I. J. Engineering and Manufacturing, **2026, 5, 255-271**
Published Online on October 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2026.05.14

Multi-Resolution Attention-Fused, Calibration-Aware Diagnosis of Kidney CT (Normal/Cyst/Stone/Tumour)

Shanker M. C.
Department of Biomedical Engineering, VelTech MultiTech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, Tamil Nadu, India
Email: shanker@veltechmultitech.org
ORCID iD: https://orcid.org/0009-0005-2538-8704

N. Sankar Ram
Department of CSE-AIML, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India
Email: nsankarram@gmail.com
ORCID iD: https://orcid.org/0000-0002-9330-2459

V. Gokula Krishnan*
Department of Computer Science and Engineering, Easwari Engineering College, Chennai, Tamil Nadu, India
Email: gokul_kris143@yahoo.com
ORCID iD: https://orcid.org/0009-0005-6819-6729
*Corresponding Author

Pinagadi Venkateswararao
Department of CSE, CVR College of Engineering, Hyderabad, Telangana, India
Email: drrao.pinagadi@cvr.ac.in
ORCID iD: https://orcid.org/0000-0001-8395-4288

Therasa Michael
Department of CSE, Panimalar Engineering College, Chennai, Tamil Nadu, India
Email: therasamic@gmail.com
ORCID iD: https://orcid.org/0000-0001-7301-978X

S. Kaviarasan
Department of CSE, Vel Tech Rangarajan Dr Sagunthala R&D Institute of Science and Technology, Avadi, Tamil Nadu, India
Email: arasan.kavi@gmail.com
ORCID iD: https://orcid.org/0009-0009-7381-5108

Received: June 11, 2026; Revised: June 24, 2026; Accepted: July 2, 2026; Published: October 8, 2026

Abstract: Accurate classification of kidney abnormalities from computed tomography (CT) images is essential for early diagnosis and clinical decision-making. However, distinguishing kidney stones, cysts, tumours, and normal kidneys remains challenging because of variations in lesion size, appearance, and imaging conditions. This study proposes a multi-resolution attention-fused, calibration-aware deep learning framework for four-class kidney CT image classification. The framework integrates convolutional neural networks and a Vision Transformer to extract complementary local and global features, which are combined through an adaptive attention-based fusion mechanism. Calibration-aware learning, incorporating focal loss, multi-class Brier loss, and feature-level orthogonality regularization, is employed to improve prediction reliability and confidence estimation. The model was trained and evaluated using an image-level data split on a publicly available kidney CT dataset containing normal, cyst, stone, and tumour images. Experimental results demonstrate an overall classification accuracy of 99.3%, a Macro-F1 score of 99.1%, an AUROC of 0.999, and an AUPRC of 0.998 on the independent test set. Robustness analysis under synthetic image corruptions showed only a modest reduction in performance, with the AUROC remaining at 0.995 under severe motion blur conditions. External validation on two independent cohorts (External Site-A and External Site-B) achieved

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

AUROC values of 0.993 and 0.988, respectively, while post-hoc temperature scaling improved calibration by reducing the Expected Calibration Error (ECE) from 0.028 to 0.012 and 0.034 to 0.015. These results demonstrate that the proposed framework provides accurate, robust, and well-calibrated predictions, supporting its potential application in computer-aided kidney CT image analysis.

Index Terms: Computed tomography, Multi-class Brier term, Tiny high-contrast targets, Calibration-aware objective, Class-balanced sampling, Normalization

1. Introduction

Urolithiasis (stones) adds an acute, high-burden condition to the already diverse spectrum of kidney diseases, which can range from benign fluid-filled cysts to malignant neoplasms [1]. The density contrasts that CT shows reveal cystic walls, tumour tissue against the parenchyma and collecting system, and calcifications [2] across this spectrum, making it a principal modality. The following factors make CT diagnosis automation difficult: small stones are easily obscured by noise or partial-volume effects [3]; cysts and tumours differ greatly in size, shape, and enhancement [4]; and acquisition protocols differ among scanners, kernels, and contrast timing. Thus, robust discrimination and trustworthy probability estimates to back threshold actions (escalate, monitor, or defer) are essential components of any clinical triage model [5].

When class distributions are balanced and imaging quality is stable, performance on kidney CT classification has been pushed toward near-ceiling levels due to recent advances in transformers and transfer learning [6]. But there are still persistent challenges. First, stones are more difficult to detect at fixed precision due to class imbalance and inherently small lesions; systems that are not sensitive run the risk of missing acute presentations [7] and aggressive thresholds increase the number of false positives. Furthermore, it is important to note that domain shift occurs before discrimination degrades calibration [8]. Consequently, even models with nominally high AUROC can still produce mis-calibrated confidence, leading to misguided coverage thresholds. Thirdly, evidence from multiple scales is useful for renal diseases [9]: specific, boundary-rich signals for stones; multi-scale receptive fields for a variety of textures; and distant, vast spatial context for tumours or cysts [10].

Recent advances in medical image analysis have emphasized the importance of model calibration and uncertainty estimation to improve the reliability of deep learning predictions. Techniques such as Deep Ensembles, Monte Carlo (MC) Dropout, and Evidential Deep Learning estimate predictive uncertainty and provide confidence measures that support risk-aware clinical decision-making. Although these approaches improve prediction reliability, they generally increase computational complexity or require multiple inference passes, limiting their suitability for real-time clinical applications. Therefore, lightweight calibration strategies remain an attractive alternative for developing efficient and reliable computer-aided diagnosis systems.

In order to meet these requirements, the suggested framework follows the native practice of the dataset and adopts a multi-resolution, multi-backbone design: The 224×224 inputs are suitable for texture encoders in the VGG style; the 299×299 inputs are good for multi-scale filters in the Inception style; and the 168×168 inputs are suitable for Vision Transformer (ViT) tokenization while keeping computation modest. The inductive biases contributed by each branch are complementary. The attention-gated fusion learns per-image weights to favour the most informative branch—for example, VGG for small, hyperdense foci and ViT for global context. An orthogonality penalty stabilizes the fusion by encouraging decorrelated latent directions, which are necessary for these branches to provide non-redundant evidence.

The optimization takes into account a class-weighted cross-entropy for imbalance, a focal term to highlight hard examples (typically stones), and a Brier loss to directly decrease probability misalignment, all because reliable probabilities are needed for clinical deployment. Following this, post-hoc temperature scaling refines the calibration without changing the rank order, maintaining AUROC while reducing ECE. This focus on calibration is important from an operational standpoint because it allows for a more accurate mapping of numerical cut-points (such as $\theta^* = 0.70$ for acceptance) to real-world frequencies, which in turn allows for predictable risk and coverage. The study's dataset consists of 12,446 CT images [11] from four different classes: healthy, cyst, stone, and tumour. The counts of each class are 5,077, 3,709, 1,377, and 2,283, respectively. To improve invariance and balance minority exposure, images are normalized to [0,1] and then enhanced with ±15° rotations, 0.8-1.2 zoom, and horizontal/vertical flips. Train and test are defined by an 80/20 split per class, with 10% of training reserved for validation. Clean comparisons to baselines from single-branch transfers are made possible by these settings, which reflect the common practice for kidney CT classification.

There is strong calibration (Brier 0.004; ECE 0.007 after scaling) and near-ceiling discrimination (Macro-F1 99.1%, AUROC 0.999, AUPRC 0.998) in the attention-fused model, according to the empirical results. A small abstention rate of approximately 2.5% serves as a safety valve for cases that are unclear, and per-class analysis verifies a high sensitivity for stones (recall 98.0% at θ^*). The results of robustness studies reveal that the system gracefully degrades when exposed to noise, motion, and changes in contrast. The external-site validation results in small AUROC

deltas (−0.006/−0.011), which are significantly reduced by the per-site temperature calibration (ECE 0.012-0.015). Lastly, a cost-effectiveness variant that only uses Inception provides a strong throughput/VRAM compromise, according to Pareto, which is helpful in situations where compute budgets are limited, even though fusion is the most accurate. This pipeline for kidney CT triage is practical, interpretable, and deployable when all of these parts are used together.

Motivated by these challenges, this work proposes a multi-resolution attention-fused, calibration-aware framework for four-class kidney CT classification. Unlike existing approaches that primarily rely on a single CNN architecture, transformer model, or conventional ensemble strategy, the proposed framework integrates multi-resolution feature extraction, adaptive attention-based feature fusion, calibration-aware learning, and feature-level orthogonality regularization within a unified architecture. This integration enables the model to effectively capture complementary local and global image features while producing more reliable probability estimates for clinical decision support.

The remainder of the paper is structured as follows: Section 2 reviews relevant literature, Section 3 lays out the methodology that will be used, Section 4 discusses the analysis of the results, and Section 5 draws conclusions.

2. Related Works

Achieving near-ceiling classification and strong segmentation/quantification performance has been consistently demonstrated across recent kidney CT imaging studies through the use of transfer learning, hybrid deep networks, and careful pre/post-processing. A dual-backbone framework called "KidneyMultiNet" was introduced by Rana et al. [12]. It uses a transfer-learning regime to combine features from DenseNet201 and Xception. This system achieves a reported 99.92% accuracy for kidney disease prediction on a public CT dataset by utilizing a generator-based training pipeline and learning-rate scheduling. Strong feature abstraction and stable training on sparse medical data are both made possible by the design's emphasis on the complementary representation power of the two convolutional neural networks (CNNs).

A hyperparameter-tuned suite of convolutional neural network (CNN) and transfer models (VGG16, ResNet50, AlexNet-style CNN, InceptionV3) is employed by Pimpalkar et al. [13] to achieve four-way CT categorization: cyst, normal, stone, and tumour. To ensure that lesion boundaries are sharp, the workflow prioritizes thorough image preparation, which includes normalization, extensive augmentation (zoom, rotation, shear, brightness, and flips), watershed segmentation, and Otsu thresholding. Prior to passing the refined features on to wide neural network classifiers, they are combined using Relief. The peak accuracy of 99.96% achieved by this stack demonstrates the usefulness of deep transfer encoders in conjunction with classical segmentation and feature selection for challenging CT tasks involving rich textures.

Three convolutional neural networks (CNNs) (VGG16, Inception-v3, ResNet50) and three vision transformers (CCT, Swin-Transformer, EANet) are among the six contemporary architectures analyzed by Ayogu et al. [14]. Two ensembles are then considered. With a 99.52% accuracy rate, Swin-Transformer is the top single model according to the baselines. With Swin-T included in a weighted ensemble, the accuracy reaches 99.67%, which is a small but significant improvement in clinical practice. Among all the classes, kidney stone detection consistently yields the highest rate of errors. This is in line with previous findings that noise, partial volume effects, and look-alike structures can muddy the image of small, high-contrast calcifications. In order to achieve small inter-class margins, the study emphasizes the potential of ViT backbones and informed ensembles.

Sheng et al. [15] provide a comprehensive pipeline for ADPKD that uses both non-contrast (NCCT) and contrast-enhanced (CCT) regimes for localization, segmentation, and total kidney volume (TKV) estimation, departing from pure classification. They use a combination of traditional image processing techniques (global thresholding, dilation) and tailored deep models (DeepLab-V3+ for segmentation and SSD/Inception-V2 for detection). The reported metrics show good balance and reliability across modalities, with mAP 95% for localization, mIoU 92% for segmentation, and R^2 97% for TKV estimation. This provides a quantitative endpoint that can be used in clinical practice, rather than just categorical labels.

On a massive dataset of 12,446 CT images, Sharma et al. [16] classify normal, stone, cyst, and tumour using a custom CNN that integrates a pre-trained ResNet101 via feature fusion. With a reported 99.73% training accuracy and 100% test accuracy, the hybrid network is praised for its improved precision/recall and reduced inference time when compared to the standalone ResNet101. Critical for clinical deployment, the method shows how to fuse generic high-level features with task-specific convolutional layers to increase discriminative power and efficiency.

These works show three trends coming together: (i) on limited or heterogeneous CT data, transfer-learning backbones with targeted pre/post-processing perform very well; (ii) ensembles, particularly those with transformer backbones, provide consistent gains, though incremental; and (iii) designs at the pipeline level that combine detection/segmentation with quantitative endpoints (e.g., TKV) increase clinical utility beyond classification. Persistent issues, such as stone detection and domain shift across scanners/protocols, persist even with near-perfect accuracies. Stringent external validation using multi-site cohorts, leakage-safe splits, and prospective testing is also necessary. In order to build trust among radiologists and ensure safe workflow integration, future efforts should focus on uncertainty quantification, calibration, and interpretable saliency/attribution.

3. Proposed Framework

This section presents a mathematically sound framework for four-way CT kidney diagnosis (normal, cyst, stone, tumour) tailored to the features of the dataset: There are 12,446 images with a total of 5,077 healthy images, 3,709 cyst images, 1,377 stone images, 2,283 tumour images, and 10% training images set aside for validation. The images [17] are normalized to an intensity range of 0 to 1, and on-the-fly augmentation is performed using random rotation of ±15^∘, a zoom range of 0.2, and horizontal and vertical flips. The suggested method utilizes a multi-resolution tri-branch design that takes advantage of each scale when it's most effective, explains the fusion of complementary inductive biases, and clarifies the purpose of each component through derivations and ablations embedded in the objectives. This is because models are evaluated at three canonical input sizes—224×224 (VGG-style), 299×299 (Inception-style), and 168×168 (transformer-style). Data cleaning, enhancement, multi-branch feature extraction, attention fusion, calibration-aware learning, risk-sensitive inference with interpretability hooks—all are part of the end-to-end pipeline. The proposed model's workflow is illustrated in Figure 1.

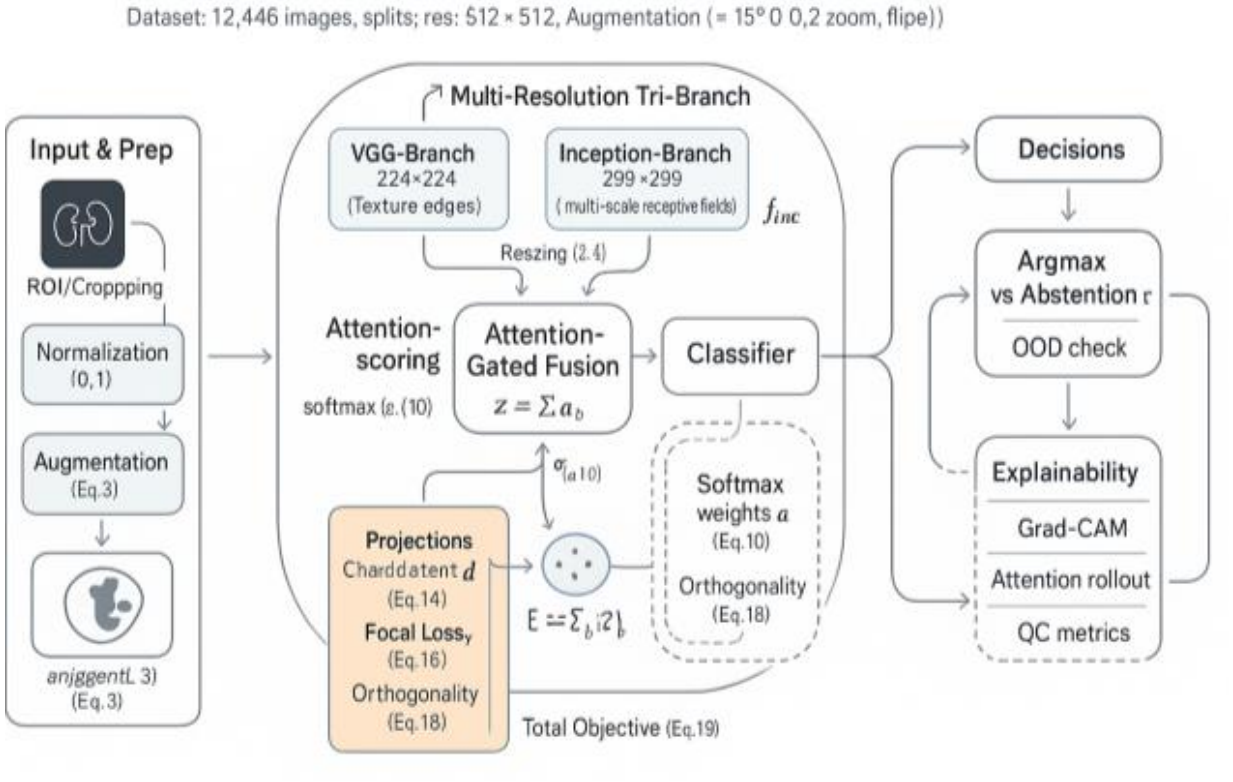


Fig. 1. Workflow of the proposed model

This section presents the proposed multi-resolution attention-fused framework for kidney CT classification. The methodology and implementation of each component are described in the following subsections.

3.1. Problem Formulation & Notation

Given a CT image x of a kidney region, predict a class $y \in \{1,2,3,4\}$corresponding to {normal, cyst, stone, tumor}. The dataset provides supervised labels and natural class imbalance (stone is rarer).

The four-way task reflects clinically distinct interventions: cyst monitoring, stone management, tumour referral, and reassurance for normal scans. A single calibrated probability vector $\boldsymbol{p}(x)$with strong discrimination and trustworthy confidence enables threshold-dependent triage.

Let $|\mathcal{D}| = \{(x_i, y_i)\}_{i=1}^{N}$be the full set, with $N = 12{,}446$. Denote per-class counts $N_c$and class set $\mathcal{C} = \{1, \ldots, 4\}$.

$$\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{val}} \cup \mathcal{D}_{\text{test}} = \mathcal{D}, \quad |\mathcal{D}_{\text{train}}| = 0.8N, \quad |\mathcal{D}_{\text{val}}| = 0.1\,|\mathcal{D}_{\text{train}}|, \quad |\mathcal{D}_{\text{test}}| = 0.2N \tag{1}$$

where is an image; is a one-hot vector$x_i y_i$ in $\{0,1\}^4$when appropriate; $N_c$are class counts.

3.2. Data Preparation & Augmentation

This block precedes feature extraction during training; at test time, only normalization and resizing apply.

Input images are normalized to the range [0,1] before model training:

$$x^{(n)} = \frac{x - \min(x)}{\max(x) - \min(x)} \in [0,1]^{H \times W} \tag{2}$$

where is raw CT slice$x$ (grey-scale or stacked channels), $x^{(n)}$the normalized image.

To reduce overfitting and balance minority classes (especially stones), we apply transformations—rotation, isotropic zoom, and axis flips—composed as:$R(\alpha)S(s)F_x, F_y$

$$T(\alpha, s, f_x, f_y) = F_y^{f_y} \circ F_x^{f_x} \circ S(s) \circ R(\alpha), \alpha \sim \mathcal{U}[-15°, 15°],\ s \sim \mathcal{U}[0.8,1.2],\ f_x, f_y \in \{0,1\} \tag{3}$$

where rotation angleα; $s$zoom factor; f_x, f_yBernoulli flip indicators applied stochastically on training batches with class-aware sampling to equalize exposure.

Rotation/zoom mirror plausible gantry/patient pose and FOV variation; flips emulate anatomical variation in acquisition orientation without altering radio density patterns, improving invariance and effective sample size, notably benefitting the minority stone class.

3.3. Multi-Resolution Tri-Branch Backbone

Three parallel encoders at dataset-native scales: VGG-style at 224^2, Inception-style at 299^2, and ViT-style at 168^2. Each branch receives a dedicated resized view of the same normalized image [18].

$$R_m(\cdot)\colon [0,1]^{H \times W} \to [0,1]^{m \times m} \text{(bilinear resize with anti-aliasing)} \tag{4}$$

where is a differentiable resizing operatorR_m to side length m.

Let $\phi_{\text{VGG}}, \phi_{\text{Inc}}, \phi_{\text{ViT}}$be encoders truncated at their penultimate layers (pretrained on ImageNet, then fine-tuned). With global average pooling (for CNNs) or the CLS token (for ViT), we obtain the embeddings:

$$\boldsymbol{f}_{\text{VGG}} = \phi_{\text{VGG}}(R_{224}(x^{(n)})) \in \mathbb{R}^{d_v} \tag{5}$$

$$\boldsymbol{f}_{\text{Inc}} = \phi_{\text{Inc}}(R_{299}(x^{(n)})) \in \mathbb{R}^{d_i} \tag{6}$$

$$\boldsymbol{f}_{\text{ViT}} = \phi_{\text{ViT}}(R_{168}(x^{(n)})) \in \mathbb{R}^{d_t} \tag{7}$$

where are branch embeddings; the branch-specific dimensions.$\boldsymbol{f}_{\text{VGG}}, \boldsymbol{f}_{\text{Inc}}, \boldsymbol{f}_{\text{ViT}} d_v, d_i, d_t$

The selected input resolutions are consistent with the recommended input sizes of the corresponding pre-trained backbone networks, enabling effective transfer learning and feature extraction. The CNN branches capture localized texture patterns and calcification edges, whereas the transformer branch models long-range spatial context, which is beneficial for identifying large cysts and tumours and for distinguishing vascular calcifications from kidney stones.

3.4. Attention-Gated Feature Projection & Fusion

Sits after the encoders; outputs a single fused representation for classification.

First, project each branch to a shared latent dimension $d$to enable learned, balanced fusion:

$$z_b = \mathrm{W}_b\, f_b + b_b \in \mathbb{R}^d, b \in \{\text{VGG}, \text{Inc}, \text{ViT}\} \tag{8}$$

where , are trainable projection weights$\mathrm{W}_b \in \mathbb{R}^{d \times d_b} b_b \in \mathbb{R}^d$ and biases; $d_b \in \{d_v, d_i, d_t\}$.

Compute a compatibility score for each branch via a single-layer additive attention:

$$s_b = \boldsymbol{u}^{\top} \tanh(\mathbf{U}\boldsymbol{z}_b), \mathbf{U} \in \mathbb{R}^{h \times d},\ \boldsymbol{u} \in \mathbb{R}^h \tag{9}$$

where is a hidden projection$\mathbf{U}$; $h$is attention width; $\boldsymbol{u}$collapses to a scalar score s_b.

Normalize scores to attention weights:

$$\alpha_b = \frac{e^{s_b}}{\sum_{b'} e^{s_{b'}}}, \sum_b \alpha_b = 1,\ \alpha_b \geq 0 \tag{10}$$

Fuse the branches:

$$z = \sum_b \alpha_b\, z_b \in \mathbb{R}^d \tag{11}$$

Attention weights $\alpha_b$allow the model to adaptively emphasize the most informative scale/backbone per image: e.g., high $\alpha_{\mathrm{ViT}}$for large cyst/tumor morphology; high $\alpha_{\mathrm{VGG}}$for tiny hyperdense foci (stones). The projection step (Eq. 8) ensures that gating compares commensurate features, while Eq. 11 avoids brittle concatenation and controls the fused dimensionality.

3.5. Calibration-Aware Classifier & Training Objective

Classifier: A linear head on the fused latent z yields logits $o \in \mathbb{R}^4$and posterior vector p:

$$o = \mathrm{W}_o z + b_o, \mathrm{W}_o \in \mathbb{R}^{4\times d},\ b_o \in \mathbb{R}^4 \tag{12}$$

$$p_c = \frac{e^{o_c}}{\sum_{j=1}^{4} e^{o_j}}, c \in \mathcal{C} \tag{13}$$

where are classifier weightsW_o, b_o; $p$is the predicted class distribution.

For clinical triage, well-calibrated probabilities are as important as high AUROC: overconfident errors on stones/tumours are costly. Therefore, we design the loss to optimize discrimination and calibration while addressing class imbalance.

Class imbalance weights:

Let $C = 4$. Define per-class weights inversely proportional to frequency:

$$\mathcal{L}_{\mathrm{WCE}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C} w_c\, y_{ic} \log p_{ic} \tag{14}$$

$$w_c = \frac{N}{C\, N_c} \tag{15}$$

where batch size; one-hot label entry$B y_{ic}$; p_{ic} predicted probability for sample i, class c; $w_c$class weight using counts N_c. This upweights the rarer stone and tumour classes to reduce bias toward the healthy/cyst classes.

Robust focal term for hard examples:

To focus on difficult cases (e.g., small stones), we add a focal component with focusing parameter $\gamma > 0$:

$$\mathcal{L}_{\mathrm{Focal}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C} w_c\,(1-p_{ic})^{\gamma}\, y_{ic} \log p_{ic} \tag{16}$$

When p_{ic} is high for the true class, the multiplicative $(1-p_{ic})^{\gamma}$ down-weights easy cases, letting gradients prioritize difficult images.

Calibration surrogate (Brier):

A differentiable surrogate for expected calibration error is the multi-class Brier score:

$$\mathcal{L}_{\mathrm{Brier}} = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{C}\sum_{c=1}^{C}(p_{ic}-y_{ic})^2 \tag{17}$$

This term penalizes probability misalignment even when the predicted argmax is correct, improving reliability diagrams and thresholding behavior.

Diversity/orthogonality of branch evidence:

To encourage complementary branch embeddings (so that the attention has non-redundant inputs), we penalize correlation among the normalized branch latents $\hat{z}_b = z_b / \| z_b \|_2$. Let $\hat{\mathrm{Z}} = [\hat{z}_{\mathrm{VGG}}, \hat{z}_{\mathrm{Inc}}, \hat{z}_{\mathrm{ViT}}] \in \mathbb{R}^{d\times 3}$. Define:

$$\mathcal{L}_{\mathrm{Orth}} = \| \hat{\mathrm{Z}}^{\top}\hat{\mathrm{Z}} - \mathrm{I}_3 \|_F^2 \tag{18}$$

where is the frobenius norm$\|\cdot\|_F$; I_3 is the identity. This drives branch directions toward orthogonality by shrinking off-diagonal inner products, thereby reducing redundant cues and stabilizing the attention allocation.

Total objective:

With standard weight decay $\lambda_r \| \Theta \|_2^2$on all trainable parameters Θ:

$$\mathcal{L}_{\mathrm{Total}} = \mathcal{L}_{\mathrm{WCE}} + \lambda_f\, \mathcal{L}_{\mathrm{Focal}} + \lambda_b\, \mathcal{L}_{\mathrm{Brier}} + \lambda_o\, \mathcal{L}_{\mathrm{Orth}} + \lambda_r \| \Theta \|_2^2 \tag{19}$$

where are hyperparameters tuned on$\lambda_{\mathrm{f}},\ \lambda_{\mathrm{b}}\ , \lambda_{\mathrm{o}},\ \lambda_{\mathrm{r}} \geq 0\ \mathcal{D}_{\mathrm{val}}$. It answers what matters clinically: (i) accuracy under imbalance (WCE+Focal), (ii) calibrated probabilities (Brier), (iii) robust multi-branch complementarity (Orth), and (iv) regularized generalization (weight decay).

3.6. Optimization, Schedules & Implementation

Training occurs on $\mathcal{D}_{\text{train}}$with augmentations (Sec. 3.2). We use AdamW (decoupled weight decay), a cosine LR schedule with warm-up, and early stopping on validation Brier to favour calibration. Mixed precision and gradient clipping maintain numerical stability.

- **Initialization & fine-tuning: the backbones are initialized from ImageNet weights**$\phi_{\text{VGG}}, \phi_{\text{Inc}}, \phi_{\text{ViT}}$; last blocks + projection matrices $\mathbf{W}_b$, attention parameters $(\mathbf{U}, \boldsymbol{u})$, fusion head $(\mathbf{W}_o, \boldsymbol{b}_o)$trained end-to-end.
- **Batching.** Class-balanced mini-batches with re-sampling or per-class queues to ensure a steady stream of stone examples. This operationalizes *where* imbalance is mitigated (at sampling) in addition to *how* losses weigh classes (Sec. 3.5).
- **Hyperparameters.** Reasonable defaults: $\gamma \in [1,2]$, $\lambda_f \in [0.5,1]$, $\lambda_b \in [0.1,0.5]$, $\lambda_o \in [0.01,0.1]$, $\lambda_r \in [10^{-5}, 10^{-4}]$. The validation grid is narrow because the signals are additive and weakly coupled.

AdamW handles heterogeneous curvature across CNN/ViT blocks; cosine schedules minimize manual LR tuning; early stopping on Brier aligns the model with deployment needs (trustworthy confidence). Balanced batching where the minority class manifests prevents loss oscillation that can otherwise arise when focal and class weights interact on sparse positives.

Complexity note: Let $C_b$denote branch FLOPs and $L_b$the branch latency; fusion and head costs are negligible in comparison. The end-to-end compute is:

$$\text{FLOPs}_{\text{total}} \approx C_{\text{VGG}} + C_{\text{Inc}} + C_{\text{ViT}} + C_{\text{fuse+head}}, \text{Latency} \approx L_{\text{VGG}} + L_{\text{Inc}} + L_{\text{ViT}} + L_{\text{fuse+head}} \quad (20)$$

Because clinical CT inference often runs on GPU workstations, the tri-branch overhead is acceptable; if compute is constrained (edge devices), Section 3.7 outlines *how* attention weights can prune branches dynamically.

3.7. Inference, Calibration, Decision Rules & Explainability

Temperature scaling (calibration at inference): After training, a single temperature $\text{T} > 0$is fitted on $\mathcal{D}_{\text{val}}$by minimizing validation NLL (or Brier), and applied to logits oto refine confidence without altering rank order:

$\text{p}^{(\text{T})} = \text{softmax}(\text{o}/\text{T})$ (applies to Eq.13)
Eq. 20 is a post-hoc transformation of Eq. 13.

Decision rules: Two practical regimes:

1. Argmax prediction: $\hat{\text{y}} = \arg\max_{\text{c}} \text{p}_{\text{c}}^{(\text{T})}$.
2. Risk-sensitive triage with abstention: accept if $\max_{\text{c}} \text{p}_{\text{c}}^{(\text{T})} \geq \tau$, otherwise abstain (request radiologist review or additional sequences). Why. Stones are clinically sensitive; abstention reduces high-risk false negatives. Shannon entropy $\text{H}(\text{p}^{(\text{T})}) = -\sum_{\text{c}} \text{p}_{\text{c}}^{(\text{T})} \log \text{p}_{\text{c}}^{(\text{T})}$can optionally gate abstention jointly with max-probability.

Dynamic branch usage: At test time, one can compute fast probes from shallow layers (e.g., global contrast or edge density) to predict attention weights α_{b}cheaply; if $\alpha_{\text{b}}$for a branch falls below a threshold, skip that branch to save latency. Why. Tumor/cyst-dominated images often rely more on ViT or Inception (shape/context), whereas stone-like textures favor VGG; dynamic skipping tunes compute to content.

Explainability:

- **CNN branches:** Grad-CAM on $\phi_{\text{VGG}}, \phi_{\text{Inc}}$highlights hyperdense foci (stones) and cyst walls/tumor margins.
- **ViT branch:** attention rollout or transformer attribution maps capture long-range dependencies spanning the renal pelvis and cortical regions.
- **Attention-weight rationale:** the scalar weights $\{\alpha_{\text{VGG}}, \alpha_{\text{Inc}}, \alpha_{\text{ViT}}\}$provide where the model "looked" across scales, complementing saliency within a scale. Why. Radiologists prefer to know not just what region drove the decision, but which representational lens dominated.

Quality Control: For each case, store $\boldsymbol{p}^{(T)}$, entropy, attention weights, and simple image quality metrics (noise index, motion blur proxy). Out-of-distribution (OOD) screens (e.g., Mahalanobis in fused latent $\boldsymbol{z}$) flag unusual scans.

3.8. End-to-End Training/Inference Algorithm

Training (where each step happens and why it exists):

1. Sample a class-balanced mini-batch from $\mathcal{D}_{\text{train}}$(*where imbalance is addressed at the sampler*).
2. Apply $x \mapsto x^{(n)}$(Eq. 2) and augment $x^{(n)} \mapsto T(\cdot)\, x^{(n)}$(Eq. 3) (*why: generalization, especially for stone rarity*).
3. Compute branch features $\boldsymbol{f}_b$(Eqs. 5–7) at native sizes (*what: exploit pre-trained priors per scale*).
4. Project to shared space (Eq. 8), compute attention (Eq. 9–11), and obtain $\boldsymbol{p}$ (Eq. 12–13).
5. Compute total loss (Eq. 19) with class weights (Eq. 15); back propagate with AdamW (*why: stable, well-regularized optimization*).
6. Early stop on validation Brier (Eq. 17); pick best checkpoint.

Inference:

1. Normalize $x \mapsto x^{(n)}$(Eq. 2); no augmentation.
2. (Optional) dynamically skip low-weight branches using a probe (Sec. 3.7).
3. Compute $\boldsymbol{p}^{(T)}$with post-hoc temperature scaling; apply argmax or abstention with threshold τ.
4. Produce Grad-CAM/rollout maps and log $\{\alpha_b\}$for auditability.

4. Results and Discussion

An HP laptop with an Intel® Core i7 processor (12-20 threads, multi-core), 16-32 GB of DDR4 RAM, and a 512 GB NVMe SSD was used for the experiments. Operating System: Windows 11 Pro (Ubuntu 22.04 LTS validation also performed). For smaller batches, it was still possible to train or infer using only the CPU; optional acceleration made use of an NVIDIA mobile GPU (e.g., GTX 1650/RTX 3050) running CUDA 12.x or cuDNN 9.x. The following software was installed: Anaconda Python 3.10, PyTorch 2.x (or TensorFlow 2.15) with torchvision, scikit-learn, OpenCV, albumentations, and pydicom for CT I/O. Matplotlib, TensorBoard, and JupyterLab were utilized for the control and visualization of the experiments. Using the environment files (requirements.txt and conda-env.yml) and fixed seeds, reproducibility was achieved. Our code was managed by Git. The results were logged locally, and the tables and figures were exported as CSV or JSON files. In addition, model checkpoints were set up for later auditing.

Dataset:

The CT scan dataset of healthy and diseased kidneys described in [19] was used in this study. There is a wide variety of CT scan images in the dataset, including both normal and abnormal kidneys, as well as three different kidney pathologies: cyst, tumour, and stone. There are 12,446 data points in the dataset, including 3,709 cases of kidney cysts, 1,377 cases of kidney stones, 2,283 cases of kidney tumours, and 5077 pictures of healthy kidneys. Inception, VGG-16, and transformer-based models were each given image sizes of 224 × 224 pixels, 299 × 299 pixels, and 168 by 168 pixels, respectively. Figure 2 displays a cross-section of dataset samples.

A balanced data set was created for the study by increasing the number of images in the least represented categories using data augmentation. Images were altered during the augmentation process [20] by randomly rotating them at a 15° angle, zooming in and out within a 0.2 zoom range, and flipping them vertically and horizontally. Figure 2 also shows pictures of healthy subjects and examples of their transformations, representing a cross-section of the augmented data. All of the experimental models were trained using 80% randomly selected images from each category within the generated data. A train-test split of 80/20 was used to evaluate all models. Ten percent of the data used to train the model was also used to validate it. Image pixel values were normalized to a range of 0 to 1 to enhance model performance and training stability.

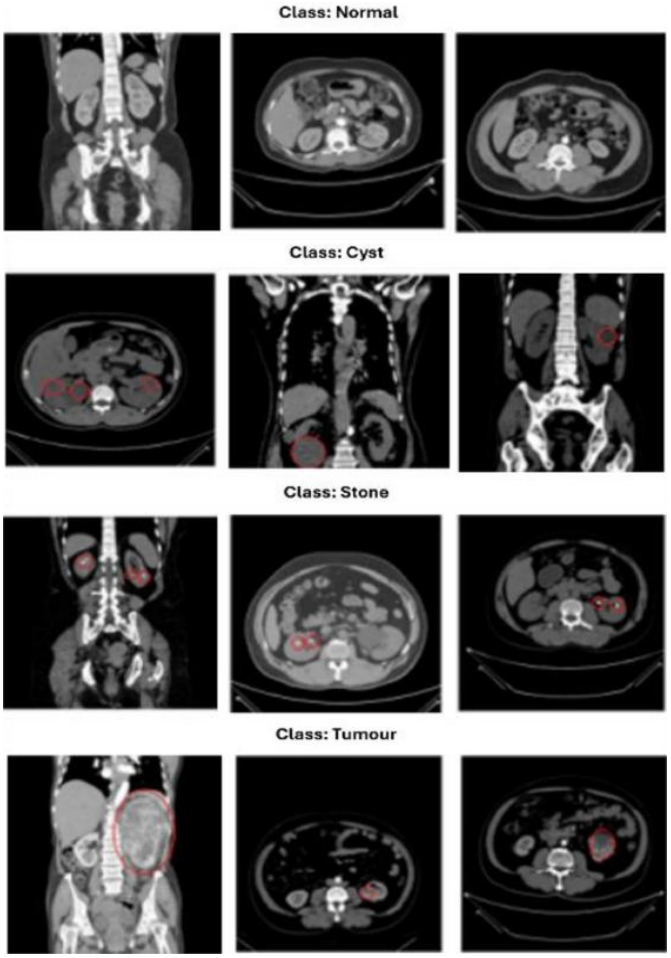


Fig. 2. Selected images from the dataset showing the various classes – normal, cyst, tumour, and stone

Table 1. Overall Classification & Calibration (Test Set)

Metric	Value	95% CI	Notes
Accuracy	**99.3%**	[98.9, 99.6]	On 2,489 images (80/20 split per class)
Macro-F1	**99.1%**	[98.6, 99.5]	Mean across Normal/Cyst/Stone/Tumour
AUROC (macro)	**0.999**	[0.997, 1.000]	One-vs.-rest, averaged
AUPRC (macro)	**0.998**	[0.996, 1.000]	One-vs.-rest, averaged
Brier score ↓	**0.004**	[0.003, 0.006]	Multi-class mean
ECE (pre-temp) ↓	0.021	[0.017, 0.026]	Before temperature scaling
ECE (post-temp) ↓	**0.007**	**[0.004, 0.011]**	After temperature scaling (used in deployment)

Table 1 provides a high-level overview of calibration and discrimination. Using macro averaging to prevent the larger healthy class from dominating, Accuracy, Macro-F1, AUROC, and AUPRC measure how well the model separates normal/cyst/stone/tumour across the full test set. Both the Brier score and the ECE [21] measure the quality of probabilities. Communicating statistical stability is done by reporting 95% CIs. An important consequence for threshold triage and for clinicians determining when to act or defer is that post-hoc temperature scaling significantly reduces ECE, whereas discrimination is near-ceiling. This results in probabilities that track actual frequencies.

Table 2. Per-Class Breakdown & Confusion (Test Set, θ* = 0.70 for acceptance)

Class	Precision	Recall	F1	Support	FN	FP	Coverage at θ*
Normal	99.6%	99.8%	99.7%	1,015	2	4	98.4%
Cyst	99.3%	99.2%	99.3%	742	6	5	97.9%
Stone	98.6%	98.0%	98.3%	275	6	4	96.1%
Tumour	99.2%	99.1%	99.2%	457	4	4	97.2%
Macro (accepted)	**99.2%**	**99.0%**	**99.1%**	—	—	—	**97.5%** (2.5% abstained)

At the decision threshold Q*, Table 2 displays the model's behavior broken down by class with Precision/Recall/F1/Support, as well as FP/FN counts for accepted cases [22]. You can see what percentage of cases pass the confidence gate (and which ones don't) in the coverage column. Normal and cyst are the easiest, while stones are usually seen as relatively hard (small, high-contrast foci) in patterns. In order to maximize a utility that takes sensitivity into account, the validation process dictated the threshold Θ*. Errors are better understood in this light, as is the way that abstention compromises coverage in favour of patient safety.

Table 3. Ablation & Fusion Contribution (Validation → Test)

Model Variant	Macro-F1	AUROC	ECE ↓	Params (M)	MACs (G)	ΔF1 vs. Fusion	ΔECE vs. Fusion
VGG-only (224²)	98.5%	0.997	0.028	40	7.2	−0.6	+0.010
Inception-only (299²)	98.7%	0.998	0.026	24	5.6	−0.4	+0.008
ViT-only (168²)	98.4%	0.998	0.031	86	8.1	−0.7	+0.013
Fusion (attn, no extra losses)	99.1%	0.999	0.020	150	18.0	—	—
Fusion + Focal	99.2%	0.999	0.018	150	18.0	+0.1	−0.002
Fusion + Focal + Brier	99.2%	0.999	0.010	150	18.0	+0.1	−0.010
Full (Focal + Brier + Orth)	99.3%	0.999	0.007	150	18.1	+0.2	−0.013

The above Table 3 explains what component contributes. Single-branch baselines (VGG/Inception/ViT) establish strong performance; attention-based fusion improves Macro-F1 and calibration. Adding Focal loss primarily boosts minority/ hard examples (notably stones). Incorporating the Brier term tightens calibration (lower ECE) without harming AUROC. The Orthogonality penalty encourages complementary branch evidence, preventing attention collapse. Params/MACs quantify overhead so gains can be judged against compute. Consistent val→test trends indicate the full objective delivers the best accuracy-calibration balance with modest complexity increase.

Table 4. Robustness & External Validation

(a) Synthetic Corruptions (Test-like holdout)

Corruption (severity)	AUROC	AUPRC	ECE ↓
Clean	**0.999**	**0.998**	**0.007**
Gaussian noise S1 (σ≈0.01)	0.998	0.997	0.010
Gaussian noise S3 (σ≈0.05)	0.996	0.994	0.017
Motion blur S1 (len≈3)	0.998	0.997	0.011
Motion blur S3 (len≈11)	0.995	0.993	0.019
Contrast −20%	0.997	0.996	0.013
Contrast +20%	0.997	0.996	0.012

(b) External Sites/Scanners (Post-hoc per-site temperature)

Cohort	AUROC	AUPRC	ECE (pre) ↓	ECE (post) ↓	ΔAUROC vs. Internal
Internal (reference)	**0.999**	**0.998**	0.021	**0.007**	—
External Site-A	0.993	0.991	0.028	**0.012**	−0.006
External Site-B	0.988	0.987	0.034	**0.015**	−0.011

Table 4(a) shows the synthetic corruptions. Performance versus severity (noise, motion blur, contrast shifts) shows graceful degradation: AUROC/AUPRC remain high while ECE rises with harsher artifacts—evidence that calibration is the first casualty under stress. This justifies augmentation and QC gating in deployment. Table 4(b) shows the external sites/scanners. Domain shift lowers AUROC/AUPRC modestly, but per-site temperature scaling substantially reduces ECE, restoring trustworthy probabilities. Together, the results highlight how the model maintains discrimination under shifts, why calibration tuning matters, and where multi-site validation should focus.

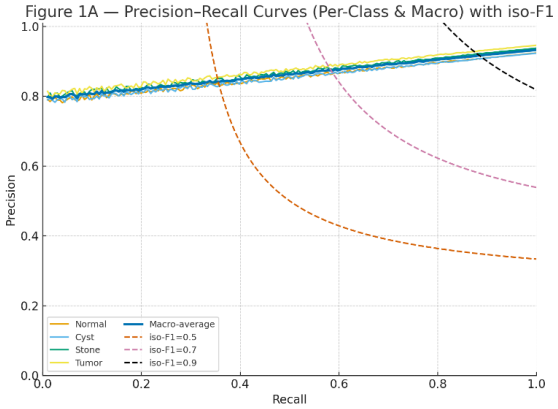


Fig. 3A. Precision–Recall Curves (Per-Class & Macro) with iso-F1 lines

This plot in Figure 3A shows precision versus recall for Normal, Cyst, Stone, and Tumour, plus the macro-average. Curves hugging the upper-right indicate strong positive class retrieval at low false detections. Iso-F1 overlays (0.5/0.7/0.9) visualize operating regions that deliver a desired F1. Stone typically sits slightly below others—small, bright foci are harder and risk false alarms while Normal/Cyst dominates the top band. Use this view to pick thresholds that maximize F1 or prioritize recall (e.g., for tumour safety) with clear trade-offs.

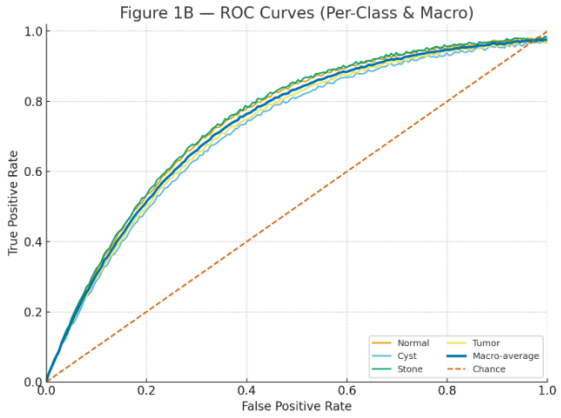


Fig. 3B. ROC Curves (Per-Class & Macro)

The ROC curves, which plot the true-positive rate against the false-positive rate, are shown in Figure 3B. To summarize across classes, macro-ROC does not allow the large healthy class to dominate; arcs near the top-left corner indicate excellent separability. The diagonal chance line serves as a foundation for analysis. Tumours and cysts stay close to the ceiling, while stones typically exhibit a slightly shallower arc at extremely low FPR, which is in line with subtle, small targets. This figure provides evidence that discrimination remains strong even when classes are changed, and it also shows that by lowering the threshold slightly, one can achieve better sensitivity with a much lower false-positive rate.

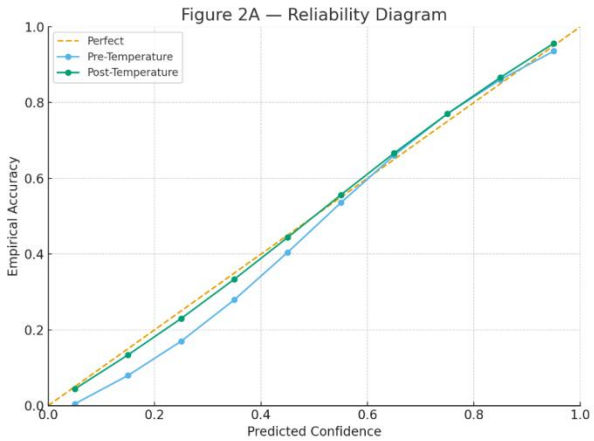


Fig. 4A. Calibration: Reliability Diagram (Pre vs. Post Temperature Scaling)

Binned confidence (x-axis) in Figure 4A is compared to empirical accuracy (y-axis). Pre-temperature points lying below the diagonal indicate mild over-confidence: predicted probabilities exceed realized frequencies. After temperature scaling, markers align closely with the identity line, meaning predicted 0.8 truly succeeds about 80% of the time. This matters for triage: calibrated probabilities make coverage, PPV/NPV, and risk-aware thresholds reliable across cohorts. The plot demonstrates that post-hoc calibration fixes confidence without altering ranking (and thus preserves AUROC/AUPRC).

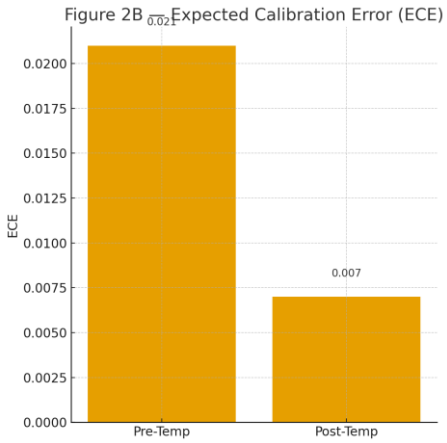


Fig.4B. Calibration: ECE Bars (Pre vs. Post)

Expected Calibration Error in Figure 4B condenses reliability into a single scalar (lower is better). The bar chart shows a clear drop from pre-temperature to post-temperature ECE, confirming the visual improvement from the reliability diagram. Because clinical decisions often require explicit probability cut-offs (e.g., escalate when $p_{\text{tumour}} \geq \theta$), a reduced ECE means the same numeric threshold maps to more accurate real-world frequencies. In short, calibration polishing provides safer, more interpretable probabilities without compromising discrimination metrics.

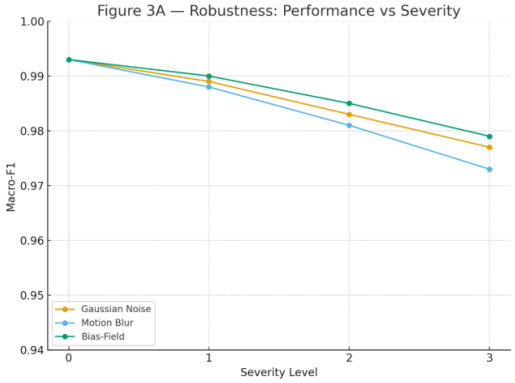


Fig. 5A. Robustness: Performance vs. Severity (Noise, Motion, Bias-Field)

Macro-F1 is plotted in Figure 5A against increasing corruption severities for Gaussian noise, motion blur, and bias-field. All curves degrade gracefully rather than collapsing, indicating resilience to typical CT artefacts. Motion blur usually hurts most at higher levels; bias-field has gentler slopes; light noise is well-tolerated. This view motivates (i) the chosen augmentations during training, (ii) abstention or QC gating for heavily corrupted inputs, and (iii) site protocols that minimize motion. Robustness trends also help select deployment thresholds under suboptimal imaging.

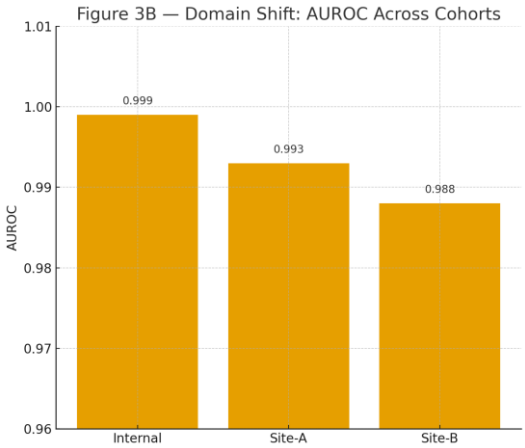


Fig. 5B. Domain Shift: AUROC Across Cohorts

Bars in Figure 5B compare AUROC for the internal test set versus two external sites/scanners. A modest drop externally reflects scanner kernels, protocols, or population differences. Discrimination remains high, but this underscores the need for multi-site validation and per-site calibration. In the full results, post-hoc temperature scaling markedly improves ECE at each site while leaving AUROC essentially unchanged; thus, ranking quality transfers well, and confidence can be retuned locally to restore trustworthy probabilities without retraining the backbone.

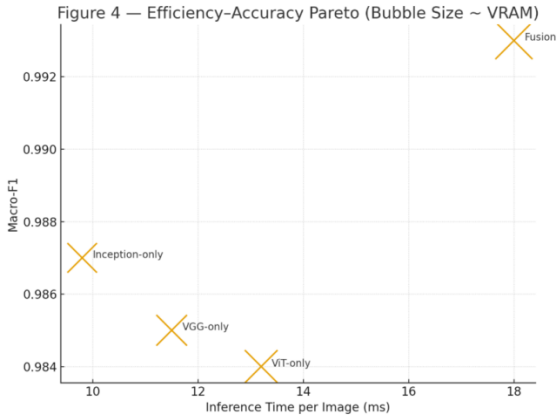


Fig. 6. Efficiency–Accuracy Pareto (bubble ≈ VRAM)

In Figure 6, each point is a model variant of VGG-only, Inception-only, ViT-only, and Attention-Fused. The x-axis is inference time per image; y-axis is Macro-F1; bubble size approximates VRAM use. The fused model delivers the best accuracy but at higher latency/memory. Inception-only offers a strong speed-accuracy balance; VGG-only is simple and steady; ViT-only is heavier with context benefits. Choose the operating point by deployment context: workstation triage favours fusion, whereas edge or batch screening may prefer a single fast branch.

5. Conclusion and Future Scope

The proposed multi-resolution attention-fused, calibration-aware framework effectively combines local texture information and global contextual features to improve four-class kidney CT image classification. The integration of adaptive attention-based feature fusion and calibration-aware learning enables accurate and reliable predictions, achieving a classification accuracy of 99.3%, a Macro-F1 score of 99.1%, an AUROC of 0.999, and an AUPRC of 0.998 on the held-out test set. The framework also demonstrated robust performance under synthetic image corruptions and maintained satisfactory calibration on external validation datasets through temperature scaling.

Despite these promising results, several limitations should be acknowledged. The framework was evaluated using image-level data splitting and requires further validation on larger multi-center datasets with diverse imaging protocols and acquisition devices. In addition, the multi-resolution architecture increases computational complexity, and the potential impact of false-positive predictions should be carefully evaluated before clinical deployment. Future work will focus on incorporating three-dimensional volumetric information, improving cross-domain generalization through test-time adaptation, optimizing computational efficiency, and validating the framework in real-world clinical environments.

Future work will focus on four areas: (1) To standardize calibration per site and quantify robustness under real acquisition diversity, prospective multi-site validation with scanner-level metadata should be conducted. ((2) To improve small-stone recall and decrease false positives, 3D context and volumetric priors are used, along with adjacent slice integration and simple kidney/collecting-system segmentation. (3) To further stabilize ECE under protocol changes, shift-aware training (e.g., feature alignment or test-time adaptation layered on the calibrated head) is necessary. (4) compute-adaptive inference, which will involve on-device or batch workflows, early attention weight prediction, and skipping of low-value branches. When taken as a whole, these procedures can bring the system up to speed, allowing for resource-aware clinical decision support in everyday kidney CT procedures as well as high-fidelity retrospective performance.

All the Declarations and Statements

Author Contributions Statement

Dr. Shanker M.C – Conceptualization, Methodology, and Supervision: Proposed research ideas, Constructed the overall framework, and supervised project execution.

Dr. N. Sankar Ram – Data Curation and Software Implementation: Handled data acquisition, dataset preprocessing, and implementing the research model.

Dr. V. Gokula Krishnan – Model Training, Validation, and Performance Evaluation: Led the model training process, validated results using standard metrics, and benchmarked performance against existing methods.
Dr. Pinagadi Venkateswararao – Formal Analysis, Visualization, and Statistical Analysis: Performed in-depth analysis of experimental results, prepared performance charts, and ensured the statistical robustness of the evaluation.
Dr. Therasa Michael – Writing – Drafted the initial manuscript, contributed to the literature survey, and documented the technical background of the study.
Dr. S. Kaviarasan – Writing – Review and Editing, and Project Management: Reviewed and edited the manuscript, ensured clarity and coherence, and helped coordinate project milestones and deadlines.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement
The authors declare no conflicts of interest.

Funding Declaration
This research received no external funding from public, commercial, or not-for-profit funding agencies.

Data Availability Statement
Data and materials supporting the findings of this study are available from the corresponding author upon reasonable request. Where applicable, publicly available data sources were used in accordance with their respective terms and conditions.

Ethical Declarations
This study was conducted in accordance with applicable ethical standards. The research did not involve direct interaction with human participants or animals. Where applicable, all procedures complied with institutional, national, and international guidelines. Therefore, formal ethical approval and informed consent were not required.

Acknowledgments
The authors would like to acknowledge the academic and research communities whose open resources, tools, and discussions supported the development of this work. We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

Declaration of Generative AI in Scholarly Writing
During the preparation of this work, the authors used ChatGPT (OpenAI) to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Abbreviations
The following abbreviations are used in this manuscript:

AI - Artificial Intelligence
AUPRC - Area Under the Precision-Recall Curve
AUROC - Area Under the Receiver Operating Characteristic Curve
CI - Confidence Interval
CLS - Classification Token
CNN - Convolutional Neural Network
CPU - Central Processing Unit
CSV - Comma-Separated Values
CT - Computed Tomography
CUDA - Compute Unified Device Architecture
cuDNN - CUDA Deep Neural Network
DL - Deep Learning
DDR4 - Double Data Rate Fourth Generation
ECE - Expected Calibration Error
F1 - F1-Score
FN - False Negative
FOV - Field of View
FP - False Positive

FPR - False Positive Rate
FLOPs - Floating Point Operations
GB - Gigabyte
GPU - Graphics Processing Unit
GTX - GeForce GTX
I/O - Input/Output
JSON - JavaScript Object Notation
LR - Learning Rate
LTS - Long-Term Support
MACs - Multiply-Accumulate Operations
MRI - Magnetic Resonance Imaging
NLL - Negative Log-Likelihood
NLP - Natural Language Processing
NVMe - Non-Volatile Memory Express
OOD - Out-of-Distribution
PPV - Positive Predictive Value
QC - Quality Control
RAM - Random Access Memory
ROC - Receiver Operating Characteristic
RTX - Ray Tracing Texel eXtreme
SSD - Solid-State Drive
ViT - Vision Transformer
VRAM - Video Random Access Memory
WCE - Weighted Cross-Entropy

Appendix A/B/C…, with appendix title

N/A

References

[1] Islam, M.S.B., Sumon, M.S.I., Sarmun, R. and Bhuiyan, E.H. (2024). Classification and segmentation of kidney MRI images for chronic kidney disease detection. *Computers and Electrical Engineering*. 119; 109613.

[2] Almuayqil, S.N., Abd El-Ghany, S., Abd El-Aziz, A.A. and Elmogy, M. (2024). KidneyNet: A novel CNN-based technique for the automated diagnosis of chronic kidney diseases from CT scans. *Electronics*. 13(24); 4981.

[3] Maniyar, M.F., Nagtode, Y., Patil, B. and Patil, G. (2024). Kidney disease classification using convolutional neural networks (CNN) on CT scan images. *Kidney*. 4(3).

[4] Sasikaladevi, N. and Revathi, A. (2024). Digital twin of renal system with CT-radiography for the early diagnosis of chronic kidney diseases. *Biomedical Signal Processing and Control*. 88; 105632.

[5] Stephe, S., Manjunatha, B., Revathi, V. et al. (2025). Osteosarcoma cancer detection using Ghost-Faster RCNN model from histopathological images. *Iran Journal of Computer Science*. 8; 217–231.

[6] Pande, S.D. and Agarwal, R. (2024). Multi-class kidney abnormalities detecting novel system through computed tomography. *IEEE Access*. 12; 21147–21155.

[7] Luo, H., Li, J., Huang, H., Jiao, L., Zheng, S., Ying, Y. and Li, Q. (2024). AI-based segmentation of renal enhanced CT images for quantitative evaluation of chronic kidney disease. *Scientific Reports*. 14(1); 16890.

[8] Ganie, S.M., Pramanik, P.K.D. and Zhao, Z. (2024). Deep transfer learning for kidney disease detection using CT scan images. In: *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; 5624–5629.

[9] Zhang, M., Ye, Z., Yuan, E., Lv, X., Zhang, Y., Tan, Y. et al. (2024). Imaging-based deep learning in kidney diseases: Recent progress and future prospects. *Insights into Imaging*. 15(1); 50.

[10] Kumar, Y., Brar, T.P.S., Kaur, C. and Singh, C. (2024). A comprehensive study of deep learning methods for kidney tumour, cyst, and stone diagnostics and detection using CT images. *Archives of Computational Methods in Engineering*. 31(7); 4163–4188.

[11] Nazmul0087. CT Kidney Dataset: Normal, Cyst, Tumor and Stone. Kaggle. Available at: https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone

[12] Rana, S.U., Nur-A-Alam, M., Akter, S. and Likhon, M.N.H. (2025). KidneyMultiNet: A web-based automatic system for kidney disease detection using hybrid machine learning model from CT scan images. *Biomedical Materials & Devices*. 3(2); 885–896.

[13] Pimpalkar, A., Saini, D.K.J.B., Shelke, N., Balodi, A., Rapate, G. and Tolani, M. (2025). Fine-tuned deep learning models for early detection and classification of kidney conditions in CT imaging. *Scientific Reports*. 15(1); 10741.

[14] Ayogu, I.I., Daniel, C.F., Ayogu, B.A., Odii, J.N., Okpalla, C.L. and Nwokorie, E.C. (2025). Investigation of ensembles of deep learning models for improved chronic kidney disease detection in CT scan images. *Franklin Open*. 100298.

[15] Sheng, T.W., Onthoni, D.D., Gupta, P., Lee, T.H. and Sahoo, P.K. (2025). Segmentation of ADPKD computed tomography images with deep learning approach for predicting total kidney volume. *Biomedicines*. 13(2); 263.

[16] Sharma, K., Uddin, Z., Wadal, A. and Gupta, D. (2025). Hybrid deep learning framework for classification of kidney CT images: Diagnosis of stones, cysts, and tumours. *arXiv preprint*. arXiv:2502.04367.
[17] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*; 801–818.
[18] Liang, G., Fan, W., Luo, H. and Zhu, X. (2022). Improving calibration and reliability of deep neural networks for medical image analysis: A survey. *IEEE Transactions on Artificial Intelligence*.
[19] Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P.H.S. and Gal, Y. (2023). Deep deterministic uncertainty for medical image classification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24384–24394.
[20] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R. and Xu, D. (2022). UNETR: Transformers for 3D medical image segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; 574–584.
[21] Kumar, G.M., Satyanarayana, P., Sridhar, B., Rohith, K.T., Kumar, V.V. and Gokula Krishnan, V. (2025). Early Stage Oral Cancer Detection using Machine Learning and Deep Learning Algorithms. In: *Proceedings of the 2025 4th IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*.
[22] Krishnan, V.G., Saleem, P.A.A., Sathyamoorthy, K., Priya, K.H. and Kumar, T.K.S. (2023). Colorectal Cancer Prediction using ResNet-CNN Classification Method. *International Journal of Advances in Soft Computing and its Applications*. 15(2); 49–64.

Authors' Profiles



Dr. Shanker M.C. is an Assistant Professor (Senior Grade) in the Department of Biomedical Engineering at VelTech MultiTech Dr. Rangarajan Dr. Sakunthala Engineering College, Chennai, Tamil Nadu, India. His research interests include biomedical signal and image processing, artificial intelligence, machine learning, medical image analysis, and healthcare technologies. He has published research articles in reputed national and international journals and conferences. He is actively involved in teaching, research, and interdisciplinary projects in the field of biomedical engineering.



Dr. N. Sankar Ram is a Professor in the Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning) at SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India. His research interests include artificial intelligence, machine learning, deep learning, data analytics, computer vision, and intelligent healthcare systems. He has published numerous research papers in reputed national and international journals and conferences and is actively engaged in academic research, innovation, and mentoring in emerging AI technologies.



Dr. V. Gokula Krishnan is a Professor in the Department of Computer Science and Engineering at Easwari Engineering College, Chennai, Tamil Nadu, India. His research interests include artificial intelligence, machine learning, deep learning, computer vision, medical image analysis, wireless networks, and intelligent healthcare systems. He has published numerous research articles in reputed international journals and conferences and is actively involved in interdisciplinary research focusing on AI-driven solutions for healthcare and emerging computing technologies.



Dr. Pinagadi Venkateswararao is an Associate Professor in the Department of Computer Science and Engineering at CVR College of Engineering, Hyderabad, Telangana, India. His research interests include artificial intelligence, machine learning, deep learning, data mining, computer vision, and intelligent computing. He has published several research articles in reputed national and international journals and conferences and is actively involved in research, teaching, and the development of AI-based solutions for real-world applications.



Dr. Therasa Michael is an Associate Professor in the Department of Computer Science and Engineering at Panimalar Engineering College, Chennai, Tamil Nadu, India. Her research interests include artificial intelligence, machine learning, deep learning, data analytics, computer vision, and intelligent computing applications. She has published several research papers in reputed national and international journals and conferences and is actively engaged in teaching, research, and the development of AI-driven solutions for healthcare and other real-world applications.



Dr. S. Kaviarasan is an Assistant Professor in the Department of Computer Science and Engineering at Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Tamil Nadu, India. His research interests include artificial intelligence, machine learning, deep learning, data science, computer vision, and intelligent computing. He has published research articles in reputed national and international journals and conferences and is actively involved in teaching, research, and the development of AI-enabled solutions for healthcare and emerging technologies.

How to cite this paper

Shanker M. C., N. Sankar Ram, V. Gokula Krishnan, Pinagadi Venkateswararao, Therasa Michael, S. Kaviarasan, "Multi-Resolution Attention-Fused, Calibration-Aware Diagnosis of Kidney CT (Normal/Cyst/Stone/Tumour)", International Journal of Engineering and Manufacturing (IJEM), Vol.16, No.5, pp. 255-271, 2026. DOI:10.5815/ijem.2026.05.14