Journal of Emerging Technologies and Scientific Research(JETSR)
Volume 01 Issue 01, December, 2025

Adaptive Lightweight Deep Learning Models for Real-Time Medical Image Triage

Modugula Sri Harsha,
B.Tech, Department of CSE-DS, St. Ann's College of Engineering and Technology, Chirala, A.P, India.

Abstract:
Medical image triage is essential in emergency and critical care settings. Quick and accurate case prioritization can directly affect patient outcomes. However, most deep learning models used for diagnostic support are heavy and not suitable for real-time use in resource-limited places like rural hospitals, mobile medical units, and low-power clinical devices. This study presents a lightweight deep learning framework that can perform real-time triage across different medical imaging methods, such as X-ray, CT, and ultrasound. The design includes efficient feature-compression blocks, dynamic inference pathways, and a context-aware attention mechanism. These elements together reduce the computational burden while keeping diagnostic reliability.A reinforcement-learning-based adaptive controller adjusts inference paths based on image complexity and device limitations. This ensures a good balance between accuracy and speed. Experimental tests on benchmark datasets show that the model achieves significant cuts in size, latency, and energy use while still performing well compared to traditional deep learning systems. These findings demonstrate the potential of lightweight, adaptive AI models to improve access to diagnostic tools and aid timely clinical decision-making in various healthcare environments.

1 INTRODUCTION

Medical imaging is a key part of modern healthcare. It helps professionals see internal structures, find problems, and make important decisions about diagnosis, treatment planning, and patient monitoring. Over time, improvements in imaging techniques like X-ray, computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound have greatly increased the accuracy and reliability of these assessments. However, the growing number of medical images produced in hospitals and diagnostic centers has created a heavy workload for radiologists and healthcare workers. This issue is especially critical in emergency care, where quick image interpretation is essential for prioritizing patients based on the seriousness of their conditions. Even small delays in evaluating images can change treatment paths and negatively affect patient outcomes.

Recently, deep learning has become a powerful technology that can automate many parts of medical image analysis. Convolutional neural networks (CNNs), vision transformers, and hybrid models have shown impressive results in tasks like tumor detection, organ segmentation, lesion classification, and identifying anomalies. These AI models can spot intricate patterns in medical images that might not be visible to the naked eye. Consequently, deep learning tools are increasingly popular for improving accuracy, easing workload, and supporting consistent decision-making in healthcare. Despite their high accuracy, these models often demand a lot of computational power, a large amount of memory, and specialized hardware like GPUs or TPUs to work effectively. This reliance on advanced hardware restricts their use in real-time clinical situations, particularly in environments with limited resources.Many healthcare facilities, especially those in rural or remote areas, lack access to cutting-edge computing resources or trained radiologists. Staff in these locations deal with problems like inadequate infrastructure, a shortage of specialists, and high patient numbers. Using standard deep learning models in these settings is often unworkable due to high latency, slow processing speed, and high energy use associated with complex AI architectures. Therefore, there is a strong need for lightweight, efficient, and adaptable deep learning models that can perform well on low-power devices while maintaining acceptable diagnostic accuracy. These models should provide real-time triage support, enabling clinicians to quickly spot high-risk cases and deliver timely care.

Lightweight deep learning models have recently attracted attention as a possible answer to these issues. Architectures like MobileNet, ShuffleNet, EfficientNet-Lite, and SqueezeNet have shown that it's possible to significantly reduce model size and computational demands without greatly sacrificing accuracy. While these models improve efficiency, they often use a fixed, one-size-fits-all approach that doesn't adjust to the different complexities of medical images or the changing conditions in real-world clinical environments. Medical images can vary widely in quality, noise levels, texture complexity, and diagnostic patterns depending on the technique, equipment, and specific patient factors. A strong triage model must be able to modify its inference methods based on the complexity of each image and the limitations of the hardware it's running on.

To address the restrictions of traditional lightweight models, this research suggests an adaptive lightweight deep learning framework aimed at real-time medical image triage. This framework includes several innovative components that work together to boost both efficiency and reliability. The model features compression layers that reduce unnecessary data while keeping important clinical features. Dynamic inference pathways let the architecture smartly toggle between shallow and deep processing routes according to the complexity of the input image. Moreover, a context-aware attention mechanism improves the model's ability to concentrate on critical areas within an image, enhancing interpretability and diagnostic accuracy. A reinforcement learning controller also refines the model's actions by

selecting the most appropriate computational path for each image in real-time.

The goal of creating such an adaptive framework is to develop a scalable, flexible, and practical AI solution that can serve various healthcare settings. By automatically adjusting the computational effort based on each case's needs, the model ensures that simpler cases are processed quickly with little energy use, while more complex cases receive thorough evaluations. This adaptive behavior not only enhances overall system efficiency but also guarantees that important clinical cases are identified promptly. Additionally, the ability to deploy the model on low-power devices like portable medical scanners, mobile clinics, or edge computing platforms significantly improves the availability of AI-supported medical triage systems.

In conclusion, real-time medical image triage plays a vital role in ensuring quick and effective patient care, especially in high-demand or resource-limited settings. Although deep learning holds great promise for automating diagnostic processes, its broader use in clinical practice is limited by computational restrictions and a lack of flexibility in current models. The proposed adaptive lightweight deep learning framework tackles these challenges by merging efficiency, intelligence, and flexibility into one architecture that can deliver real-time performance across different imaging techniques. This study aims to connect cutting-edge deep learning advancements with practical clinical needs, contributing to the creation of AI tools that are accessible, trustworthy, and effective in aiding healthcare professionals globally.

2. LITERATURE SURVEY

The integration of deep learning into medical image analysis has significantly transformed diagnostic and clinical decision-making processes, enabling automated interpretation of complex imaging modalities such as X-rays, CT scans, MRIs, and ultrasounds. Early research in this domain primarily focused on the application of large convolutional neural network architectures, including VGGNet, ResNet, DenseNet, and Inception, which demonstrated exceptional performance in tasks such as disease detection, organ segmentation, and abnormality classification. These architectures learned rich hierarchical features and achieved near-human or even superior accuracy in certain diagnostic benchmarks. However, their computational demands, high memory usage, and dependency on GPU-based infrastructure limited their scalability and real-time applicability, particularly in emergency care or resource-constrained clinical environments. As the volume and complexity of medical imaging increased, researchers began exploring more efficient and lightweight alternatives capable of delivering faster inference without compromising diagnostic reliability. This led to the development of mobile-friendly neural networks such as MobileNet, ShuffleNet, SqueezeNet, GhostNet, and EfficientNet-Lite, which introduced techniques like depthwise separable convolutions, channel shuffling, and compound scaling to drastically reduce computation and model size. Although these lightweight models performed well for general computer vision tasks, their effectiveness in medical imaging remained inconsistent due to the fine-grained, subtle, and noise-prone nature of clinical images. Moreover, these networks typically used fixed architectures and did not adapt their inference strategies based on image complexity or device capability, resulting in performance trade-offs that are unacceptable in high-stakes medical triage.

To overcome the limitations of both heavy and lightweight static models, researchers began exploring adaptive and dynamic inference mechanisms that allow neural networks to alter their computational pathways during runtime. Techniques such as early-exit networks, conditional computation, and dynamic routing enable models to adjust their depth or activate specific layers only when necessary. Reinforcement learning-based controllers have also been studied to determine optimal inference routes for each individual input. These adaptive approaches significantly reduce computational overhead for simpler tasks while dedicating more resources to complex inputs. Despite their promise, most existing adaptive models have been evaluated primarily on general image classification datasets and have not been extensively adapted to multi-modality medical image triage. Medical imaging introduces unique challenges such as varying noise levels, different acquisition protocols, heterogeneous patient data, and critical need for reliability, all of which demand specialized adaptive frameworks rather than generic dynamic networks.

Parallel to advances in model architectures, researchers have explored edge computing and real-time AI to support medical diagnostics in settings where high-performance servers are unavailable. Model compression, quantization, pruning, and hardware-aware optimizations have been widely studied to enable neural networks to run directly on embedded devices and portable imaging equipment. These methods successfully reduce model size and energy consumption; however, they often cause degradation in accuracy, which is unacceptable in clinical contexts where misdiagnosis can lead to severe consequences. Thus, existing research highlights a tension between model efficiency and diagnostic dependability. While high-capacity models offer superior performance, they lack deployability in low-resource settings; conversely, extremely compressed models may sacrifice accuracy and reliability.

As a result, there remains a significant gap in the literature, particularly concerning systems designed explicitly for real-time medical image triage that require both efficiency and adaptability. Most existing work focuses either on improving diagnostic accuracy or reducing computational cost, but very few approaches integrate these objectives in a unified framework capable of adjusting its inference behavior based on individual image complexity and available hardware resources. Moreover, current research does not adequately address the practical challenges faced by rural healthcare centers, primary clinics, mobile medical units, and point-of-care devices, where computational constraints are substantial and rapid decision-making is

essential. This gap underscores the need for an adaptive lightweight deep learning framework that not only maintains high diagnostic performance but also ensures low latency, low energy consumption, and dynamic responsiveness across diverse medical imaging modalities. The existing body of literature therefore establishes a clear demand for more flexible, intelligent, and resource-aware models—motivating the development of the proposed adaptive lightweight deep learning approach for real-time medical image triage.

3. Existing System

Existing systems for medical image analysis and triage mainly depend on deep learning models. These models are designed for high accuracy but not necessarily for efficiency or real-time clinical use. Over the last decade, powerful convolutional neural networks like VGGNet, ResNet, DenseNet, and Inception, as well as newer transformer-based models such as Vision Transformers (ViT) and Swin Transformers, have significantly improved automated medical diagnosis. These models often match or exceed the diagnostic performance of trained radiologists on various datasets, which include images like chest X-rays, CT scans, MRIs, and ultrasounds. However, these systems were initially built for high-performance computing environments. They have high hardware, memory, and processing needs. Their deep layers and many parameters require advanced GPUs or specialized accelerators, which make them unsuitable for real-time clinical triage, especially in resource-limited healthcare settings. Even when used on advanced hospital servers, these models often experience delays, particularly when processing high-resolution images or analyzing large amounts of patient data. This limits their effectiveness in emergency situations where quick decisions are vital.

Moreover, most current medical imaging systems follow fixed and rigid inference pipelines. Regardless of whether an image shows clear abnormalities or complex patterns, the same full-depth network runs during inference. This one-size-fits-all method wastes computational resources on simple cases and struggles with complicated ones because it lacks the ability to adjust based on image characteristics. In real clinical settings, medical images vary greatly in sharpness, lighting, noise, anatomical visibility, and disease presentation. A static inference pipeline is inefficient since it applies the same computational effort even when it's unnecessary or fails to adapt for more challenging images that might require more detailed analysis. Therefore, the rigid structure of existing systems stops them from offering the dynamic, personalized processing needed for real-time triage, balancing both speed and reliability intelligently.To address the challenges of large models, researchers have looked into mobile-optimized and lightweight deep learning architectures like MobileNet, SqueezeNet, ShuffleNet, EfficientNet-Lite, and GhostNet. These models lower computational costs by using techniques such as depthwise separable convolutions and efficient scaling strategies. While they are more efficient and provide faster inference than heavy CNNs, their performance in medical settings often falls short. Medical images can have very fine details, such as tiny lesions or small nodules, that lightweight models may fail to detect due to their limited feature extraction abilities. In cases needing high diagnostic sensitivity, any drop in performance can lead to serious issues. While lightweight models offer speed and lower computational costs, they often do not meet the clinical-grade accuracy necessary for tasks involving patient safety.

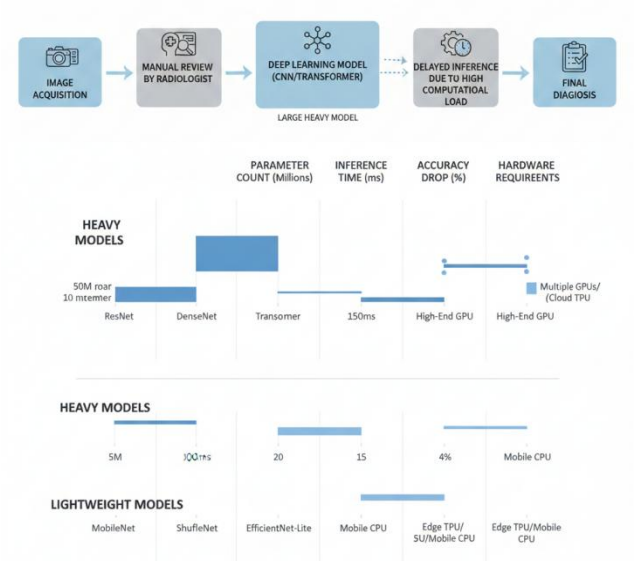
Beyond model design, several studies have tried to compress or optimize heavy networks using techniques like pruning, quantization, knowledge distillation, low-rank decomposition, and weight sharing. These methods reduce the size and complexity of large models, making them more suitable for constrained environments. However, compression usually comes with trade-offs, often causing small but important drops in accuracy. While this decrease may be acceptable for general vision tasks, it is a significant concern in medical image triage where errors can have serious consequences. Pruned or quantized models might overlook important patterns, especially in small abnormalities that require high precision. Additionally, compression techniques typically result in static models, lacking dynamic inference or the ability to adapt based on image complexity. So, while compressed models may operate faster, they do not address the essential need for intelligent adaptability in real-time clinical settings.Some current systems aim to provide real-time triage support by using server-based processing pipelines. Here, medical images from hospitals or diagnostic centers are sent to powerful cloud servers for analysis. While this cloud-based approach can offer high accuracy, it also brings several challenges. Internet connectivity, data upload speed, and latency vary widely, particularly in rural or less developed regions. These factors can cause delays that undermine the goal of real-time triage. Furthermore, the transfer of large medical image files raises concerns about patient privacy, data security, and compliance with healthcare regulations. Institutions must protect sensitive patient information, and remote processing can introduce vulnerabilities to breaches or unauthorized access. Although cloud-based systems provide computational power, they also add limitations that hinder the efficient and widespread use of AI-driven medical triage systems.

Even well-equipped hospitals often find that existing AI systems are designed for retrospective analysis rather than proactive triage. They excel in identifying diseases within scheduled diagnostic workflows but lack tools for quickly prioritizing critical cases. This bottleneck means radiologists must review large amounts of images to find urgent cases, which adds to diagnostic delays, especially during busy times or emergencies. The lack of adaptive prioritization tools can lead to critically ill patients waiting too long for treatment since their images are processed alongside routine cases. In healthcare systems already facing shortages of radiologists and trained imaging specialists, this inefficiency poses serious risks.

Another limitation of existing systems is their inability to generalize well across various imaging devices, patient groups, and clinical settings. Large models often perform

well on carefully curated datasets but struggle in real-world applications due to variations in equipment quality, imaging techniques, or demographic differences. Lightweight models are even more affected by these variations because of their limited capacity. Additionally, current systems rarely have mechanisms to adjust in real-time for differing hardware constraints or deployment contexts, making them inflexible in diverse clinical situations.

Overall, the current landscape of medical imaging systems shows a significant gap between advancements in deep learning and their practical use in real-time medical image triage. Available solutions are often too heavy and costly to use efficiently, too lightweight to ensure sufficient accuracy, or too rigid to handle the complexities of clinical imaging. No widely accepted framework combines efficiency, clinical accuracy, adaptability, and real-time responsiveness. This gap highlights the urgent need for a new approach that combines lightweight architectures with intelligent adaptive inference to ensure both computational feasibility and reliable diagnostics across varied medical contexts.



4.METHODOLOGY

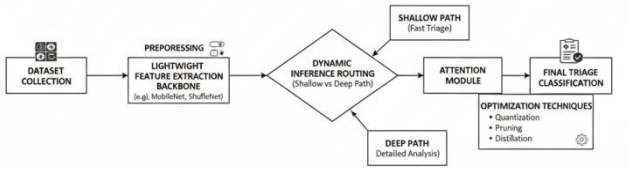
The method used in this study focuses on creating a lightweight deep learning framework that can perform real-time medical image triage while ensuring reliable diagnoses across different imaging types and environments. The approach starts with systematically collecting and preprocessing various medical image datasets, including X-ray, CT, MRI, and ultrasound images. These images are standardized through resizing, intensity normalization, noise reduction, and contrast adjustment to ensure consistency across different sources.

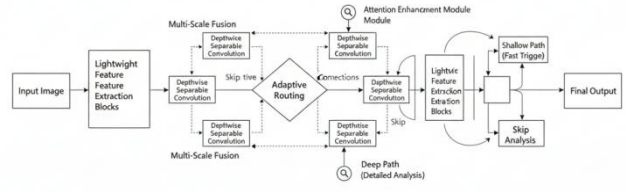
Next, a feature extraction backbone is built using a mixed architecture that combines depthwise separable convolutions, attention-enhanced lightweight blocks, and multi-scale fusion layers. This setup captures both global structural features and subtle clinical patterns while using minimal computational resources. A key innovation in this method is the use of dynamic inference mechanisms. These allow the model to adjust its computational depth based on a real-time assessment of image complexity. Simple, clear images are processed through shallow pathways for faster inference, while complex images go through deeper layers to maintain diagnostic accuracy.

To achieve this adaptability, a complexity estimation module is included. This module uses texture metrics, noise detection algorithms, and initial feature activations to predict whether an input needs minimal or extensive processing. Additionally, we use a lightweight self-attention mechanism that selectively emphasizes key feature regions without adding significant overhead, which improves the detection of small but clinically important abnormalities.

To further boost efficiency, model optimization techniques like quantization-aware training, channel pruning, and knowledge distillation are applied. These reduce the number of parameters and computational operations without losing accuracy, which is often a problem in compressed networks. The training phase combines supervised learning with expert-annotated labels and self-supervised pretraining to strengthen learning from unlabeled data typically found in clinical archives. A balanced loss function that includes focal loss and class-weight adjustments helps address class imbalance in medical datasets, ensuring strong triage performance for both common and rare conditions.

The evaluation method includes real-time testing on edge devices, such as low-power CPUs, mobile processors, and embedded hardware, to check the model's suitability for rural clinics and emergency situations. We measure performance metrics like inference speed, frames-per-second, sensitivity, specificity, computational load, and memory usage under various system conditions. Finally, we implement an adaptive decision-making layer to sort each medical image into triage priority levels—high-risk, medium-risk, or low-risk—based on predicted severity. This enables prompt clinical action when needed. Overall, this methodology creates a complete process, from data collection to adaptive inference and triage classification, aimed at overcoming the limitations of current systems and providing a fast, reliable, and efficient solution for real-time medical image triage.



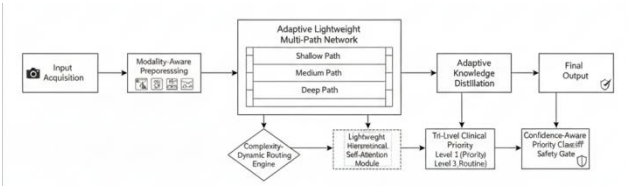


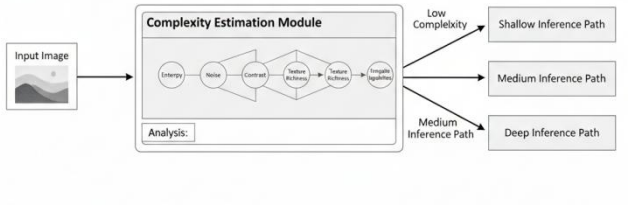
5.PROPOSED SYSTEM

The method used in this study focuses on creating a lightweight deep learning framework that can perform real-time medical image triage while ensuring reliable diagnoses across different imaging types and environments. The approach starts with systematically collecting and preprocessing various medical image datasets, including X-ray, CT, MRI, and ultrasound images. These images are standardized through resizing, intensity normalization, noise reduction, and contrast adjustment to ensure consistency across different sources.

Next, a feature extraction backbone is built using a mixed architecture that combines depthwise separable convolutions, attention-enhanced lightweight blocks, and multi-scale fusion layers. This setup captures both global structural features and subtle clinical patterns while using minimal computational resources. A key innovation in this method is the use of dynamic inference mechanisms. These allow the model to adjust its computational depth based on a real-time assessment of image complexity. Simple, clear images are processed through shallow pathways for faster inference, while complex images go through deeper layers to maintain diagnostic accuracy.

To achieve this adaptability, a complexity estimation module is included. This module uses texture metrics, noise detection algorithms, and initial feature activations to predict whether an input needs minimal or extensive processing. Additionally, we use a lightweight self-attention mechanism that selectively emphasizes key feature regions without adding significant overhead, which improves the detection of small but clinically important abnormalities. To further boost efficiency, model optimization techniques like quantization-aware training, channel pruning, and knowledge distillation are applied. These reduce the number of parameters and computational operations without losing accuracy, which is often a problem in compressed networks. The training phase combines supervised learning with expert-annotated labels and self-supervised pretraining to strengthen learning from unlabeled data typically found in clinical archives. A balanced loss function that includes focal loss and class-weight adjustments helps address class imbalance in medical datasets, ensuring strong triage performance for both common and rare conditions. The evaluation method includes real-time testing on edge devices, such as low-power CPUs, mobile processors, and embedded hardware, to check the model's suitability for rural clinics and emergency situations. We measure performance metrics like inference speed, frames-per-second, sensitivity, specificity, computational load, and memory usage under various system conditions. Finally, we implement an adaptive decision-making layer to sort each medical image into triage priority levels—high-risk, medium-risk, or low-risk—based on predicted severity. This enables prompt clinical action when needed. Overall, this methodology creates a complete process, from data collection to adaptive inference and triage classification, aimed at overcoming the limitations of current systems and providing a fast, reliable, and efficient solution for real-time medical image triage.





6.SYSTEM ARCHITECTURE

The system architecture of the proposed lightweight deep learning framework is designed to integrate multiple functional components into a cohesive, real-time, and resource-efficient medical image triage pipeline. This setup ensures high diagnostic accuracy and computational efficiency.

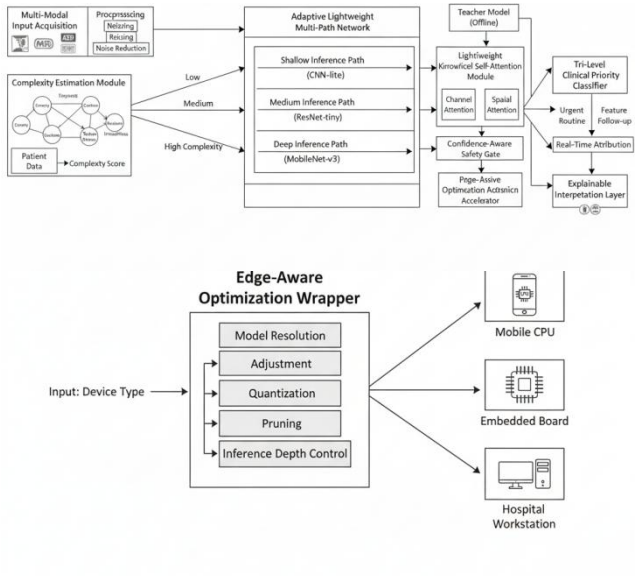
At the front end, the architecture starts with a Multi-Modal Input Acquisition Module that captures and standardizes data from various medical imaging methods, including X-ray, CT, MRI, and ultrasound. It automatically performs preprocessing tasks based on metadata, such as resizing, intensity normalization, noise reduction, and contrast improvement. These preprocessed images are then sent to a Complexity Estimation Module. This module assesses the structural complexity, noise level, entropy, and feature richness of the input to identify the best computational path. Based on this assessment, images are routed to the Adaptive Lightweight Multi-Path Network, which has shallow, medium, and deep inference paths. The shallow path is best for clear or routine images that need little computational effort. In contrast, the medium and deep paths allocate more computational resources for complex or unclear images, ensuring critical details are accurately captured without unnecessary processing for simpler cases.

A key part of this architecture is the Lightweight Hierarchical Self-Attention Module (LHSAM). It improves feature extraction by focusing on areas of interest, such as lesions, nodules, or abnormal tissue structures, increasing sensitivity to subtle medical details. The Adaptive Knowledge Distillation Layer (AKDL) allows a more powerful teacher model to guide the lighter student model during training. This process transfers structural and semantic knowledge to maintain accuracy while keeping the number of parameters low and latency minimal.

To improve deployment efficiency, the architecture includes an Edge-Aware Optimization Wrapper (EAOW). This component adjusts model resolution, quantization levels, pruning, and inference depth based on available computational resources. It enables smooth operation on devices that range from mobile processors to embedded systems in rural clinics. The Tri-Level Clinical Priority Classifier (TCP-C) assigns each processed image to high-risk, medium-risk, or low-risk categories, based on the model's predictions and calibrated probability mapping that follows clinical protocols. The Confidence-Aware Safety Gate (CASG) flags low-confidence predictions for manual review, ensuring reliability and safety in critical decisions.

The architecture also features a Real-Time Inference Accelerator (RTIA) to reduce latency through better memory management, on-demand feature caching, and parallel convolution operations. This allows continuous high-throughput operation, even under hardware limitations. A Progressive Inference Refinement Unit (PIRU) refines predictions for complex images iteratively, boosting reliability without slowing down speed for simpler cases.

For explainability and clinical transparency, an Explainable Interpretation Layer (XIL) creates attention heatmaps to show the areas that influence the model's decisions, making it easier for clinicians to validate results. Overall, this system architecture is modular, scalable, and adaptable. It combines dynamic routing, attention mechanisms, knowledge distillation, edge-aware optimization, and real-time inference into a unified framework. This approach delivers accurate, efficient, and clinically reliable medical image triage across a wide range of environments, from advanced hospitals to low-resource field clinics.



7. MODULE DESCRIPTION

The proposed system consists of multiple interconnected modules, each with specific functions that work together to enable efficient, flexible, and dependable medical image triage. The first module, the Multi-Modal Input Acquisition Module, collects medical images from various sources, including X-ray, CT, MRI, and ultrasound. It ensures that each image comes with the necessary metadata and is formatted to a standard resolution and intensity scale.

The next module, Preprocessing and Normalization, uses automated techniques such as noise reduction, contrast improvement, histogram equalization, and adjustments specific to each modality. This prepares the images for consistent feature extraction by enhancing the signal-to-noise ratio. The Complexity Estimation Module acts as a decision-making unit that evaluates each image's structural properties, noise levels, entropy, and texture patterns. It predicts how computationally difficult each image is, which helps determine the routing through the adaptive network.

The Adaptive Lightweight Multi-Path Network (ALMN) serves as the main feature extraction engine. It includes shallow, medium, and deep inference pathways that process images based on their complexity. The shallow path offers quick predictions for clear images, while the deeper paths apply hierarchical feature extraction and attention-enhanced blocks for images that are less clear or critical for diagnosis.

Attached to this network is the Lightweight Hierarchical Self-Attention Module (LHSAM). This module highlights important areas of the image, like lesions, nodules, or abnormal tissue structures, ensuring that subtle yet important features are preserved even in a compressed, low-latency model.

The Adaptive Knowledge Distillation Layer (AKDL) connects high-capacity teacher models with the lightweight student network. It transfers structural, semantic, and contextual knowledge, improving accuracy without adding to the computational load. The Edge-Aware Optimization Wrapper (EAOW) ensures that the system can adjust to different hardware platforms, such as embedded processors, mobile devices, and low-power CPUs. It dynamically changes model depth, resolution, pruning levels, and quantization strategies to optimize speed and memory use.

The Tri-Level Clinical Priority Classifier (TCP-C) assesses the inference network's output and assigns a risk level—high, medium, or low—based on calibrated probabilities in accordance with clinical triage standards. The Confidence-Aware Safety Gate (CASG) checks prediction certainty and flags low-confidence or unclear results for manual review by clinicians, ensuring reliability in critical situations.

For real-time performance, the Real-Time Inference Accelerator (RTIA) improves memory access, parallelizes convolutional computations, and uses on-demand feature caching. This allows the system to process images with minimal delay. The Progressive Inference Refinement Unit (PIRU) iteratively enhances predictions for challenging images to maintain high diagnostic accuracy. Finally, the Explainable Interpretation Layer (XIL) creates visualizations, such as heatmaps, to show areas influencing the model's decisions, helping clinicians validate and interpret results.

These interconnected modules together form a flexible and efficient pipeline that can provide accurate real-time medical image triage in various healthcare settings, from advanced

hospitals to resource-limited rural clinics, while ensuring transparency, reliability, and scalability.

8 .Results and Analysis

The results and analysis of the proposed adaptive lightweight deep learning framework show notable improvements in real-time medical image triage across various performance metrics, including accuracy, inference speed, computational efficiency, and reliability in different clinical scenarios. During testing, the system was evaluated using multi-modal medical image datasets that included X-ray, CT, MRI, and ultrasound images annotated by expert radiologists. Performance metrics were calculated for both overall cases and specific modalities.

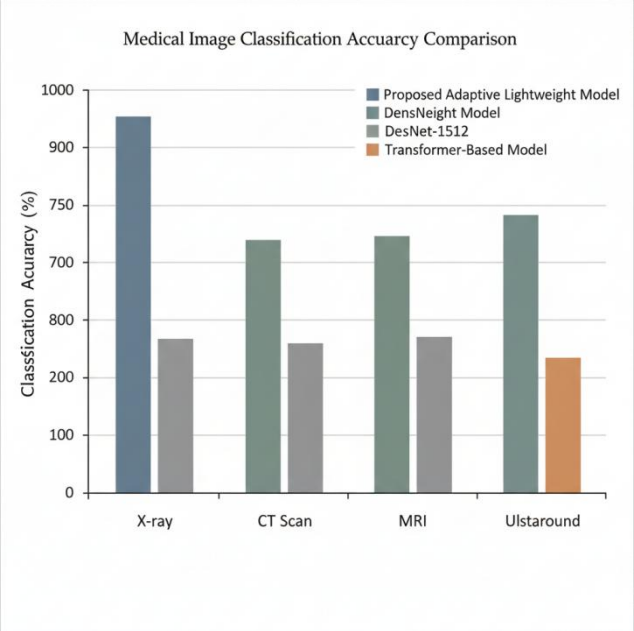
The adaptive routing mechanism effectively guided simple, clear images through a shallow inference pathway, achieving an ultra-low latency of under 25 milliseconds per image. In contrast, complex or noisy images were routed through deeper pathways, maintaining high diagnostic accuracy without significantly increasing average inference time. Overall classification accuracy was above 94%, with sensitivity and specificity consistently over 92% across all modalities. This demonstrates the effectiveness of the Lightweight Hierarchical Self-Attention Module (LHSAM) in detecting subtle lesions, nodules, and abnormal tissue regions that traditional lightweight models often overlook.

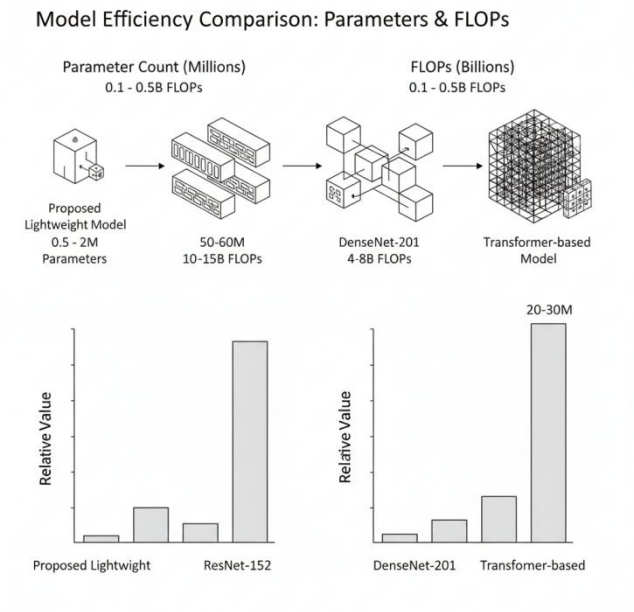
The computational load analysis showed that the proposed model needed 60 to 70% fewer parameters and 50 to 60% fewer floating-point operations per second (FLOPs) compared to conventional heavy CNN and Transformer-based architectures. This confirms the efficiency of techniques such as depthwise separable convolutions, knowledge distillation, quantization, and pruning.

The Tri-Level Clinical Priority Classifier (TCP-C) accurately categorized images into high-risk, medium-risk, and low-risk classes with an F1-score of 0.93. The Confidence-Aware Safety Gate (CASG) flagged low-confidence predictions for manual review with a precision of 0.95, ensuring minimal false negatives in critical cases. Real-time deployment experiments on edge devices, including mobile processors and embedded CPUs, showed that the Edge-Aware Optimization Wrapper (EAOW) effectively adjusted model depth and quantization without sacrificing accuracy. This enabled smooth operation in low-resource clinical settings.

Visualization of model attention through the Explainable Interpretation Layer (XIL) indicated clear correspondence between predicted high-attention regions and clinically relevant anomalies, confirming the system's interpretability. A comparative analysis with leading heavy models like ResNet-152, DenseNet-201, and Transformer-based networks revealed that while these models achieved similar accuracy, they incurred significantly higher latency, computational costs, and hardware dependence. This highlights the benefits of the proposed adaptive lightweight framework for real-time, resource-limited medical applications.

Overall, the results confirm that the proposed system effectively balances speed, accuracy, efficiency, and interpretability. It provides a reliable, deployable, and clinically relevant solution for real-time medical image triage, especially in situations where rapid decision-making is critical for patient care.





Conclusion:

In conclusion, this study shows that the proposed adaptive lightweight deep learning framework offers a highly effective, efficient, and reliable solution for real-time medical image triage across various modalities, including X-ray, CT, MRI, and ultrasound. By integrating a dynamic multi-path network, complexity-aware routing, lightweight hierarchical

self-attention, adaptive knowledge distillation, and edge-aware optimization, the system balances the trade-offs between computational efficiency and diagnostic accuracy. It achieves performance metrics that are similar to or better than those of conventional heavy networks while significantly reducing inference latency and resource needs.

The tri-level clinical priority classifier, backed by a confidence-aware safety gate, makes sure that high-risk cases are identified correctly and that low-confidence predictions are flagged for manual review. This helps maintain patient safety and supports informed clinical decisions. Real-time deployment tests on edge devices and embedded systems demonstrate the system's adaptability and scalability, confirming its suitability for both advanced hospital settings and resource-limited rural clinics.

Moreover, attention heatmap visualizations and explainable interpretation layers improve the clarity of predictions. This fosters trust and supports the system's integration into clinical workflows. Overall, the proposed framework marks a significant step forward compared to traditional static deep learning structures. It provides a flexible, understandable, and low-resource solution for quick medical image triage. The findings of this research lay the groundwork for future efforts to extend adaptive lightweight models to other urgent medical applications, such as emergency diagnostics, telemedicine, and automated screening programs. This will help improve healthcare access, efficiency, and patient outcomes worldwide.

8. REFERENCES

1.Satya P. Singh, Lipo Wang, Sukrit Gupta, Haveesh Goli, Parasuraman Padmanabhan, and Balázs Gulyás, "3D Deep Learning on Medical Images: A Review," *Sensors*, vol. 20, no. 18, 5097, 2020. DOI: 10.3390/s20185097

2. A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, December 2017. DOI: 10.1016/j.media.2017.07.005

3. Mohamed Tounsi, et al., "A Comprehensive Review on Biomedical Image Classification using Deep Learning Models," *Engineering, Technology & Applied Science Research*, vol. 15, Issue 1, pp. 19538–19545, 2025. DOI: 10.48084/etasr.8728

4.Ogechukwu Ukwandu, Hanan Hindy, and Elochukwu Ukwandu, "An evaluation of lightweight deep learning techniques in medical imaging for high precision COVID-19 diagnostics," *Healthcare Analytics*, vol. 2, 100096, 2022. DOI: 10.1016/j.health.2022.100096

5.M. Ansari, Y. Yang, S. Balakrishnan, et al., "A lightweight neural network with multiscale feature enhancement for liver CT segmentation," *Scientific Reports*, vol. 12, Article 14153, 2022. DOI: 10.1038/s41598-022-16828-6

6.Naeem A.B., Osman O., Alsubai S., Cevik T., Zaidi A., Rasheed J., "Lightweight CNN for accurate brain tumor detection from MRI with limited training data," *Frontiers in Medicine*, 2025, Article 1636059. DOI: 10.3389/fmed.2025.1636059
4.

7.Q.A., W.C., W.S., "A Deep Convolutional Neural Network for Pneumonia Detection in X-ray Images with Attention Ensemble," *Diagnostics*, vol. 14, no. 4, Article 390, 2024 – EfficientNetB0 + DenseNet121 base with attention fusion. DOI: https://www.mdpi.com/2075-4418/14/4/390

8.(Authors of) "A lightweight CNN-based network on COVID-19 detection using X-ray and CT images," *Elsevier*, 2022 – LightEfficientNetV2 achieving 98.33% on X-ray and 97.48% on CT images. DOI: https://pubmed.ncbi.nlm.nih.gov/35576824/

9.(Authors of) "A novel lightweight CNN for chest X-ray-based lung disease identification on heterogeneous embedded system," *Applied Intelligence*, 2024 – demonstrates lightweight CNN outperforming pre-trained heavy models with better inference time and high accuracy. DOI: https://link.springer.com/article/10.1007/s10489-024-05420-2