasoning that a smaller dataset reaches a comparable number of gradient updates in fewer epochs for a fixed batch size, and remains considerably more prone to overfitting under extended training than a dataset roughly two orders of magnitude larger; the shorter pelvic schedule is accordingly a safeguard against overfitting on a comparatively small cohort rather than a sign of a less thorough training process. All other optimization settings are held constant across both tasks, so that any performance difference reported in Chapter 6 between the CXR and pelvic fracture results reflects genuine differences in task difficulty and dataset composition rather than differences in how each task's model was optimized. All models in this research, including both streams, the segmentation model, and GateKeeper, were trained on a single NVIDIA A100 GPU accessed through Google Colab.

5.5 Evaluation Framework

Evaluating SecondOpinion requires more than the standard set of classification metrics used to assess a conventional diagnostic model, since its central claim is not simply that it classifies accurately, but that it does so while activating its more expensive anatomy-guided pathway only when needed. This section defines both halves of that evaluation: the classification metrics used to assess diagnostic performance, applied identically to SecondOpinion and to every baseline it is compared against in Chapter 6, and the activation-

rate accounting used specifically to assess whether GateKeeper's conditional behavior actually tracks task difficulty in the way this research argues it should. Every metric is stated formally below; the results themselves are reported only in Chapter 6, and this section defines exclusively how those results are to be measured and interpreted.

5.5.1 Classification Metrics

SecondOpinion and every model it is compared against in Chapter 6 are evaluated using the same six metrics: Accuracy, Precision, Recall, Specificity, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUROC). All six are computed per fold and then averaged across five-fold cross-validation, consistent with the evaluation protocol described for both datasets in Chapter 4, together with 95% confidence intervals, giving a sense of how consistent each reported metric is across folds rather than reporting only a single point estimate that could be sensitive to how a particular fold happened to be composed.

For a given fold, let TP, TN, FP, and FN denote the number of true positive, true negative, false positive, and false negative predictions respectively, using the standard convention of treating the positive class as the diagnostically relevant class for a given task, disease presence for the CXR classes and fracture for the pelvic tasks. Accuracy is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.13}$$

Precision, the fraction of predicted-positive cases that are genuinely positive, is defined as

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5.14}$$

Recall, also referred to as sensitivity, the fraction of genuinely positive cases correctly identified as such, is defined as

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5.15}$$

Specificity, the fraction of genuinely negative cases correctly identified as such, is defined as

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5.16}$$

F1 score is computed slightly differently between the two tasks, following an established convention rather than an inconsistency introduced by this research. For the CXR

classification task, F1 is computed in its standard form, as the harmonic mean of Precision and Recall,

$$\mathrm{F1}_{\mathrm{CXR}} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{5.17}$$

For the pelvic fracture detection task, F1 is instead computed as the harmonic mean of Recall and Specificity,

$$\mathrm{F1}_{\mathrm{pelvic}} = 2 \cdot \frac{\mathrm{Recall} \cdot \mathrm{Specificity}}{\mathrm{Recall} + \mathrm{Specificity}} \tag{5.18}$$

This departure from the standard definition in Equation 5.17 is deliberate and is adopted specifically to match the convention already established by PelFANet and by the DRR20 and ImageNet+DRR20 baselines it was compared against [103], ensuring that any F1 comparison drawn in Chapter 6 between SecondOpinion and this research's own prior pelvic work, or the external baselines PelFANet itself was benchmarked against, is measuring the same underlying quantity rather than two differently defined scores that happen to share a name.

AUROC summarizes classification performance across every possible decision threshold, rather than at a single fixed operating point the way Accuracy, Precision, Recall, and Specificity each implicitly do. At a given threshold τ, the classifier's True Positive Rate, equivalent to Recall, and False Positive Rate, equal to $1 - \mathrm{Specificity}$, are computed as defined above; sweeping τ across its full range traces out the Receiver Operating Characteristic (ROC) curve, and AUROC is the area beneath it,

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\, d(\mathrm{FPR}) \tag{5.19}$$

interpretable as the probability that a randomly chosen positive case is assigned a higher predicted probability of being positive than a randomly chosen negative case, making it a threshold-independent measure of how well a model separates the two classes overall, in contrast to the four preceding metrics, each of which reflects performance at whatever single threshold, typically 0.5, was used to binarize the model's output.

Given a metric value $\hat{m}$ computed as described above, its 95% confidence interval across the $k = 5$ folds of cross-validation is computed using the standard z-approximation for a mean of independent estimates,

$$\mathrm{CI}_{95\%} = \hat{m} \pm 1.96 \cdot \frac{s}{\sqrt{k}} \tag{5.20}$$

where s is the sample standard deviation of the metric's per-fold values, and 1.96 is the critical value of the standard normal distribution corresponding to a 95% confidence level. This interval is reported alongside every metric's point estimate throughout Chapter 6, for both SecondOpinion and every baseline it is compared against.

Beyond these six standard classification metrics, Params and FLOPs are also reported, both as static per-path values, the fixed cost of evaluating Stream A alone, Stream B alone, or the full fused pathway, and as activation-weighted averages that account for how often each path is actually invoked at inference. Given the activation rate ρ defined in Section 5.5.2, the activation-weighted FLOPs of the full SecondOpinion pipeline are

$$\text{FLOPs}_{\text{weighted}} = \text{FLOPs}_{\text{A}} + \rho \cdot \text{FLOPs}_{\text{B+fusion}} \tag{5.21}$$

where FLOPs_{A} is the fixed cost incurred on every sample by Stream A and GateKeeper together, and $\text{FLOPs}_{\text{B+fusion}}$ is the additional cost incurred only when Stream B and the fusion module are activated. This activation-weighted accounting is what allows SecondOpinion's actual computational cost, as opposed to its theoretical worst-case cost if Stream B were always activated, to be reported and compared meaningfully against an always-on dual-stream baseline, for which $\rho = 1$ by construction and Equation 5.21 reduces to the simple, unconditional sum of both streams' costs. Static Params and FLOPs figures for the segmentation model itself are reported separately from the classification pipeline's own cost, since the segmentation stage runs once, offline, prior to classification, rather than being repeated at inference for every classification query.

5.5.2 GateKeeper Activation Rate

Activation rate, denoted ρ, is defined for a given fold of size n as the fraction of samples in that fold for which GateKeeper's output satisfies $\alpha < 0.5$, and Stream B and the fusion module are consequently evaluated,

$$\rho = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[\alpha_i < 0.5] \tag{5.22}$$

This is reported as a single aggregate percentage per task in Chapter 6, computed as an average of Equation 5.22 over the same five-fold cross-validation protocol used for the classification metrics above, rather than as a raw count specific to any one fold, so that the reported activation rate reflects a stable property of GateKeeper's behavior on a given task rather than an artifact of a particular fold's composition.

Activation rate serves a purpose beyond simply quantifying computational savings, and beyond its direct role in Equation 5.21. Because this research's central architectural claim, established across Chapter 1 and revisited throughout Chapter 3, is that pelvic fracture detection sits at a harder point on a diagnostic difficulty gradient than CXR classification, with invisible fracture cases harder still, activation rate offers a direct, quantitative test of whether GateKeeper's learned gating behavior actually tracks that same difficulty gradient rather than activating Stream B at some roughly uniform rate regardless of task. A gating mechanism trained purely for correctness prediction, as GateKeeper is, with no explicit awareness of task identity or difficulty built into its objective, would be expected under this research's central hypothesis to activate Stream B more often precisely on the harder tasks, since Stream A's own unaided accuracy should be lower there, and

consequently the correctness target $\hat{g}$ defined in Equation 5.9 should be zero for a larger share of cases, directly increasing ρ by Equation 5.22. Chapter 6 examines whether the activation rates observed across the CXR, visible fracture, and invisible fracture tasks bear out this expectation.

Activation rate is also read alongside the Params and FLOPs accounting described in Section 5.5.1, since the two are directly linked by construction through Equation 5.21: a higher activation rate on a harder task necessarily corresponds to a higher activation-weighted average cost for that task, moving SecondOpinion's realized computational cost closer to the always-on ceiling as task difficulty increases. This coupling is treated in Chapter 6 not as an incidental byproduct but as further evidence bearing on the same underlying question examined through activation rate alone, whether GateKeeper's conditional behavior is allocating the more expensive anatomical reasoning pathway specifically to the cases that need it, rather than either under-activating on hard cases or over-activating on easy ones.

Taken together, the four components described across this chapter, the dual-stream architecture, GateKeeper, the two-phase training strategy, and the evaluation framework built around activation-weighted cost and activation rate, form a single coherent answer to the gap Chapter 3 closed with. Where every prior architecture surveyed in that chapter applied its most expensive reasoning unconditionally, SecondOpinion applies it only when GateKeeper's correctness-supervised judgment indicates a given case actually needs it, and every design choice in this chapter serves that one distinction. The dual-stream architecture keeps Stream B's anatomical reasoning genuinely separable from Stream A's fast primary path, rather than entangling the two so that separating them at inference time would be architecturally impossible. GateKeeper is trained against ground-truth correctness rather than an unsupervised confidence proxy specifically because a gate this report's later chapters intend to interpret as tracking real diagnostic difficulty needs to be trained against a signal that actually reflects that difficulty, not one that merely correlates with it under favorable conditions. The two-phase training strategy exists to prevent GateKeeper and the fused pathway from being optimized against a moving target, learning to fuse before there is a stable signal worth fusing on. And the evaluation framework, in particular the activation-weighted FLOPs accounting and the activation rate metric itself, exists so that the central claims this report makes in Chapters 6 and 7, that conditional gating recovers most of always-on performance at a fraction of the cost, and that its activation behavior tracks genuine task difficulty, can be measured directly rather than argued for qualitatively. Chapter 6 reports what these four components produce when trained and evaluated across the three diagnostic tasks introduced in Chapter 4.

Chapter-6: Result

Chapter 5 described SecondOpinion's architecture, its two-phase training procedure, and the evaluation framework used to assess it. This chapter reports the results of that evaluation across both tasks introduced in Chapter 4: the unified five-class chest X-ray dataset, and the two subsets of the AMERI pelvic fracture dataset, Visible Fracture (VIS) and Invisible Fracture (INVIS). Results are reported using the classification metrics, confidence interval convention, Params/FLOPs accounting, and activation rate definition formalized in Section 5.5, applied consistently across every configuration and every task.

This chapter is organized as follows. Section 6.1 gives an overview of the four internal configurations evaluated throughout the chapter and clarifies how they relate to one another and to the external baselines each task is compared against. Section 6.2 reports CXR classification results. Section 6.3 reports pelvic fracture detection results, treating the VIS and INVIS subsets separately given their different roles in training and evaluation. Section 6.4 reports the efficiency and activation-rate analysis central to this research's conditional-computation claim. Section 6.5 presents a qualitative GradCAM analysis of two representative cases. Interpretation of these results, their relationship to the prior approaches traced in Chapter 3, and the limitations of the present evaluation are addressed separately in Chapter 7; this chapter reports what was measured, without argument as to what it means.

6.1 Overview

Four internal configurations are evaluated across every task in this chapter, each isolating a different component of SecondOpinion's architecture. **Stream A only** evaluates Stream A's classification head in isolation, with Stream B, GateKeeper, and the fusion module entirely absent from the forward pass; this configuration is architecturally equivalent to a conventional single-stream baseline operating on the raw image alone, and serves as the lower bound against which the value of anatomical guidance can be measured. **Stream B only** evaluates Stream B's classification head in isolation on the anatomy-guided input alone, with Stream A absent; because Stream A only and Stream B only share an identical architecture and differ solely in the input they receive, raw image versus anatomy-guided representation, any performance gap between them isolates the value of anatomical guidance as an input choice, independent of any gating or fusion behavior. **Always-On** evaluates the full dual-stream architecture with GateKeeper's gating mechanism disabled entirely: both streams and the fusion module are evaluated unconditionally on every sample, exactly as a conventional, always-on dual-stream design of the kind traced through Chapter

3, PelFANet included, would operate. Always-On therefore represents the performance ceiling SecondOpinion is measured against, the best result the architecture can produce if computational cost is not a constraint. **SecondOpinion** is the full system as described in Chapter 5, with GateKeeper active and hard gating applied at inference exactly as defined in Equation 5.12.

These four configurations are reported together in a single table for each task, following the convention already established in the SecondOpinion paper and consistent with how comparable ablation tables are presented in Chapter 3. Beyond these four internal configurations, each task is additionally compared against a set of external baselines specific to that task, drawn either from the general medical imaging literature or, for the pelvic tasks, directly from the prior approaches traced in Chapter 3. These external comparisons are reported in a separate subsection from the internal ablation results in every section below, keeping the question of how SecondOpinion's own components contribute to its performance analytically distinct from the question of how SecondOpinion compares against previously published work.

6.2 CXR Classification Results

6.2.1 Ablation Results

Table 6.1 reports Accuracy, Precision, Recall, Specificity, F1, and AUROC for all four internal configurations on the unified five-class CXR dataset, averaged over five-fold cross-validation with 95% confidence intervals computed according to Equation 5.20.

Table 6.1: CXR classification results across internal configurations. Mean over 5-fold CV; 95% CI shown in the row below each result.

Configuration	Acc.	Prec.	Rec.	Spec.	F1	AUROC
Stream A only	93.25%	95.02%	93.56%	95.01%	0.943	0.981
	±0.32%	±0.40%	±0.53%	±0.52%	±0.0044	±0.0061
Stream B only	94.06%	95.53%	94.28%	95.56%	0.949	0.983
	±0.19%	±0.22%	±0.61%	±0.58%	±0.0040	±0.0042
Always-On	**99.05%**	**98.56%**	**98.32%**	**99.42%**	**0.984**	**0.993**
	±0.06%	±0.16%	±0.24%	±0.02%	±0.0018	±0.0054
SecondOpinion	98.41%	97.41%	97.13%	98.91%	0.973	0.990
	±0.07%	±0.26%	±0.08%	±0.02%	±0.0013	±0.0046

Stream B only outperforms Stream A only across every one of the six metrics reported, with the largest margins appearing in Specificity (95.56% against 95.01%) and Precision (95.53% against 95.02%). Always-On achieves the highest value on every metric among all four configurations, reaching 99.05% Accuracy and 0.993 AUROC. SecondOpinion reaches 98.41% Accuracy and 0.990 AUROC, placing it above both single-stream configurations on every metric and below Always-On on every metric, with the gap to Always-On ranging from 0.14 percentage points on Specificity to 1.19 percentage points on Recall.

6.2.2 Comparison Against External Baselines

Table 6.2 reports the same six metrics for three external baselines drawn from the CXR architecture progression traced in Chapter 3, alongside SecondOpinion and Always-On repeated from Table 6.1 for direct comparison.

Table 6.2: CXR classification results against external baselines. Mean over 5-fold CV; 95% CI shown in the row below each of SecondOpinion's and Always-On's results.

Model	Acc.	Prec.	Rec.	Spec.	F1	AUROC
EfficientNetB3 [98]	93.35%	95.99%	93.81%	95.11%	0.949	0.987
CheXNet [98]	94.12%	94.76%	93.05%	95.81%	0.939	0.988
CheXNet-CBAM [106]	96.21%	96.14%	94.67%	97.55%	0.954	0.989
Always-On	**99.05%**	**98.56%**	**98.32%**	**99.42%**	**0.984**	**0.993**
	±0.06%	±0.16%	±0.24%	±0.02%	±0.0018	±0.0054
SecondOpinion	98.41%	97.41%	97.13%	98.91%	0.973	0.990
	±0.07%	±0.26%	±0.08%	±0.02%	±0.0013	±0.0046

Among the three external baselines, CheXNet-CBAM records the highest value on every metric, reaching 96.21% Accuracy and 0.989 AUROC. SecondOpinion's Accuracy, Precision, Recall, Specificity, F1, and AUROC all exceed CheXNet-CBAM's corresponding values, with the AUROC margin measuring 0.001 and the Accuracy margin measuring 2.20 percentage points. SecondOpinion's activation rate on this task, and the computational cost associated with it, are reported separately in Section 6.4.

6.3 Pelvic Fracture Detection Results

6.3.1 Visible Fracture (VIS) Results

The VIS subset of the AMERI dataset is used for both training and evaluation of all four internal configurations and all pelvic external baselines, under stratified five-fold cross-validation as described in Chapter 4. F1 for every result in this section is computed according to Equation 5.18, the harmonic mean of Recall and Specificity, following the convention established in Section 5.5.1.

6.3.1.1 Ablation Results

Table 6.3 reports the four internal configurations on the VIS subset.

Stream B only outperforms Stream A only on every metric except Accuracy, where the two configurations differ by only 0.20 percentage points. Always-On records the highest value on five of the six metrics, with SecondOpinion recording the highest Recall relative to Always-On being the sole exception, where Always-On's 89.80% exceeds SecondOpinion's

Table 6.3: Visible fracture (VIS) results across internal configurations. Mean over 5-fold CV; 95% CI shown in the row below each result.

Configuration	Acc.	Prec.	Rec.	Spec.	F1	AUROC
Stream A only	86.40%	85.90%	84.00%	73.30%	0.763	0.867
	±3.8%	±4.3%	±9.0%	±20.3%	±0.121	±0.020
Stream B only	86.60%	86.20%	85.80%	76.60%	0.806	0.871
	±3.6%	±4.2%	±7.8%	±9.6%	±0.071	±0.008
Always-On	**92.60%**	**93.90%**	89.80%	**90.70%**	**0.901**	**0.951**
	±2.7%	±0.9%	±1.9%	±6.4%	±0.027	±0.006
SecondOpinion	91.40%	93.00%	87.90%	88.30%	0.879	0.949
	±3.5%	±1.5%	±2.9%	±6.5%	±0.022	±0.011

87.90%. SecondOpinion's F1, 0.879, is 2.2 points below Always-On's 0.901, and its AUROC, 0.949, is 0.002 below Always-On's 0.951.

6.3.1.2 Comparison Against External Baselines

Table 6.4 reports three external pelvic fracture detection baselines drawn from Chapter 3, alongside SecondOpinion and Always-On repeated for comparison. ImageNet+DRR20, DRR20, and PelFANet report only F1 and AUROC in their original evaluations; Accuracy, Precision, Recall, and Specificity are accordingly left blank for these three rows, consistent with how they were originally reported.

Table 6.4: Visible fracture (VIS) results against external baselines. Mean over 5-fold CV; 95% CI shown in the row below each of SecondOpinion's and Always-On's results.

Method	Acc.	Prec.	Rec.	Spec.	F1	AUROC
ImageNet+DRR20 [55]	–	–	–	–	0.8390	0.9280
DRR20 [55]	–	–	–	–	0.8520	0.9290
PelFANet [103]	88.68%	92.49%	92.21%	78.33%	0.8471	0.9334
Always-On	**92.60%**	**93.90%**	**89.80%**	**90.70%**	**0.901**	**0.951**
	±2.7%	±0.9%	±1.9%	±6.4%	±0.027	±0.006
SecondOpinion	91.40%	93.00%	87.90%	88.30%	0.879	0.949
	±3.5%	±1.5%	±2.9%	±6.5%	±0.022	±0.011

PelFANet records the highest Recall among all rows in Table 6.4, 92.21%, exceeding both Always-On's 89.80% and SecondOpinion's 87.90%. SecondOpinion's F1, 0.879, exceeds PelFANet's 0.8471, DRR20's 0.8520, and ImageNet+DRR20's 0.8390. SecondOpinion's AUROC, 0.949, exceeds PelFANet's 0.9334, DRR20's 0.9290, and ImageNet+DRR20's 0.9280, and sits 0.002 below Always-On's 0.951.

6.3.2 Invisible Fracture (INVIS) Results

The INVIS subset, 35 cases confirmed via 3D-CT but showing no visible sign of fracture on the X-ray itself, is used exclusively for evaluation, as described in Chapter 4. None of the four internal configurations, nor PelFANet among the external baselines, are trained on any INVIS case; every result reported in this subsection reflects generalization from VIS-trained models applied to a held-out, harder subset. F1 for every result in this section is again computed according to Equation 5.18.

6.3.2.1 Ablation Results

Table 6.5 reports the four internal configurations on the INVIS subset.

Table 6.5: Invisible fracture (INVIS) results across internal configurations, evaluated as a held-out generalization set. Mean over 5-fold CV; 95% CI shown in the row below each result.

Configuration	Acc.	Prec.	Rec.	Spec.	F1	AUROC
Stream A only	76.60%	77.90%	82.10%	71.60%	0.756	0.744
	±4.1%	±6.8%	±3.8%	±13.2%	±0.063	±0.050
Stream B only	77.00%	80.10%	82.10%	72.60%	0.763	0.752
	±3.7%	±6.8%	±3.8%	±12.3%	±0.057	±0.036
Always-On	**85.40%**	**89.90%**	83.50%	**85.30%**	**0.842**	**0.901**
	±0.8%	±1.1%	±5.1%	±3.0%	±0.016	±0.020
SecondOpinion	84.60%	89.20%	83.20%	83.30%	0.830	0.898
	±0.9%	±1.7%	±5.1%	±5.1%	±0.026	±0.020

Stream B only exceeds Stream A only on every metric except Recall, where both configurations record an identical 82.10%. Always-On records the highest value on five of the six metrics; Recall is again the exception, with Always-On's 83.50% exceeding SecondOpinion's 83.20% by only 0.30 percentage points, the narrowest margin between the two configurations on any metric in this table. SecondOpinion's F1, 0.830, is 1.2 points below Always-On's 0.842, and its AUROC, 0.898, is 0.003 below Always-On's 0.901.

6.3.2.2 Comparison Against External Baselines

Table 6.6 reports the same three external baselines evaluated on the INVIS subset, alongside SecondOpinion and Always-On.

PelFANet records the highest Recall among all rows in Table 6.6, 84.35%, exceeding Always-On's 83.50% and SecondOpinion's 83.20%. SecondOpinion's F1, 0.830, exceeds PelFANet's 0.8123, DRR20's 0.7860, and ImageNet+DRR20's 0.7210. SecondOpinion's AUROC, 0.898, exceeds PelFANet's 0.8688, DRR20's 0.8002, and ImageNet+DRR20's 0.7140, and sits 0.003 below Always-On's 0.901.

Table 6.6: Invisible fracture (INVIS) results against external baselines, evaluated as a held-out generalization set. Mean over 5-fold CV; 95% CI shown in the row below each of SecondOpinion's and Always-On's results.

Method	Acc.	Prec.	Rec.	Spec.	F1	AUROC
ImageNet+DRR20 [55]	–	–	–	–	0.7210	0.7140
DRR20 [55]	–	–	–	–	0.7860	0.8002
PelFANet [103]	82.29%	88.36%	84.35%	78.33%	0.8123	0.8688
Always-On	**85.40%**	**89.90%**	**83.50%**	**85.30%**	**0.842**	**0.901**
	±0.8%	±1.1%	±5.1%	±3.0%	±0.016	±0.020
SecondOpinion	84.60%	89.20%	83.20%	83.30%	0.830	0.898
	±0.9%	±1.7%	±5.1%	±5.1%	±0.026	±0.020

6.4 Efficiency and Activation Analysis

6.4.1 Params and FLOPs Across Configurations

Table 6.7 reports Params and FLOPs for the Single Stream Only and Always-On configurations, which are architecturally fixed and identical across all three tasks, alongside SecondOpinion's activation-weighted Params and FLOPs, computed according to Equation 5.21, for each of the three tasks individually.

Table 6.7: Parameters and FLOPs per configuration. Single Stream Only and Always-On are architecturally fixed across tasks; SecondOpinion's figures are activation-weighted per task.

Configuration	Params (M)	FLOPs (G)	Activation
Single Stream Only	4.18	0.400	–
Always-On	10.68	0.855	100%
SecondOpinion – CXR	4.78	0.442	9.23%
SecondOpinion – VIS	5.75	0.510	24.12%
SecondOpinion – INVIS	7.15	0.608	45.71%

Single Stream Only, corresponding architecturally to either Stream A only or Stream B only from the ablation tables in Sections 6.2 and 6.3, since both share an identical backbone and differ only in input, records 4.18M Params and 0.400 GFLOPs. Always-On, evaluating both streams and the fusion module on every sample, records 10.68M Params and 0.855 GFLOPs, roughly two and a half times Single Stream Only's Params and just over double its FLOPs, the difference accounting for the added cost of the second backbone, the fusion module, and GateKeeper itself. SecondOpinion's activation-weighted Params and FLOPs increase monotonically from the CXR task (4.78M, 0.442G) through VIS (5.75M, 0.510G) to INVIS (7.15M, 0.608G), tracking the corresponding increase in Activation reported in the same table's rightmost column. Reported costs for all three SecondOpinion rows cover the classification pipeline only, following the convention established in Section 5.5.1, and exclude the upstream segmentation model described in Section 5.2.1, which runs offline prior to inference.

6.4.2 Activation Rate Across the Difficulty Gradient

Activation rate, computed according to Equation 5.22 and averaged across five-fold cross-validation, is 9.23% on the CXR task, 24.12% on the VIS task, and 45.71% on the INVIS task, repeated here from the rightmost column of Table 6.7 for direct reference alongside the accuracy figures below. Table 6.8 reports AUROC for Stream A only and Always-On across all three tasks, alongside the AUROC gap between them, computed as Always-On's AUROC minus Stream A only's AUROC for each task and expressed in points.

Table 6.8: AUROC gap between Stream A only and Always-On across tasks, alongside GateKeeper's activation rate for each task.

Task	Stream A only AUROC	Always-On AUROC	AUROC Gap (pts)	Activation Rate
CXR	0.981	0.993	1.18	9.23%
VIS	0.867	0.951	8.40	24.12%
INVIS	0.744	0.901	15.78	45.71%

The AUROC gap between Stream A only and Always-On increases monotonically across the three tasks, from 1.18 points on CXR to 8.40 points on VIS to 15.78 points on INVIS. Activation rate increases across the same three tasks in the same order, from 9.23% to 24.12% to 45.71%.

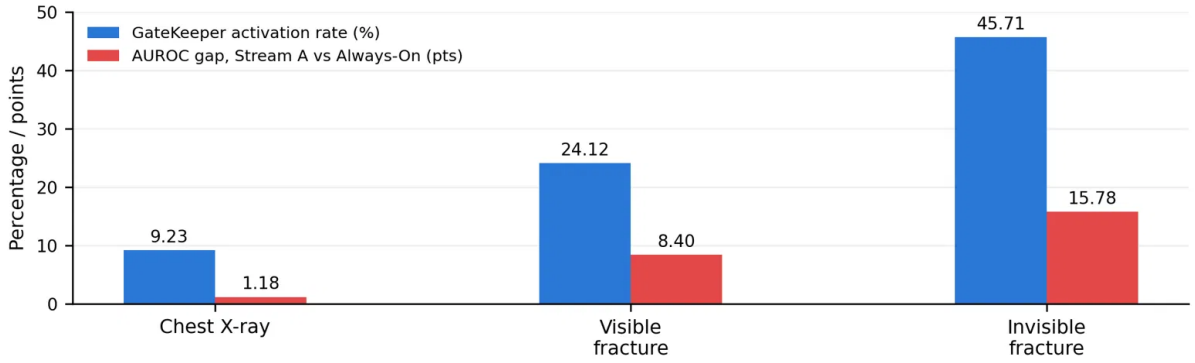

Figure 6.1: GateKeeper's activation rate and the AUROC gap between Stream A only and Always-On, plotted together across the three tasks.

Figure 6.1 plots activation rate and the AUROC gap together across the three tasks. Both quantities are lowest on CXR, at 9.23% and 1.18 points respectively, rise to 24.12% and 8.40 points on VIS, and reach their highest values on INVIS, at 45.71% and 15.78 points. The two quantities increase across the same three tasks in the same order.

6.5 Qualitative Analysis (GradCAM)

Two representative cases, each correctly classified by SecondOpinion, are examined using Gradient-weighted Class Activation Mapping (GradCAM) to visualize which regions of the input most influenced the model's prediction.

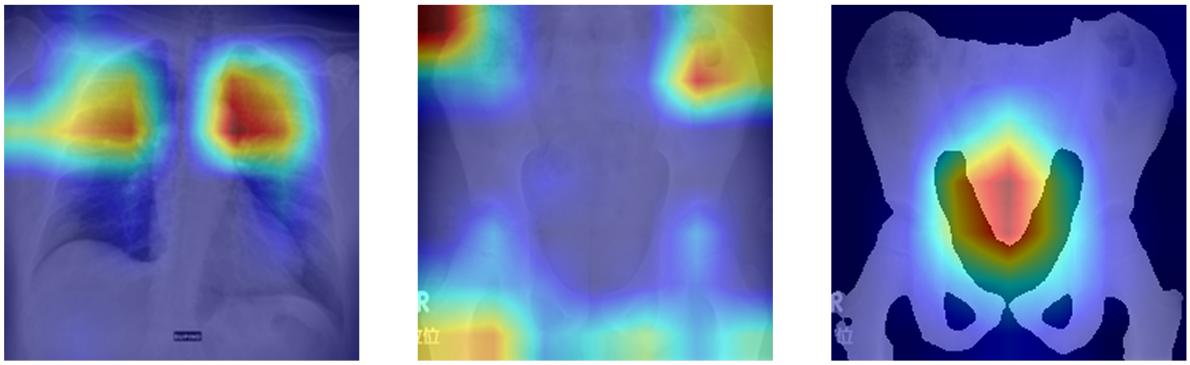

Figure 6.2: GradCAM visualizations for two correctly classified cases. (a) Chest X-ray case, Stream B inactive: Stream A's attention is concentrated on both lung fields. (b) Invisible fracture case, Stream B active: Stream A's attention prior to fusion is concentrated on regions outside the pelvic bone structure. (c) The same invisible fracture case after fusion: attention shifts to the central bone region.

Figure 6.2(a) shows a correctly classified COVID-19 chest X-ray case for which GateKeeper's output α exceeded 0.5, and Stream B was accordingly not activated. Stream A's GradCAM attention in this case is concentrated on both lung fields.

Figure 6.2(b) and (c) show a correctly classified invisible fracture case for which GateKeeper's output α fell below 0.5, activating Stream B and the fusion module. Prior to fusion, shown in (b), Stream A's attention is concentrated on regions outside the pelvic bone structure, in the background corners of the image. After fusion, shown in (c), attention shifts to the central bone region.

Across all three diagnostic tasks introduced in Chapter 4, and every internal configuration and external baseline described in Section 6.1, the results in this chapter follow a consistent shape. On CXR, SecondOpinion's accuracy and AUROC sat within roughly a percentage point of Always-On while activating Stream B on 9.23% of cases, and outperformed every external baseline evaluated, CheXNet-CBAM included. On both pelvic subsets, SecondOpinion and Always-On traded a Recall shortfall against PelFANet for gains in Specificity, F1, and AUROC, a pattern that held on VIS and, with a narrower Recall gap, on INVIS as well. Params, FLOPs, and activation rate moved together and in the same order across all three tasks, rising from CXR through VIS to INVIS, and the AUROC gap between Stream A only and Always-On rose in that same order alongside them. The GradCAM analysis in Section 6.5 illustrated this same pattern at the level of two individual cases, one where Stream B stayed inactive and one where it activated and visibly redirected attention toward the pelvic bone structure after fusion.

These are the measurements this work rests its three research questions on. Chapter 7 takes up the interpretation this chapter has deliberately withheld, situating each of the results reported above against RQ1 through RQ3, against the prior approaches traced in Chapter 3, and against the specific limitations of the evaluation this chapter reports.

Chapter-7: Discussion

Chapter 6 reported what SecondOpinion measured against; this chapter interprets what those measurements mean, situates them against the prior approaches traced in Chapters 2 and 3, and states plainly where the present evaluation falls short of a complete answer. Section 7.1 briefly frames the shift from reporting to interpretation. Section 7.2 works through the main quantitative results task by task. Section 7.3 turns specifically to GateKeeper's activation behavior and what it does and does not demonstrate about the correctness-supervision design argued for in Chapter 5. Section 7.4 discusses the two GradCAM cases from Chapter 6 with appropriate restraint given their number. Section 7.5 situates SecondOpinion against the prior approaches this research has built on and departed from. Section 7.6 draws out the broader implications of the framework beyond the two tasks it was tested on. Section 7.7 states the limitations of the present work directly, distinguishing limitations this research inherits from PelFANet's own stated scope from limitations specific to SecondOpinion's own design and evaluation.

7.1 Overview

Chapter 6 was deliberately restricted to description: what each configuration scored on each metric, how activation rate varied across tasks, what a GradCAM overlay showed for two specific cases. That restriction was useful for keeping the results themselves uncontaminated by argument, but a set of numbers does not interpret itself. This chapter asks the questions Chapter 6 set aside.

- Does the gap between SecondOpinion and Always-On represent an acceptable cost for the computation it avoids, or a real deficiency in what SecondOpinion accomplishes on the cases it does route through Stream B alone?
- Does the consistent Recall shortfall against PelFANet across both pelvic subsets undermine the case for conditional gating, or is it a separate architectural trade-off that has little to do with gating at all?
- Does the fact that activation rate rises in step with task difficulty actually demonstrate that GateKeeper's correctness-supervision objective works as intended, or is it equally consistent with a simpler, less interesting explanation?

These are the questions this chapter works through, returning repeatedly to the four

objectives stated in Chapter 1 as the standard against which SecondOpinion's results should be judged.

7.2 Interpretation of Main Results

7.2.1 The Accuracy-Efficiency Trade-off (CXR)

On the CXR task, SecondOpinion reaches 98.41% Accuracy and 0.990 AUROC while activating Stream B on only 9.23% of cases, against Always-On's 99.05% Accuracy and 0.993 AUROC with Stream B activated unconditionally. The gap on every metric in Table 6.1 is small in absolute terms, the widest being 1.19 percentage points on Recall, and this smallness is worth taking seriously rather than treating as an incidental footnote. SecondOpinion is not being asked to match Always-On's performance while paying Always-On's cost; it is being asked to approximate that performance while paying a small fraction of it, since the great majority of CXR cases never reach Stream B at all. Read this way, the 0.64 percentage point Accuracy gap and the 0.003 AUROC gap are not a shortfall to be explained away, but the explicit price of the computational saving GateKeeper is designed to produce, and a price small enough that it does not obviously outweigh the saving it buys.

What makes this trade-off more than merely acceptable, though, is where the gap is concentrated. If GateKeeper were gating on some signal only loosely related to actual case difficulty, the natural expectation would be a meaningfully larger accuracy gap, since a substantial number of genuinely hard cases would slip through ungated and be misclassified by Stream A alone. That the gap stays this narrow at a 9.23% activation rate is itself indirect evidence, developed more fully in Section 7.3, that GateKeeper is not gating arbitrarily but is concentrating its limited activations specifically on the cases where Stream A's unaided judgment is least trustworthy. The alternative interpretation, that CXR classification is simply an easy enough task that almost any gating policy would produce a similarly narrow gap, cannot be ruled out from the CXR results in isolation; it is precisely for this reason that the pelvic results, and the cross-task comparison in Section 7.3.2, matter more than the CXR results do on their own for establishing whether GateKeeper's gating is doing real diagnostic work rather than coasting on an easy task.

Against the three external CXR baselines in Table 6.2, SecondOpinion exceeds CheXNet-CBAM, the strongest of the three, on every one of the six reported metrics, while activating its more expensive pathway on under a tenth of cases. This comparison should be read carefully rather than as a claim that SecondOpinion is simply a better classifier than CheXNet-CBAM in some general sense. CheXNet-CBAM is a single-stream architecture with no access to anatomical guidance at all; SecondOpinion's advantage over it plausibly reflects the value of anatomical guidance on the minority of cases where it is invoked, compounding on top of a primary stream that is already competitive on its own. The comparison is nonetheless informative: it establishes that conditional anatomical guidance, applied selectively, is not merely a cheaper alternative to an unconditional attention mechanism like CBAM, but one capable of exceeding it while incurring its added cost on a small minority of inputs rather than on every one.

7.2.2 The Recall Trade-off (Pelvic Tasks)

The pelvic results in Tables 6.4 and 6.6 present a more complicated picture than the CXR results, and it is worth being direct about the part of that picture that does not favor SecondOpinion. On both VIS and INVIS, PelFANet records the highest Recall among all baselines and configurations compared, 92.21% on VIS against SecondOpinion's 87.90%, and 84.35% on INVIS against SecondOpinion's 83.20%. This is not a marginal difference on VIS specifically; a 4.31 percentage point Recall gap means SecondOpinion is missing a meaningfully larger share of genuine fracture cases than PelFANet does on the same evaluation subset.

What tempers this shortfall, without erasing it, is that it is not isolated to SecondOpinion. Always-On, the unconditional dual-stream configuration built from exactly the same two backbones and the same cross-attention fusion mechanism SecondOpinion uses when Stream B is activated, trails PelFANet on Recall by a comparable margin on both subsets, 89.80% against 92.21% on VIS and 83.50% against 84.35% on INVIS. Since Always-On has GateKeeper disabled entirely and evaluates both streams on every case exactly as PelFANet does, this rules out gating as the source of the Recall gap: whatever is producing lower Recall relative to PelFANet is present in the fusion architecture itself, independent of whether that fusion is applied conditionally or unconditionally. The comparison instead points toward the two architectures' differing fusion strategies, cross-attention fusion applied once, late, in SecondOpinion and Always-On, against PelFANet's iterative CBAM-based exchange across eight FABlocks distributed throughout the network, as the more plausible source of the difference, though establishing this with confidence would require an ablation this research does not report, and is noted as such in Section 7.7.

The other side of this trade-off is where SecondOpinion and Always-On both pull ahead. On both VIS and INVIS, SecondOpinion's Specificity, F1, and AUROC all exceed PelFANet's corresponding values, and by a wider margin than the Recall gap runs in the opposite direction. SecondOpinion trails PelFANet's Recall by 4.31 points on VIS but exceeds its Specificity by 10.0 points, 88.30% against 78.33%; a similar pattern holds on INVIS, where SecondOpinion trails Recall by 1.15 points but exceeds Specificity by 5.0 points. This is a real trade-off rather than a straightforward improvement, and it is worth naming what that trade-off means clinically before moving on: PelFANet is comparatively more willing to flag ambiguous cases as fractures, catching more true fractures at the cost of more false alarms on normal cases, while SecondOpinion and Always-On are comparatively more conservative, missing a somewhat larger share of true fractures in exchange for substantially fewer false positives on cases that are actually normal. Whether this is a favorable trade depends on the clinical context in which either model is deployed and the relative cost of a missed fracture against an unnecessary follow-up, a question this research does not attempt to adjudicate; what can be said is that F1, computed in this pelvic context as the harmonic mean of Recall and Specificity per Equation 5.24, rewards exactly this kind of balance between the two, and it is on this balanced metric, not on Recall alone, that SecondOpinion and Always-On outperform PelFANet on both subsets.

7.2.3 Generalization to Invisible Fractures

The INVIS results carry a different evidentiary weight than the CXR or VIS results, because of what they are testing. No configuration evaluated in Chapter 6, and PelFANet before it, is trained on a single invisible fracture case; every metric reported on INVIS reflects a model trained exclusively on visible fractures being applied to a subset explicitly curated to lack the visual signs that training was based on. A model that performs well here is not demonstrating that it has memorized or overfit to some pattern specific to the INVIS subset, since it never saw that subset during training; it is demonstrating that whatever it learned from visible fracture cases transfers to a genuinely different presentation of the same underlying pathology.

Read against this framing, two results from Chapter 6 stand out together. First, the Recall gap between SecondOpinion and Always-On narrows considerably on INVIS relative to VIS, from 1.90 percentage points on VIS to 0.30 percentage points on INVIS, the narrowest margin between the two configurations on any metric reported for either pelvic subset. Second, and more strikingly, the AUROC gap between Stream A only and Always-On, the gap that activation rate is meant to track as described in Section 7.3.2, reaches its largest value of the three tasks specifically on INVIS, 15.78 points, against 8.40 on VIS and 1.18 on CXR. Taken together, these two observations describe a task where Stream A's unaided judgment degrades the most severely of the three tasks studied, where anatomical guidance consequently matters the most, and where SecondOpinion's conditional gating responds by activating Stream B on the largest share of cases of any task evaluated, 45.71%, nearly the majority of the subset. The narrowing Recall gap against Always-On on this specific subset is consistent with GateKeeper activating Stream B often enough, and on close enough to the right cases, that SecondOpinion's behavior on INVIS converges toward Always-On's more closely than it does on either of the two easier tasks, even though SecondOpinion is still incurring less than half the Always-On cost on this subset by Table 6.7's Params and FLOPs accounting.

This should not be overstated. A 35-case subset, discussed further as a limitation in Section 7.7.1, does not license strong claims about how reliably this convergence would hold on a larger or more varied invisible fracture cohort. What can be said with more confidence is narrower and still meaningful: on the one task in this research's evaluation where the gap between an anatomy-blind and an anatomy-guided prediction is largest, GateKeeper's conditional gating narrows the gap between itself and the always-on ceiling relative to how that same comparison behaves on easier tasks, which is exactly the behavior the difficulty-gradient argument in Chapter 1 and Section 5.3.1's correctness-supervision rationale would predict, without that behavior being separately engineered into GateKeeper's training objective at any point.

7.3 GateKeeper and the Difficulty Gradient

7.3.1 Activation Rate as Evidence for Correctness-Supervised Gating

Section 5.3.1 argued for training GateKeeper as a binary correctness classifier rather than relying on Stream A's own softmax confidence, on the grounds that softmax confidence is a poorly calibrated proxy for actual correctness and would risk failing to activate Stream B precisely on the cases where Stream A is confidently wrong. That argument was made on principle, before any results existed to test it against. Chapter 6's activation-rate results give the argument its first empirical check.

If GateKeeper's correctness-supervision objective were failing to track genuine case difficulty, in the way an unsupervised confidence-based gate risks doing, there is no architectural reason the resulting activation rate should bear any particular relationship to how difficult a task actually is; a poorly calibrated gate could just as easily activate at a roughly constant rate across tasks of different difficulty, or activate more on the easier task than the harder one, or fluctuate without any discernible pattern from one task to the next. What is observed instead is a monotonic increase across exactly the ordering this research's own difficulty gradient predicts: 9.23% on CXR, the easiest of the three tasks by every accuracy figure in Chapter 6, rising to 24.12% on VIS, and reaching 45.71% on INVIS, the hardest task by the same measure. This ordering was not built into GateKeeper's training objective anywhere; Equation 5.9's binary cross-entropy loss against the correctness target defined in Equation 5.11 has no notion of task identity, difficulty gradient, or CXR versus pelvic domain built into it at all. The three activation rates emerge purely from GateKeeper being trained separately on each task against that task's own correctness signal, and the fact that they land in the order they do is evidence, not proof, that the correctness signal itself is capturing something real about how often Stream A's unaided prediction can actually be trusted on a given task.

The comparison this argument leans on most heavily is not any single activation rate in isolation, but the relative ordering across all three. A 45.71% activation rate on INVIS alone would be a plausible outcome under several different explanations, including one in which GateKeeper is simply miscalibrated toward over-activation generally, activating too often even where it need not. What that alternative explanation cannot easily account for is why the same miscalibration would not equally inflate the CXR activation rate to something well above 9.23%. A gate that is well-calibrated to task-specific correctness produces exactly the pattern observed, low activation where Stream A is already reliable, high activation where it is not, tracking a different equilibrium on each task; a gate that is uniformly biased in one direction would not.

7.3.2 The Coupling Between Activation Rate and the AUROC Gap

Figure 6.1 and Table 6.8 make a related but distinct point more directly than activation rate does on its own. The AUROC gap between Stream A only and Always-On, 1.18

points on CXR, 8.40 on VIS, 15.78 on INVIS, is a direct measure of how much anatomical guidance actually helps on a given task, independent of GateKeeper or gating entirely, since it compares two configurations that never involve GateKeeper's decision at all. That this gap rises in lockstep with activation rate across the same three tasks, in the same order, is a stronger and more specific claim than the previous subsection's argument alone, because it connects GateKeeper's learned behavior directly to an independently measurable quantity: the actual diagnostic value anatomical guidance provides on each task.

This coupling is the clearest available evidence, short of a direct causal intervention this research does not attempt, that GateKeeper is allocating its more expensive pathway roughly in proportion to how much that pathway is actually worth on a given task, rather than in proportion to some other property of the task that happens to correlate with difficulty by coincidence. It is worth being precise about what this comparison does and does not establish. It does not establish that GateKeeper activates the correct individual samples within a task, since activation rate and the AUROC gap are both aggregate, task-level statistics rather than sample-level ones; a gate that activated on a large, poorly targeted subset of a hard task's cases could in principle produce a similar aggregate pattern to one that activates precisely on the cases where Stream A errs. What it does establish is the coarser, task-level claim this research's difficulty-gradient argument actually depends on: that GateKeeper's aggregate activation behavior, summed across an entire task, scales with how much that task benefits from anatomical guidance overall, which is the specific claim Objective 4 in Chapter 1 sets out to test.

7.4 Qualitative Evidence from GradCAM

The single GradCAM figure examined in Section 6.5 covers two cases, one from each task, and any interpretation drawn from it needs to be scaled to match that: two cases illustrate a mechanism, they do not establish how reliably that mechanism generalizes across a dataset, a limitation already acknowledged directly in Section 7.7.3 below and worth restating here rather than glossing over in the interpretation itself.

With that scaling in mind, the two cases are still worth reading closely, because they show something the aggregate metrics in Chapter 6 cannot: what fusion is actually doing to the model's spatial attention on a single, specific case, rather than what fusion does to a score averaged over hundreds of cases. In the chest X-ray case, where GateKeeper did not activate Stream B, Stream A's attention is already concentrated on both lung fields, precisely where a diagnostically relevant finding for a chest condition should be located, and the case was classified correctly using this attention alone. This is consistent with, though not proof of, the idea that GateKeeper's decision not to activate Stream B on this case reflected a genuine absence of need rather than an underestimate of the case's difficulty: Stream A's own attention pattern, examined independently of GateKeeper's decision, was already well-localized.

The invisible fracture case is the more informative of the two, precisely because it is the case where the mechanism SecondOpinion is built around is directly visible in a single image. Prior to fusion, Stream A's attention sits in the background corners of the radiograph,

away from any bone structure at all, which is a plausible explanation, at least for this one case, for why an invisible fracture, one with no obvious cortical disruption to fix attention on, would be difficult for an anatomy-blind stream to localize correctly in the first place. After fusion, attention shifts to the central bone region, and the case is classified correctly. Read narrowly, this single case is consistent with the claim that cross-attention fusion, when GateKeeper activates it, is doing something more specific than simply averaging in a second opinion; it appears, in this instance, to be redirecting Stream A's spatial focus toward the anatomically relevant region it had originally missed. Read too broadly, the same observation risks becoming a claim this research has not tested systematically, namely that this redirection happens reliably across the 45.71% of INVIS cases GateKeeper actually activates Stream B for, a claim that would require GradCAM analysis across that full activated subset rather than one representative case, and which is accordingly left as a direction for future interpretability work rather than a conclusion drawn here.

7.5 Situating SecondOpinion Against Prior Approaches

7.5.1 Against PelFANet and the Anatomy-Guided Lineage

Chapter 3 traced a lineage of anatomy-guided pelvic fracture detection work culminating in PelFANet, and Chapter 5 was explicit that SecondOpinion departs from every design in that lineage on one specific architectural axis: anatomical guidance is applied conditionally rather than unconditionally. The results in Chapter 6 are the first point in this work where that departure can be assessed against actual numbers rather than argued for on principle alone.

The comparison worth foregrounding here is not SecondOpinion against PelFANet directly, since the two differ in more than gating alone, but Always-On against PelFANet, since Always-On isolates the effect of SecondOpinion's cross-attention fusion mechanism from the effect of conditional gating entirely, evaluating both streams unconditionally exactly as PelFANet does. On this comparison, discussed already in Section 7.2.2, Always-On exceeds PelFANet on Specificity, F1, and AUROC on both pelvic subsets, while trailing on Recall. This establishes that the architectural lineage this research continues from PelFANet, replacing PelFANet's iterative CBAM-based FABlock fusion with a single late cross-attention step, is not simply a lighter-weight version of the same idea that sacrifices performance for simplicity; on the balanced metrics F1 and AUROC, it improves on PelFANet's own reported results even before conditional gating is introduced at all. Gating is then layered on top of this already-competitive fusion architecture, and SecondOpinion's results relative to Always-On, discussed fully in Section 7.2.1, show that this layer can be added while recovering a large majority of Always-On's performance at a fraction of its cost, rather than gating being a source of the performance difference from PelFANet in the first place.

7.5.2 Against Early-Exit and Conditional-Computation Literature

The early-exit literature reviewed in Chapter 2 establishes conditional computation as an idea with real prior traction, but one that has predominantly relied on gating signals derived from a network's own unsupervised output confidence [20, 21], with more recent work specifically showing that aligning a gate's training dynamics with its inference-time policy improves reliability over gates trained only as a byproduct of the underlying task loss [22]. GateKeeper's correctness-supervision design, argued for in Section 5.3.1 and examined empirically in Section 7.3.1 above, is this research's own answer to that same alignment problem, applied to a different architectural setting than the early-exit literature typically considers, gating between two anatomically distinct streams rather than gating an early exit out of a single, depth-wise network.

The activation-rate results discussed in Section 7.3 are, in this light, indirect evidence bearing on a question the early-exit literature has raised in principle but that this research is positioned to test empirically in a new setting: whether a correctness-supervised gate, aligned by construction with the decision it is meant to make, actually behaves more sensibly across tasks of differing difficulty than an unsupervised confidence threshold would be expected to. This research does not report a direct ablation against an unsupervised-confidence version of GateKeeper, a comparison noted as a limitation in Section 7.7.2, so this remains a comparison against the literature's general finding rather than against a controlled variant built and tested within this research itself.

7.5.3 Against Confidence-Gated Dual-Path Designs

Chapter 2's research gap was framed specifically against T-DuMpRa, a teacher-guided, dual-path, multi-prototype retrieval-augmented framework in which a confidence signal governs routing between a retrieval-based path and a directly learned representation path for fine-grained medical image classification [23]. T-DuMpRa establishes that confidence-gated routing between two computational paths is a workable idea in medical image classification outside the early-exit literature specifically, which is precisely why it was worth citing as the closest existing precedent to SecondOpinion's own gated, dual-path structure.

The distinction Chapter 2 draws, and that this research's results now give some empirical weight to, is that T-DuMpRa's two paths are not built around an anatomical guidance distinction the way Stream A and Stream B are, and its confidence-gated routing was not evaluated for whether it tracks a difficulty gradient across structurally distinct tasks the way GateKeeper's activation rate is examined in Section 7.3. SecondOpinion's contribution relative to this precedent is accordingly narrow and specific rather than a wholesale departure from it: applying a comparable gated dual-path idea to a setting where one path is explicitly anatomy-guided, and testing whether the resulting gate's activation behavior tracks task difficulty in a way that generalizes across genuinely different diagnostic domains, chest radiography and pelvic radiography, rather than within a single domain alone.

7.6 Broader Implications

Beyond the two specific tasks this research evaluates, the results in Chapter 6 point toward a broader argument about how anatomy-guided architectures might be built going forward, one that extends past pelvic fracture detection and chest disease classification specifically. The consistent finding across both tasks, that GateKeeper's activation rate scales with how much anatomical guidance actually helps on that task, as measured independently through the Stream A versus Always-On AUROC gap, suggests that the choice to invoke anatomical guidance conditionally need not be treated as a task-specific engineering decision made anew for each new diagnostic problem. If a correctness-supervised gate trained the same way, on GateKeeper's own architecture and training objective, produces sensible activation behavior on two structurally different tasks, chest disease classification with five classes and no fracture-specific structure at all, and pelvic fracture detection with an explicit visible-versus-invisible difficulty split, without any task-specific modification to the gating mechanism itself, this is at least suggestive evidence that the same conditional-gating recipe could transfer to other anatomy-guided settings in medical imaging without needing to be redesigned from scratch for each one.

This has a direct practical implication for how anatomy-guided dual-stream architectures might be deployed rather than only researched. Every prior anatomy-guided design traced in Chapter 3, and the broader dual-stream literature reviewed in Chapter 2, pays the full cost of a second stream on every single case that passes through the model, a cost that scales linearly with deployment volume regardless of how many of those cases actually needed the second opinion. A clinical deployment processing a large volume of chest radiographs, the overwhelming majority of which are unambiguous by SecondOpinion's own 9.23% CXR activation rate, would under an always-on architecture pay the anatomical-guidance cost on every one of those unambiguous cases as a matter of course. The efficiency argument this work makes is accordingly not only a matter of research convenience, faster experimentation cycles or lower compute budgets during model development, but a claim about what a production diagnostic pipeline built on this architecture would actually cost to run at scale, where the difference between paying for anatomical guidance on 9% of cases against 100% of cases compounds directly with deployment volume.

There is a second, more speculative implication worth naming here without overstating it. The difficulty-gradient framing this research adopts, chest disease classification easier than visible pelvic fracture detection easier than invisible pelvic fracture detection, was constructed deliberately across three tasks selected in advance to test whether activation rate would track it. A natural further question, outside the scope of what this research evaluates but suggested by its results, is whether GateKeeper's activation rate could itself serve as a diagnostic signal about task or dataset difficulty in settings where that difficulty is not already known in advance, rather than only being examined after the fact as this research does. This possibility is not tested here and is noted only as a direction the present results point toward, not a claim this research has established.

7.7 Limitations

7.7.1 Dataset Scale and Private Data

The INVIS subset used throughout this research's pelvic evaluation contains 35 cases, and despite 95% confidence intervals being reported alongside every metric computed on it in Chapter 6, point estimates drawn from a subset this small should be read with appropriate caution; a small number of individual cases classified differently would shift the reported metrics by a larger margin on INVIS than the same number of cases would on either the considerably larger VIS subset or the CXR dataset. This limitation is not introduced by this research; it is inherited directly from the AMERI dataset's own construction, and PelFANet's original evaluation, discussed in Chapter 3, was subject to the same constraint on the same underlying subset. It is restated here rather than treated as resolved, since nothing about SecondOpinion's own evaluation protocol changes the underlying sample size the INVIS conclusions in Section 7.2.3 rest on.

A related and broader constraint is that the AMERI pelvic dataset is private, collected at a single institution, Steel Memorial Hirohata Hospital, under a specific IRB approval described in Chapter 4, and is not available for independent replication outside a data-sharing request to that institution. Every pelvic result in this research, and in PelFANet before it, reflects performance on data from a single imaging site, a single patient population, and a single set of acquisition protocols and equipment. Whether SecondOpinion's pelvic results, including the activation-rate pattern central to Section 7.3, would hold on pelvic radiographs acquired at a different institution, with different imaging equipment, patient demographics, or radiographic conventions, is a question this research's evaluation cannot answer, since no external pelvic dataset was evaluated. This research's own stated scope, noted in Chapter 1, explicitly excludes AMERI-to-PXR150 cross-dataset validation of the kind explored for segmentation alone in the Beyond the Pelvic Ring study reviewed in Chapter 3; that exclusion was a deliberate scoping decision made to keep this research's pelvic evaluation focused on the same institution's own visible-to-invisible generalization question, but it means the cross-institution generalization question remains genuinely open rather than answered negatively or positively by anything reported here.

7.7.2 Single Backbone and Gating Architecture

Every result in Chapter 6 was produced using a single fixed architectural choice at every level of the pipeline described in Chapter 5: EfficientNet-B0 for both streams, a U-Net with MiT-B0 encoder for segmentation, and a specific, fixed MLP design for GateKeeper's own architecture. This research does not report an ablation varying any of these choices independently, for instance substituting a different backbone family for Stream A or Stream B, varying GateKeeper's depth or width, or testing an unsupervised-confidence variant of GateKeeper against the correctness-supervised version actually used, a specific instance of this same limitation already noted in Section 7.5.2.

This matters for how broadly the conclusions of Sections 7.2 and 7.3 should be read. The argument that correctness-supervised gating tracks task difficulty, and the argument that

cross-attention fusion outperforms PelFANet's iterative CBAM fusion on balanced metrics while trailing on Recall, are both established here for one specific instantiation of each architectural choice. It remains an open question, not addressed by this research, whether these findings are properties of the general design principles argued for in Chapter 5, correctness supervision as a gating strategy, late asymmetric cross-attention as a fusion strategy, or whether they are specific to the particular backbone and gate architecture this research happened to build them with.

7.7.3 Limited Qualitative Evaluation

The GradCAM analysis in Section 6.5, discussed in Section 7.4 above, covers two cases, chosen as representative examples of the two behaviors most central to this research's argument, Stream B remaining inactive on a straightforward case and Stream B correcting a poorly localized prediction on a difficult one. Two cases are sufficient to illustrate that these behaviors occur, but not to establish how consistently they occur across the full set of cases GateKeeper activates Stream B for. In particular, the claim advanced cautiously in Section 7.4, that fusion redirects Stream A's attention toward anatomically relevant regions specifically on cases where that redirection is needed, was illustrated on one INVIS case and has not been verified across the 45.71% of the INVIS subset GateKeeper actually activates Stream B for at inference. A systematic interpretability evaluation, aggregating GradCAM or a comparable attribution method across every activated case in a given task rather than a small number of hand-selected examples, would be needed to state that claim with the same confidence given to the quantitative results in Section 7.2.

7.7.4 Scope Exclusions

Beyond the AMERI-to-PXR150 cross-dataset exclusion already discussed in Section 7.7.1, this research's evaluation of the segmentation pipeline described in Section 5.2.1 is indirect throughout. The segmentation model's own output quality, whether measured through Dice Similarity Coefficient, Intersection over Union, or any other segmentation-specific metric, is never reported directly in this research; the segmentation model's contribution is assessed only through its downstream effect on Stream B's and the fused pathway's classification performance in Chapter 6. This mirrors how the segmentation stage was treated in PelFANet itself, from which this research's segmentation architecture and training protocol are carried forward largely unchanged, as described in Section 5.2.1, but it means that a specific and answerable question, how accurate the anatomical masks feeding Stream B actually are on either dataset, is left unanswered by this research's own evaluation, even though that accuracy plausibly bears directly on how much diagnostic value Stream B is able to extract from the anatomical input it is given.

Chapter-8: Conclusion

8.1 Overview

This research set out from a single motivating question, stated in Chapter 1: can a diagnostic model learn to allocate additional anatomical reasoning only when it is actually needed, rather than applying the same fixed computation to every case regardless of difficulty? That question was broken down into three specific research questions in Chapter 1, concerning whether conditional gating can match always-on performance at a fraction of the cost, whether a correctness-supervised gate's activation behavior tracks genuine diagnostic difficulty, and whether that tracking behavior generalizes across diagnostically distinct domains. Chapter 2 situated these questions at the intersection of two established but previously separate lines of work, anatomy-guided dual-stream architectures and conditional computation, and identified the specific gap neither had addressed together. Chapter 3 traced this work's own research lineage toward the same underlying question from a different direction, arriving at PelFANet as the immediate architectural predecessor still bound by the always-on assumption this research departs from. Chapters 5 through 7 then answered the three research questions directly: Chapter 5 proposed SecondOpinion and GateKeeper as the system capable of generating evidence relevant to them, Chapter 6 measured that evidence across two diagnostic tasks of increasing difficulty, and Chapter 7 interpreted those measurements against the three questions this research set out to answer. This chapter closes the report by revisiting those research questions explicitly, summarizing what was found, considering the work's broader social, environmental, and ethical dimensions, and outlining the directions Chapter 7's limitations leave open for future work.

8.2 Research Questions Answered

Chapter 1 posed three research questions this work set out to answer. Each is restated below alongside a direct account of the evidence that answers it, drawn from the results reported in Chapter 6 and interpreted in Chapter 7.

RQ1: Can anatomical guidance be applied conditionally, through a learned gating mechanism, and match the diagnostic performance of an unconditional, always-on dual-stream architecture at a substantially lower computational cost? This question is answered affirmatively across all three diagnostic tasks evaluated in this research. On the CXR task, SecondOpinion reached 98.41% Accuracy and 0.990

AUROC against Always-On's 99.05% and 0.993, a gap under 1.2 percentage points on every metric, while activating its anatomy-guided stream on only 9.23% of cases and incurring, by Table 6.7's activation-weighted accounting, just 52% of Always-On's FLOPs. On the pelvic tasks, SecondOpinion trailed Always-On by similarly narrow margins, 2.2 points of F1 on VIS and 1.2 points on INVIS, while incurring 60% and 71% of Always-On's FLOPs respectively. In every case, the performance retained relative to the always-on ceiling was disproportionately large relative to the computational cost actually paid, a relationship examined directly in Section 7.2.1 and treated there as the explicit, deliberate cost of conditional computation rather than an unexplained shortfall. RQ1 is answered by this consistent pattern: conditional gating did not merely approximate always-on performance by coincidence on one task, but did so across every task this research evaluated, at costs that scaled with how much anatomical guidance that task actually required.

RQ2: Does a gating mechanism's activation rate, when trained explicitly for correctness prediction rather than relying on unsupervised confidence, correlate with the underlying diagnostic difficulty of the case or task it is applied to? This question is also answered affirmatively, and with evidence more specific than RQ1's alone. Section 7.3.2 established that GateKeeper's activation rate, 9.23% on CXR, 24.12% on VIS, and 45.71% on INVIS, rose in the same order and in step with an entirely independent measure of task difficulty, the AUROC gap between Stream A only and Always-On, 1.18, 8.40, and 15.78 points across the same three tasks respectively. Because this second quantity is computed without any involvement of GateKeeper or gating whatsoever, comparing it against activation rate provides a check on RQ2 that does not rely on activation rate alone, and the two quantities moved together, as illustrated in Figure 6.1. This coupling was argued in Section 7.3.1 to be difficult to reconcile with any explanation other than GateKeeper's correctness-supervision objective genuinely tracking how much a given task benefits from anatomical guidance, since a uniformly miscalibrated gate would be expected to inflate or deflate activation on every task roughly alike, rather than scaling specifically with each task's own measured difficulty.

RQ3: Does this relationship between activation behavior and task difficulty hold consistently across diagnostically distinct domains, rather than being an artifact specific to a single task or dataset? This question is answered affirmatively as well, on the strength of the specific comparison this research was designed to make possible. The activation-rate pattern examined under RQ2 was not observed within a single domain alone; it held both across a between-domain comparison, chest disease classification against pelvic fracture detection, two tasks differing in imaging modality, class structure, and clinical context entirely, and across a within-domain comparison, visible against invisible pelvic fracture cases, two subsets of the same dataset differing specifically in how visually apparent the underlying pathology is. That the same correctness-supervised gating mechanism, trained separately and without any task-specific modification on each of the three tasks, produced activation behavior consistent with the difficulty gradient in both comparisons is the evidence this research offers for RQ3. A relationship that held only within one of these two comparisons, for instance tracking visible-versus-invisible difficulty but not chest-versus-pelvic difficulty, would have left RQ3 open; that it held across both is what allows this research to conclude that GateKeeper's activation behavior reflects a property of the correctness-supervision design itself, examined further as a broader implication in Section 8.6, rather than a pattern specific to the particular dataset or task it happened to be tested on first.

8.3 Summary of Key Findings

Taken together, the results reported in Chapter 6 and interpreted in Chapter 7 support a small number of central findings, beyond the direct answers to the three research questions above. On the CXR task, SecondOpinion approximates the performance of an unconditional dual-stream model within a fraction of a percentage point on most metrics, while activating its anatomy-guided stream on only 9.23% of cases, exceeding every external CXR baseline evaluated in the process, including CheXNet-CBAM, the strongest of the three. On both pelvic subsets, SecondOpinion and its always-on counterpart trade a modest Recall shortfall against PelFANet for a larger gain in Specificity, producing higher F1 and AUROC on both VIS and INVIS despite this trade-off, a pattern traced in Section 7.2.2 to the fusion architecture itself rather than to gating, since Always-On, with gating disabled entirely, shows the same trade-off against PelFANet that SecondOpinion does. On the invisible fracture subset specifically, evaluated purely as a generalization test since no configuration in this research was ever trained on an invisible fracture case, SecondOpinion's Recall gap against the always-on ceiling narrows to its smallest margin of any task studied, alongside the largest AUROC gap between anatomy-blind and anatomy-guided prediction observed across the three tasks, consistent with GateKeeper allocating anatomical reasoning where this research's own difficulty-gradient argument predicts it is needed most.

8.4 Social Impact Analysis

This research was designed with an eye toward how conditional computation of this kind could extend the practical reach of anatomy-guided diagnostic support, particularly in settings where computational resources cannot be assumed to be abundant. A radiology setting without access to large, always-available compute infrastructure benefits disproportionately from a diagnostic architecture whose typical cost is a small fraction of its worst-case cost, since the majority of cases in such a setting can be served at the cost of a single lightweight backbone rather than a full dual-stream evaluation on every case. In this sense, the efficiency SecondOpinion demonstrates is not only a matter of computational convenience during research but a design property with a direct bearing on where anatomy-guided diagnostic support could plausibly be deployed at all.

The clinical analogy underlying SecondOpinion's design, a fast initial read escalated to a second opinion only on genuinely ambiguous cases, was also chosen deliberately to keep the system's role framed appropriately relative to a clinician's own judgment, rather than as a replacement for it. GateKeeper's decision to invoke Stream B mirrors a clinician's own decision to seek a second reading, a workflow radiologists already practice routinely, and this research's framing treats SecondOpinion as a tool that could plausibly sit inside that existing workflow rather than one that displaces the clinical judgment it is meant to support. Considered this way, the social benefit this research aims toward is one of extending access to anatomically informed diagnostic assistance to a wider range of clinical settings, without asking those settings to bear the full computational cost that prior always-on anatomy-guided architectures would require of every case.

8.5 Environmental Impact and Sustainability

Conditional computation carries a direct environmental implication alongside its computational one, since a lower average inference cost per case corresponds, in principle, to lower average energy draw per case at deployment scale. This research measured that cost specifically through Params and FLOPs, reported per configuration and per task in Table 6.7, rather than through direct energy or carbon accounting, and this distinction is worth stating plainly: FLOPs are a widely used and informative proxy for computational cost, but they are a proxy, and the results in this research should be read as establishing a favorable direction for energy efficiency rather than a direct energy measurement.

Read as that proxy, the results are nonetheless meaningful. SecondOpinion's activation-weighted FLOPs, computed according to Equation 5.14, ranged from 0.442 GFLOPs on the CXR task to 0.608 GFLOPs on the hardest task evaluated, INVIS, against the Always-On configuration's fixed 0.855 GFLOPs regardless of task, meaning SecondOpinion's realized computational cost stayed at 52% to 71% of the always-on ceiling across every task this research evaluated, without ever reaching that ceiling even on the task where anatomical guidance was needed most. At the deployment scale a diagnostic system of this kind is intended for, where a single architecture may be applied across a large and continuing volume of cases, a computational saving of this magnitude compounds directly with volume, and this research's conditional-computation design was chosen specifically with that compounding relationship, and the sustainability benefit it implies, in mind.

8.6 Ethical Considerations

This research was conducted with explicit attention to the ethical dimensions its data sources and its intended application area both raise. The AMERI pelvic fracture dataset used throughout the pelvic side of this research was collected and used under Institutional Review Board approval (IRB No. 2019-1-52) from Steel Memorial Hirohata Hospital, Japan, as described in Chapter 4, and every use of that dataset in this research, including its incorporation into SecondOpinion's training and evaluation, was conducted within the terms of that approval. The public datasets used on the CXR side of this research, described in the same chapter, were likewise used in accordance with their respective release terms and licenses.

Beyond data governance, this research's own framing was chosen deliberately to situate SecondOpinion as a decision-support tool rather than an autonomous diagnostic authority. The clinical analogy invoked throughout Chapter 5, a second opinion sought on a difficult case rather than a final, unilateral verdict, was adopted specifically because it reflects the role this research intends such a system to occupy: assisting a clinician's own judgment on cases flagged as needing further scrutiny, not replacing that judgment. This framing was carried through consistently, from the architecture's own name to the way its results are interpreted in Chapter 7, and reflects this research's position that AI-assisted diagnostic tools of this kind are most appropriately developed and evaluated as an aid to clinical decision-making, positioned alongside a clinician's expertise rather than in place of it.

8.7 Future Work

Chapter 7 identified several specific limitations in the present evaluation, and each points toward a concrete direction for future work rather than remaining an open-ended concern.

Broader Backbone and Gating Ablations. Section 7.7.2 noted that every result in this research was produced with a single fixed backbone family, EfficientNet-B0, and a single fixed GateKeeper architecture. A natural extension is to test whether the correctness-supervision and cross-attention fusion principles argued for in Chapter 5 hold when instantiated with different backbone families for Stream A and Stream B, and to test an unsupervised-confidence variant of GateKeeper directly against the correctness-supervised version used throughout this research, isolating the specific contribution of correctness supervision that RQ2 currently answers only indirectly, through the coupling argument in Section 7.3.2 rather than a controlled comparison against an alternative gating strategy built and tested within this research itself.

Cross-Institution and Cross-Dataset Validation. Section 7.7.1 noted that the AMERI dataset reflects a single institution's patient population and imaging protocol, and that AMERI-to-PXR150 cross-dataset validation was explicitly placed outside this research's scope. Future work extending SecondOpinion's pelvic evaluation to PXR150 or to an entirely independent pelvic radiograph cohort would test whether RQ3's answer, that activation behavior tracks difficulty consistently across domains, extends further still to a comparison across institutions and acquisition protocols, beyond the two within-institution domains this research evaluates.

Systematic Interpretability Evaluation. Section 7.7.3 noted that this research's GradCAM analysis covers two representative cases rather than a systematic sample. Future work aggregating GradCAM or a comparable attribution method across the full set of cases GateKeeper activates Stream B for, on both the CXR and pelvic tasks, would allow the qualitative claims made cautiously in Section 7.4 about fusion's effect on spatial attention to be stated with the same confidence given to this research's quantitative answers to RQ1 and RQ2.

Extending Conditional Gating to New Backbones and Modalities. More broadly, RQ3's affirmative answer across two structurally distinct tasks suggests that GateKeeper's correctness-supervision recipe may transfer to anatomy-guided settings beyond chest radiography and pelvic radiography specifically. Future work applying the same gating and fusion design to other modalities and anatomical regions would test how far this research's central finding, that a correctness-supervised gate's activation behavior tracks how much anatomical guidance actually helps on a given task, extends beyond the two tasks evaluated here.

8.8 Closing Remarks

This research began with an observation about how diagnostic effort is allocated in clinical practice: straightforward cases are resolved quickly, and only genuinely ambiguous ones

prompt a clinician to seek a second opinion. SecondOpinion was built to give a diagnostic model this same adaptive discipline, formalized through GateKeeper rather than left to a fixed rule, and evaluated across two tasks chosen specifically to test whether that discipline holds up as diagnostic difficulty and modality changes. The results reported in this research suggest that it does: a model trained to know when it needs a second opinion, and not merely trained to produce one on every case regardless, can approach the performance of a model that always asks for one, while asking for one only when the case in front of it actually warrants the additional scrutiny.

Bibliography

[1] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017.

[2] C. Jin, J. Udupa, L. Zhao, Y. Tong, D. Odhner, G. Pednekar, S. Nag, S. Lewis, N. Poole, S. Mannikeri *et al.*, "Anatomy-guided deep learning for object localization in medical images," in *Medical Imaging 2022: Image Processing*, vol. 12032. SPIE, 2022, pp. 591–597.

[3] P. Bruno, M. Macrì, and C. Dodaro, "A dual-stage deep learning framework for breast ultrasound image segmentation and classification," *Journal of Medical Systems*, vol. 49, no. 1, p. 162, 2025.

[4] J. Zafar, V. Koc, and H. Zafar, "Dual-stream contrastive latent learning generative adversarial network for brain image synthesis and tumor classification," *Journal of Imaging*, vol. 11, no. 4, p. 101, 2025.

[5] I. Kasuga *et al.*, "Evaluation of chest radiography and low-dose computed tomography as valuable screening tools for thoracic diseases," *Medicine*, vol. 101, no. 29, p. e29261, 2022.

[6] O. Ruuskanen, E. Lahti, L. C. Jennings, and D. R. Murdoch, "Viral pneumonia," *The Lancet*, vol. 377, no. 9773, pp. 1264–1275, 2011.

[7] UNICEF, "Statistical tables. the state of the world's children 2016," UNICEF report, pp. 107–172, 2016.

[8] World Health Organization, "Tuberculosis," https://www.who.int/news-room/fact-sheets/detail/tuberculosis, 2025, accessed 2026-03-04.

[9] GBD 2021 Lower Respiratory Infections and Antimicrobial Resistance Collaborators, "Global, regional, and national incidence and mortality burden of non-COVID-19 lower respiratory infections and aetiologies, 1990–2021: a systematic analysis from the global burden of disease study 2021," *The Lancet Infectious Diseases*, vol. 24, no. 9, pp. 974–1002, 2024.

[10] S. A. Duzgun, G. Durhan, F. B. Demirkazik, M. G. Akpinar, and O. M. Ariyurek, "COVID-19 pneumonia: the great radiological mimicker," *Insights into Imaging*, vol. 11, no. 1, p. 118, 2020.

[11] B. Ferede, A. Ayenew, and W. Belay, "Pelvic fractures and associated injuries in patients admitted to and treated at emergency department of Tibebe Ghion

specialized hospital, Bahir Dar university, Ethiopia," *Orthopedic Research and Reviews*, pp. 73–80, 2021.

[12] H. Abdelrahman, A. El-Menyar, H. Keil, A. Alhammoud, S. I. Ghouri, E. Babikir, M. Asim, M. Muenzberg, and H. Al-Thani, "Patterns, management, and outcomes of traumatic pelvic fracture: insights from a multicenter study," *Journal of Orthopaedic Surgery and Research*, vol. 15, no. 1, p. 249, 2020.

[13] J. Ohla, P. Walus, M. Wicinski, B. Malkowski, B. Turon, A. Jablonski, M. Gawryjolek, K. Kellett, and J. Zabrzynski, "Pelvic fractures in adults and the importance of associated injuries: a current multi-disciplinary approach," *Clinics and Practice*, vol. 15, no. 7, p. 130, 2025.

[14] G. S. Connor, G. McGwin Jr, P. A. Maclennan, J. E. Alonso, and L. W. Rue III, "Early versus delayed fixation of pelvic ring fractures," *The American Surgeon*, vol. 69, no. 12, pp. 1019–1024, 2003.

[15] J. P. Grimes, P. M. Gregory, H. Noveck, M. S. Butler, and J. L. Carson, "The effects of time-to-surgery on mortality and morbidity in patients following hip fracture," *The American Journal of Medicine*, vol. 112, pp. 702–709, 2002.

[16] E. R. Benjamin *et al.*, "The trauma pelvic X-ray: not all pelvic fractures are created equally," *The American Journal of Surgery*, vol. 224, pp. 489–493, 2022.

[17] M. Weitz, C. Schwartz, and M. H. Scheinfeld, "Radiologic blind spots in hip and pelvic radiographs," *Emergency Radiology*, vol. 30, no. 5, pp. 569–575, 2023.

[18] A. Pinto, D. Berritto, A. Russo, F. Riccitiello, M. Caruso, M. P. Belfiore, V. R. Papapietro, M. Carotti, F. Pinto, A. Giovagnoni *et al.*, "Traumatic fractures in adults: missed diagnosis on plain radiographs in the emergency department," *Acta Bio Medica: Atenei Parmensis*, vol. 89, no. Suppl 1, pp. 111–123, 2018.

[19] J. R. Soto, C. Zhou, D. Hu, A. C. Arazoza, E. Dunn, and P. Sladek, "Skip and save: utility of pelvic X-rays in the initial evaluation of blunt trauma patients," *The American Journal of Surgery*, vol. 210, no. 6, pp. 1076–1081, 2015.

[20] N. Passalis, J. Raitoharju, A. Tefas, and M. Gabbouj, "Efficient adaptive inference for deep convolutional neural networks using hierarchical early exits," *Pattern Recognition*, vol. 105, p. 107346, 2020.

[21] M. Wołczyk, B. Wójcik, K. Bałazy, I. T. Podolak, J. Tabor, M. Śmieja, and T. Trzcinski, "Zero time waste: Recycling predictions in early exit neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2516–2528, 2021.

[22] S. Mokssit, O. Karrakchou, A. Mousist, and M. Ghogho, "Confidence-gated training for efficient early-exit neural networks," 2025.

[23] Z. Tang and S. Zhao, "T-DuMpRa: Teacher-guided dual-path multi-prototype retrieval augmented framework for fine-grained medical image classification," 2026.

[24] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," pp. 248–255, 2009.

[26] F. Israq, S. Wasi, S. B. Noor, S. T. Bhuiyan, S. B. Alam, and R. Rahman, "Ultrasound image-based liver disease detection using transfer learning: A comparative analysis of cnn and their variants," in *2026 7th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*. IEEE, 2026, pp. 744–750.

[27] M. S. Choudhury, R. Rahman, S. T. Bhuiyan, S. Wasi, and S. B. Alam, "Deep learning for accurate classification of pulmonary edema severity in chest x-rays: A comparative evaluation of cnn models with grad-cam feature visualization," in *2025 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*. IEEE, 2025, pp. 72–77.

[28] M. Z. Iqbal, H. Haque, F. Jamil, R. Rahman, S. T. Bhuiyan, R. Rahman, A. Islam, and S. B. Alam, "Benchmarking cnn architectures using transfer learning for lipomatous tumor classification: A comprehensive study," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2026, pp. 171–182.

[29] N. E. J. Hema, M. S. Choudhury, S. Wasi, S. T. Bhuiyan, F. Israq, N. Anan, S. B. Alam, A. Islam, and R. Rahman, "A comprehensive benchmark study of transfer learning-based deep learning models for brain tumor classification," in *2026 5th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*. IEEE, 2026, pp. 1–6.

[30] S. A. Kotha, A. A. A. Shanto, M. Esha, S. T. Bhuiyan, H. Khatun, R. Rahman, A. Islam, and S. B. Alam, "Learning from both eyes: A bilateral ensemble approach for retinal disease classification," in *2025 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2025, pp. 480–485.

[31] S. T. Bhuiyan, H. Khatun, S. K. Mazumder, F. Israq, R. Islam, R. Rahman, A. Islam, and S. B. Alam, "The role of model architecture in breast cancer classification: A comparison of cnn and transformer paradigms on ultrasound data," in *2025 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2025, pp. 356–361.

[32] S. Sarkar, F. Yesmin, M. A. S. Tasin, F. Israq, S. T. Bhuiyan, R. Rahman, S. B. Alam, and A. Islam, "Effect of augmentations on cnn based models for congenital heart disease detection from chest x-rays," in *2025 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2025, pp. 438–443.

[33] M. T. Islam, M. Alam, J. S.-U. I. Smaron, J. H. Setu, S. T. Bhuiyan, R. Rahman, A. Mahmud, S. B. Alam, and A. Islam, "Comparative evaluation of deep learning models for multi-disease classification using reflective dental imaging," in *2025 IEEE 23rd Student Conference on Research and Development (SCOReD)*. IEEE, 2025, pp. 1–6.

[34] A. Iqbal, M. R. N. Hridoy, I. H. Mim, R. Islam, S. T. Bhuiyan, S. Wasi, F. Sarkar, and S. B. Alam, "A morphology-aware preprocessing pipeline for improved glioblastoma segmentation in brain mri using hybrid segformer," in *2026 IEEE 2nd International*

Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN). IEEE, 2026, pp. 1–6.

[35] M. Yousuf, A. Yasin, M. Mohammad, F. Israq, S. T. Bhuiyan, R. Rahman, S. B. Alam, and A. Islam, "Deep learning framework for automated pulmonary nodule identification and anatomical segmentation in thoracic ct scans," in *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN)*. IEEE, 2026, pp. 1–6.

[36] M. J. H. Mekat, K. S. Nawal, J. O. D. Rozario, S. T. Bhuiyan, M. A. B. Khaled, M. R. Rahman, A. Islam, and S. B. Alam, "Deep learning-based stenosis segmentation in x-ray angiography: Vision transformers vs. cnns," in *2025 28th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2025, pp. 3553–3558.

[37] M. Z. Iqbal, R. Rahman, K. T. Nahar, S. T. Bhuiyan, A. Islam, S. B. Alam, R. Rahman, and M. A. Amin, "A cross-modal benchmark of deep learning encoders for automated liver segmentation," in *2025 28th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2025, pp. 3282–3287.

[38] S. Absar, I. Ahmed, S. Sakib, N. Sharma, S. T. Bhuiyan, S. B. Alam, A. Islam, and R. Rahman, "Multi-strategy optimization of u-net variants for orthopantomogram segmentation," in *2025 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)*. IEEE, 2025, pp. 17–22.

[39] M. Z. Iqbal, S. T. Bhuiyan, R. Islam, H. Khatun, S. K. Mazumder, R. Rahman, A. Islam, and S. B. Alam, "A two-stage deep learning framework for liver and hepatocellular carcinoma segmentation on mri," in *2025 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)*. IEEE, 2025, pp. 359–364.

[40] M. F. A. Sayeedi, R. Rahman, S. T. Bhuiyan, S. Wasi, A. Islam, S. B. Alam, and A. Rahman, "Route, retrieve, reflect, repair: Self-improving agentic framework for visual detection and linguistic reasoning in medical imaging," *arXiv preprint arXiv:2601.08192*, 2026.

[41] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," pp. 2097–2106, 2017.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2016.

[43] 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 (CVPR)*, 2017, pp. 4700–4708.

[44] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: inverted residuals and linear bottlenecks," pp. 4510–4520, 2018.

[45] M. Tan and Q. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," pp. 6105–6114, 2019.

[46] M. Constantinou, T. Exarchos, A. G. Vrahatis, and P. Vlamos, "COVID-19 classification on chest X-ray images using deep learning methods," *International Journal of Environmental Research and Public Health*, vol. 20, no. 3, p. 2035, 2023.

[47] H. A. Sanghvi, R. H. Patel, A. Agarwal, S. Gupta, V. Sawhney, and A. S. Pandya, "A deep learning approach for classification of COVID and pneumonia using DenseNet-201," *International Journal of Imaging Systems and Technology*, vol. 33, no. 1, pp. 18–38, 2023.

[48] F. M. J. M. Shamrat, S. Azam, A. Karim, K. Ahmed, F. M. Bui, and F. De Boer, "High-precision multiclass classification of lung disease through customized MobileNetV2 from chest X-ray images," *Computers in Biology and Medicine*, vol. 155, p. 106646, 2023.

[49] N. Novalina and M. Rizkinia, "Implementation of Xception and EfficientNetB3 for COVID-19 detection on chest X-ray image via transfer learning," *International Journal of Electrical, Computer & Biomedical Engineering*, vol. 1, no. 2, pp. 168–178, 2023.

[50] M. A. Kassem, S. M. Naguib, H. M. Hamza, M. M. Fouda, M. K. Saleh, and K. M. Hosny, "Explainable transfer learning-based deep learning model for pelvis fracture detection," *International Journal of Intelligent Systems*, vol. 2023, no. 1, p. 3281998, 2023.

[51] L. Tanzi, E. Vezzetti, R. Moreno, A. Aprato, A. Audisio, and A. Massè, "Hierarchical fracture classification of proximal femur X-ray images using a multistage deep learning approach," *European Journal of Radiology*, vol. 133, p. 109373, 2020.

[52] C.-T. Cheng, Y. Wang, H.-W. Chen, P.-M. Hsiao, C.-N. Yeh, C.-H. Hsieh, S. Miao, J. Xiao, C.-H. Liao, and L. Lu, "A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs," *Nature Communications*, vol. 12, no. 1, p. 1066, 2021.

[53] Z. Zheng, B. Y. Ryu, S. E. Kim, D. S. Song, S. H. Kim, J.-W. Park, and D. H. Ro, "Deep learning for automated hip fracture detection and classification: achieving superior accuracy," *The Bone & Joint Journal*, vol. 107, no. 2, pp. 213–220, 2025.

[54] S. T. Bhuiyan, R. Rahman, S. Wasi, H. Khatun, A. Islam, A. M. Rahman, S. B. Alam, and M. A. Amin, "Pelvinext: A modality-agnostic hybrid network for pelvic imaging in women's health," *arXiv preprint arXiv:2608.20144*, 2026.

[55] R. Rahman, N. Yagi, K. Hayashi, A. Maruo, H. Muratsu, and S. Kobashi, "Enhancing fracture diagnosis in pelvic X-rays by deep convolutional neural network with synthesized images from 3D-CT," *Scientific Reports*, vol. 14, no. 1, p. 8004, 2024.

[56] Y. Ma, Y. Guo, W. Cui, J. Liu, Y. Li, Y. Wang, and Y. Qiang, "SG-TransUNet: a segmentation-guided transformer U-Net model for KRAS gene mutation status identification in colorectal cancer," *Computers in Biology and Medicine*, vol. 173, p. 108293, 2024.

[57] A. Jaus, C. Seibold, S. Reiß, L. Heine, A. Schily, M. Kim, F. H. Bahnsen, K. Herrmann, R. Stiefelhagen, and J. Kleesiek, "Anatomy-guided pathology segmentation," vol. 15008, pp. 3–13, 2024.

[58] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," 2021.

[59] J.-M. Lee, J.-Y. Park, Y.-J. Kim, and K.-G. Kim, "Deep-learning-based pelvic automatic segmentation in pelvic fractures," *Scientific Reports*, vol. 14, no. 1, p. 12258, 2024.

[60] S.-H. Lee, J. Jeon, G.-J. Lee, J.-Y. Park, Y.-J. Kim, and K.-G. Kim, "Automated association for osteosynthesis foundation and orthopedic trauma association classification of pelvic fractures on pelvic radiographs using deep learning," *Scientific Reports*, vol. 14, no. 1, p. 20548, 2024.

[61] M. Zubair, M. Hussain, M. A. Al-Bashrawi, M. Bendechache, and M. Owais, "A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical image fusion for improved diagnosis," *Computer Methods and Programs in Biomedicine*, p. 109014, 2025.

[62] M. Safari, A. Fatemi, and L. Archambault, "MedFusionGAN: multimodal medical image fusion using an unsupervised deep generative adversarial network," *BMC Medical Imaging*, vol. 23, no. 1, p. 203, 2023.

[63] J. Liu, Y. Li, X. Sun, X. Wang, and H. Luo, "MMIF: multimodal medical image fusion network based on multi-scale hybrid attention," *Computers, Materials and Continua*, vol. 85, no. 2, pp. 3551–3568, 2025.

[64] J. D. Krogue *et al.*, "Automatic hip fracture identification and functional subclassification with deep learning," *Radiology: Artificial Intelligence*, vol. 2, no. 2, p. e190023, 2020.

[65] G. Kitamura, "Deep learning evaluation of pelvic radiographs for position, hardware presence, and fracture detection," *European Journal of Radiology*, vol. 130, p. 109139, 2020.

[66] X. Zhang *et al.*, "A new window loss function for bone fracture detection and localization in x-ray images with point-based annotation," 2020.

[67] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, and T. Darrell, "A convnet for the 2020s," pp. 11 976–11 986, 2022.

[68] B. Aslan, W. Kazaka, T. Slaven, S. Chetty, and G. Nitschke, "Deep-learning classifiers for small data orthopedic radiology," in *IEEE Symposium on Computational Intelligence in Health and Medicine (CIHM)*, 2025, pp. 1–7.

[69] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

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

[71] S. Shree, R. Bhagawati, S. Chanda, and D. R. Neog, "From pixels to diagnosis: a systematic review of deep learning in femoral fracture detection and classification," *BIO Web of Conferences*, vol. 204, p. 01018, 2025.

[72] Y. Ma, J. C. Mandell, T. Rocha, M. A. Mendicuti, M. J. Weaver, and B. Khurana, "Diagnostic accuracy of pelvic radiographs for the detection of traumatic pelvic fractures in the elderly," *Emergency Radiology*, vol. 29, no. 6, pp. 1009–1018, 2022.

[73] A. K. Obaid, A. Barleben, D. Porral, S. Lush, and M. Cinat, "Utility of plain film pelvic radiographs in blunt trauma patients in the emergency department," *The American Surgeon*, vol. 72, no. 10, pp. 951–954, 2006.

[74] A. Giełczyk, A. Marciniak, M. Tarczewska, and Z. Lutowski, "Pre-processing methods in chest x-ray image classification," *PLOS One*, vol. 17, no. 4, p. e0265949, 2022.

[75] J. Han, B. Choi, J. Y. Kim, and Y. Lee, "BO-CLAHE enhancing neonatal chest x-ray image quality for improved lesion classification," *Scientific Reports*, vol. 15, no. 1, p. 4931, 2025.

[76] S. T. Bhuiyan, R. Khatun, S. K. Mazumder, F. Israq, S. Wasi, R. Rahman, A. Islam, and S. B. Alam, "Preprocessing matters: Benchmarking image enhancement techniques for pelvic fracture detection," in *2025 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2025, pp. 379–384.

[77] M. P. Sullivan, K. D. Baldwin, D. J. Donegan, S. Mehta, and J. Ahn, "Geriatric fractures about the hip: divergent patterns in the proximal femur, acetabulum, and pelvis," *Orthopedics*, vol. 37, no. 3, pp. 151–157, 2014.

[78] Z. Balogh, K. L. King, P. Mackay, D. McDougall, S. Mackenzie, J. A. Evans, A. Lyons, and S. A. Deane, "The epidemiology of pelvic ring fractures: a population-based study," *Journal of Trauma and Acute Care Surgery*, vol. 63, no. 5, pp. 1066–1073, 2007.

[79] F. Rosa, D. Buccicardi, A. Romano, F. Borda, M. C. D'Auria, and A. Gastaldo, "Artificial intelligence and pelvic fracture diagnosis on x-rays: a preliminary study on performance, workflow integration and radiologists' feedback assessment in a spoke emergency hospital," *European Journal of Radiology Open*, vol. 11, p. 100504, 2023.

[80] T. J. O'Connor and P. A. Cole, "Pelvic insufficiency fractures," *Geriatric Orthopaedic Surgery & Rehabilitation*, vol. 5, no. 4, pp. 178–190, 2014.

[81] C. Lee, J. Jang, S. Lee, Y. S. Kim, H. J. Jo, and Y. Kim, "Classification of femur fracture in pelvic x-ray images using meta-learned deep neural network," *Scientific Reports*, vol. 10, no. 1, p. 13694, 2020.

[82] N. Nakata and T. Siina, "Ensemble learning of multiple models using deep learning for multiclass classification of ultrasound images of hepatic masses," *Bioengineering*, vol. 10, no. 1, p. 69, 2023.

[83] Y. Maraş, A. Kor, S. Duran, K. Orhan, E. Atalar, E. S. Sözen, and H. H. Maraş, "Using ensemble learning and feature selection in the diagnosis of low back pain," *Rheumatology Quarterly*, vol. 2, no. 3, pp. 121–129, 2024.

[84] M. Wang, B. Zhuang, S. Yu, and G. Li, "Ensemble learning enhances the precision of preliminary detection of primary hepatocellular carcinoma based on serological and demographic indices," *Frontiers in Oncology*, vol. 14, p. 1397505, 2024.

[85] S. T. Bhuiyan, R. Rahman, S. B. Alam, H. Khatun, R. Islam, S. Wasi, M. A. R. Iftee, and A. Islam, "Patch-based deep ensemble learning for enhancing pelvic fracture detection," in *2025 28th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2025, pp. 3022–3027.

[86] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, pp. 336–359, 2020.

[87] D. Alzu'bi *et al.*, "Kidney tumor detection and classification based on deep learning approaches: a new dataset in ct scans," *Journal of Healthcare Engineering*, vol. 2022, p. 386116, 2022.

[88] S. Wasi, S. B. Alam, R. Rahman, M. A. Amin, and S. Kobashi, "Kidney tumor recognition from abdominal CT images using transfer learning," in *IEEE 53rd International Symposium on Multiple-Valued Logic (ISMVL)*, 2023, pp. 54–58.

[89] M. Liu *et al.*, "A deep learning method for breast cancer classification in pathology images," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, pp. 5025–5032, 2022.

[90] R. Sarki, K. Ahmed, H. Wang, and Y. Zhang, "Automated detection of mild and multi-class diabetic eye diseases using deep learning," *Health Information Science and Systems*, vol. 8, p. 32, 2020.

[91] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2021.

[92] J. Wasserthal *et al.*, "TotalSegmentator: robust segmentation of 104 anatomic structures in CT images," *Radiology: Artificial Intelligence*, vol. 5, p. e230024, 2023.

[93] J. Chen *et al.*, "TransUNet: transformers make strong encoders for medical image segmentation," 2021.

[94] J. Ma *et al.*, "Segment anything in medical images," *Nature Communications*, vol. 15, p. 654, 2024.

[95] S. T. Bhuiyan, R. Rahman, S. Wasi, H. Khatun, M. S. Chowdhury, R. Islam, S. K. Mazumder, N. Yagi, S. Kobashi, and S. B. Alam, "Beyond the pelvic ring: benchmarking multi-bone segmentation for roi-guided fracture detection," *Scientific Reports*, vol. 16, no. 1, p. 24508, 2026.

[96] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," pp. 234–241, 2015.

[97] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[98] S. T. Bhuiyan, R. Rahman, R. Islam, M. Haque, M. Islam, S. Kobashi, and S. B. Alam, "Lung disease diagnosis from CXR using convolutional neural network (CNN): a comparative study," *Soft Computing*, vol. 29, no. 23, pp. 6333–6345, 2025.

[99] A. A. A. Shanto, M. Esha, S. A. Kotha, R. Islam, S. Wasi, S. T. Bhuiyan, R. Rahman, and S. B. Alam, "Assessing the impact of architectural design in deep learning models for fundus-based ocular disease detection," in *IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS)*, 2025, pp. 1–6.

[100] M. A. Mahtab, S. Tasnim, M. S. Choudhury, S. Wasi, S. T. Bhuiyan, S. B. Alam, and R. Rahman, "SensaNet: a lightweight DL model for tuberculosis detection in histopathological images," in *22nd International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, 2025, pp. 1–6.

[101] S. K. Mazumder, R. Islam, S. T. Bhuiyan, R. Rahman, S. Wasi, A. Islam, and S. B. Alam, "Detection-guided lumbar spine degeneration classification using multi-slice encoded MRI," in *International Conference on Machine Learning and Cybernetics (ICMLC)*, 2025, pp. 75–80.

[102] 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.

[103] S. T. Bhuiyan, R. Rahman, S. Wasi, N. Yagi, S. Kobashi, A. Islam, and S. B. Alam, "Invisible yet detected: PelFANet with attention-guided anatomical fusion for pelvic fracture diagnosis," in *Workshop on Empowering Medical Image Computing and Research through Early-Career Expertise*. Springer, 2025, pp. 103–113.

[104] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[105] A. M. Tahir, M. E. H. Chowdhury, A. Khandakar, T. Rahman, Y. Qiblawey, U. Khurshid, S. Kiranyaz, N. Ibtehaz, M. S. Rahman, S. Al-Maadeed *et al.*, "COVID-19 infection localization and severity grading from chest X-ray images," *Computers in Biology and Medicine*, vol. 139, p. 105002, 2021.

[106] S. T. Bhuiyan, Z. Zaman, R. Rahman, M. S. Choudhury, S. Wasi, A. Islam, S. B. Alam, and S. Kobashi, "Chexnet-cbam: Enhancing large-scale medical pre-training with attention for lung disease diagnosis," in *2026 IEEE 56th International Symposium on Multiple-Valued Logic (ISMVL)*. IEEE, 2026, pp. 105–110.

[107] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al Emadi *et al.*, "Can AI help in screening viral and COVID-19 pneumonia?" *IEEE Access*, vol. 8, pp. 132 665–132 676, 2020.

[108] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. A. Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughaier, M. S. Khan, and M. E. H. Chowdhury, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Computers in Biology and Medicine*, vol. 132, p. 104319, 2021.

[109] T. Rahman, A. Khandakar, M. A. Kadir, K. R. Islam, K. F. Islam, R. Mazhar, T. Hamid, M. T. Islam, S. Kashem, Z. B. Mahbub, M. A. Ayari, and M. E. H. Chowdhury, "Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization," *IEEE Access*, vol. 8, pp. 191 586–191 601, 2020.

[110] S. T. Bhuiyan, R. Rahman, S. Wasi, R. Islam, S. Kobashi, A. Islam, and S. B. Alam, "Secondopinion: Anatomy-aware gated reasoning for efficient medical image analysis," *arXiv preprint arXiv:2608.01808*, 2026.