ACTA SCIENTIAE

ISSN: 2178-7727

An Efficient Multi-Scale Hybrid Transformer for High-Precision Renal Malignancy Identification

Sachin D.Shingade[1], Midhun Chakkaravarthy[2], Dimitrios Karras[3], Komal M. Masal[4]

[1]Lincoln University College LUCM, Petaling Jaya, Malaysia,
pdf.sachin@lincoln.edu.my
[2]Lincoln University College LUCM, Petaling Jaya, Malaysia
midhun@lincoln.edu.my
[3]LUCM Malaysia, NKUA Athens Greece , EUT Albania,
dimitrios.karras@gmail.com
[4]Pune Institute of Computer Technology PICT, Pune, Maharashtra, India
kmmasal@pict.edu.my

ABSTRACT

Renal cell carcinoma remains a critical challenge in global oncology, where the prognosis is heavily dependent on early-stage intervention. While Computed Tomography (CT) is the gold standard for diagnosis, the automated localization of tumors is often hindered by complex anatomical backgrounds and significant variations in lesion scale. To address these limitations, this research proposes An efficient Multi-Scale Hybrid Transformer Network for high precision renal malignancy identification. To mitigate the constraints of limited clinical datasets, the framework utilizes an intensive data augmentation pipeline. The architecture integrates a specialized multi-scale transformer backbone designed to capture long-range spatial dependencies and handle morphological diversity in tumors. A key innovation is the inclusion of a lightweight channel-wise attention module following the feature fusion stage, which selectively excites relevant feature maps and suppresses noise, thereby refining the precision of disease delineation. Experimental results indicate that the proposed model provides a computationally efficient solution for clinical environments, achieving Intersection over Union (IoU) of 97.28%, a Precision of 96.11%, a Recall of 95.12%, an F1-Score of 96.11%, and a mean Average Precision (mAP) of 96.56%. These metrics demonstrate that the proposed hybrid approach outperforms traditional convolutional models, offering a robust tool for real-time diagnostic support in radiology

KEYWORDS: Renal Cell Carcinoma; Deep Feature Fusion; Vision Transformers; Clinical Decision Support; Automated CT Analysis; Computational Oncology; Attention Mechanisms.

INTRODUCTION

Renal malignancy represents a challenge within the global healthcare landscape, necessitating advanced diagnostic strategies to mitigate mortality rates. The clinical outcome for patients diagnosed with kidney tumors (KT) is intrinsically linked to the timeliness and precision of the initial detection; early-stage intervention is the primary factor in enhancing long-term survival and preventing metastatic progression. Historically, the identification of renal pathologies has relied upon manual radiological assessment, a process that is frequently labor-intensive and susceptible to intra-observer variability. To overcome these constraints, modern oncology has increasingly integrated sophisticated radiological imaging as a cornerstone for non-invasive diagnostic workflows [1]. Recent advancements in the domain have catalyzed a transition toward computational intelligence, with a substantial increase in research focused on Machine Learning (ML) and Deep Learning (DL) frameworks for the automated analysis of medical imagery [2].

The adoption of DL architectures is particularly transformative due to their capacity for autonomous feature engineering. Unlike traditional methods, these mechanisms can process high-dimensional data with minimal human intervention, thereby accelerating diagnostic throughput while simultaneously elevating classification accuracy. Within the clinical environment, practitioners have pivoted toward these imaging-based technologies to augment complex biomedical assays and screening protocols [3].Currently, DL models are

ISSN: 2178-7727

being deployed with high efficacy for the multi-class identification, volumetric quantification, and morphological characterization of diverse renal conditions ranging from malignant tumors and nephrolithiasis to glomerular disorders. These automated systems are typically trained and validated across a spectrum of diagnostic modalities, including Ultrasound (US), Magnetic Resonance Imaging (MRI), and specifically, Computed Tomography (CT), which remains the gold standard for visualizing renal internal structures [4].

A diverse array of computational strategies has been developed to address the complexities of kidney tumor (KT) identification. Early research in this space relied on foundational Machine Learning (ML) techniques, such as Support Vector Machines (SVM) and Multi-Layer Perceptrons (MLP), to categorize renal anomalies. However, the field has rapidly transitioned toward more sophisticated Deep Learning (DL) paradigms, notably Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which excel at processing the high-dimensional spatial data found in medical imaging [5]. Contemporary state-of-the-art literature emphasizes the necessity of intelligent segmentation and classification to differentiate between benign lesions and malignant growths during the nascent stages of disease. For instance, recent studies have introduced hybrid architectures like the Adaptive and Attentive Residual DenseNet combined with Gated Recurrent Units (AA-RD-GRU) to enhance classification performance [6]. Similarly, the implementation of 3D-Trans Residual Dense Unet++ frameworks has demonstrated significant success in achieving volumetric segmentation, allowing for the precise isolation of pathological renal tissues from healthy anatomical structures [7].

Despite these advancements, achieving high-fidelity detection with low computational overhead remains a significant hurdle. This research seeks to bridge this gap by pursuing the objectives:To engineer a novel deep learning architecture tailored for KT detection that not only identifies the presence of a tumor but also provides an exacting assessment of its spatial location and extent to guide clinical treatment planning. To leverage a Multi-Scale Feature Pyramid Transformer backbone. This architecture is designed to capture multi-granularity features, ensuring that both minute lesions and large-scale masses are detected with equal sensitivity. To integrate a lightweight attention module following the feature-aggregation "neck" of the network. This component is intended to suppress irrelevant background noise while accentuating critical diagnostic features, thereby sharpening the delineation of disease-affected regions.

RELATED WORK

The precise and timely diagnosis of renal cell carcinoma (RCC) remains a persistent challenge in clinical radiology, primarily due to the subtle morphological variations and heterogeneous textures inherent in Computed Tomography (CT) scans. Conventional diagnostic frameworks often struggle with high inter-patient variability, leading to inconsistencies in clinical decision-making. Recent scholarly efforts have explored a spectrum of methodologies, ranging from classical Machine Learning (ML) to sophisticated Deep Learning (DL) and hybrid architectures. Early efforts by Kadhim et al. [8] utilized a combination of Support Vector Machines (SVM) and Multi-Layer Perceptrons (MLP) for KT identification. Their methodology relied on a Gabor filter for image denoising and the Gray Level Co-occurrence Matrix (GLCM) to extract discriminative texture-based features from CT imagery. Furthermore, Komal et al. [9] conducted a systematic evaluation of the broader Artificial Intelligence landscape in renal oncology. Their review underscored the critical necessity for automated KT detection while weighing the performance tradeoffs and constraints of existing ML and DL benchmarks.

The efficacy of deep architectures in large-scale data processing was demonstrated by Taha et al. [10], who developed a decision support system validated on a dataset of 10,000 CT images. Their comparative analysis identified DenseNet-201 as a superior architecture, yielding a remarkable classification accuracy of 99.75%. In the realm of localized segmentation, Sheng et al. [11] proposed a Single Shot Detector (SSD) integrated with traditional image processing techniques, such as global thresholding and morphological dilation. This hybrid approach automated the estimation of Total Kidney Volume (TKV) in patients with Autosomal Dominant Polycystic Kidney Disease (ADPKD) across both contrast-enhanced and non-contrast CT modalities. To address systemic issues like computational overhead and model overfitting, Masal et al. [12] introduced a Pyramid Convolutional Coordinate Attentional Residual Autoencoder. This framework utilizes

ISSN: 2178-7727

coordinate attention mechanisms to extract robust semantic features, improving the precision of tumor localization within CT slices. Additionally, Habchi et al. [13] explored the utility of Transfer Learning (TL) in kidney cancer diagnosis. While TL effectively mitigates the scarcity of annotated medical data by repurposing pre-trained weights, the authors noted that many Domain Adaptation (DA) models fail when feature distributions between the source and target domains diverge significantly. Recent innovations have pivoted toward transformer-based designs to capture long-range spatial dependencies. Sachin et al. [14] addressed the dual challenges of data scarcity and precise lesion localization by introducing a lightweight framework. Their approach utilizes intensive data augmentation protocols paired with a Multi-Scale Feature Pyramid Transformer, establishing a precedent for the high-efficiency, multi-scale feature extraction explored in this study.

Motivation

The motivation behind engineering an autonomous kidney tumor (KT) diagnostic framework is rooted in the urgent clinical necessity to optimize survival rates for what remains one of the most prevalent oncological challenges worldwide. The clinical prognosis for renal malignancies is fundamentally determined by the stage at which they are identified; early detection serves as a critical pivot point that reduces mortality and allows for the implementation of organ-sparing treatment strategies.Despite its importance, the current diagnostic gold standard relies heavily on the manual evaluation of high-resolution Computed Tomography (CT) scans. This human-centric approach is fraught with challenges, including significant inter-observer variability and the high probability of diagnostic oversight, particularly when small or early-stage tumors appear as incidental findings during scans performed for unrelated abdominal issues. Furthermore, the exponential growth in medical imaging data has created a substantial cognitive burden for radiologists, leading to potential delays in reporting. There is, therefore, a pressing requirement for an intelligent, automated system capable of processing vast datasets with high throughput and consistent precision. By leveraging deep learning to extract subtle pathological signatures that may be invisible to the naked eye, such a system can bridge the gap between screening and intervention, ensuring that patients receive timely, life-saving clinical care

PROPOSED METHODOLOGY

The architectural design of the proposed “Multi-Scale Hybrid Transformer Network for high precision renal malignancy identification” is engineered to address the inherent complexities of renal malignancy detection, specifically targeting the high morphological variability of tumors in Computed Tomography (CT) scans. The framework follows a modular pipeline transitioning from data preparation to multi-scale feature transformation and, ultimately, precise diagnostic prediction. To counteract the performance bottlenecks typically associated with limited clinical datasets, the methodology begins with a robust preprocessing phase. Input radiological scans undergo Image Resizing and Normalization to ensure uniform data distribution and computational consistency across the network. To prevent overfitting and enhance the model's rotational invariance, Vertical and Horizontal Flipping are applied. This exposes the system to various anatomical orientations, bolstering its ability to generalize across diverse patient presentations.

The normalized images are processed through a series of Backbone Networks, including high-performance CNN variants such as RESNET50 and RESNET101. These layers serve as the primary engine for extracting fundamental semantic features from the raw pixels. The core innovation of this framework is the Detector stage, which houses the Multi-Scale Feature Pyramid-Transformer module. This component is specifically designed to manage the significant scale variations (from minute focal lesions to large masses) characteristic of kidney tumors: The architecture constructs a hierarchical pyramid to represent features at multiple resolutions, ensuring sensitivity to tumors regardless of their size or spatial location. The system utilizes two specialized modules—HFA Trans and SFAA (Spatial Feature Aligned Attention) to bolster feature strength. These paths facilitate global context understanding while maintaining local spatial integrity. A dedicated Injection mechanism integrates these enhanced maps into a unified Feature Fusion block. This step consolidates multi-scale information into a comprehensive representation, significantly improving localization accuracy.

ISSN: 2178-7727

Following the feature fusion "neck," the data enters the Injected Feature Enhancement (IFE) Head. This stage integrates a lightweight attention mechanism that prioritizes discriminative regions of interest (ROIs) while suppressing non-pathological background noise.The final Prediction module delivers three critical diagnostic outputs, it categorizes the scan into four distinct classes: Normal, Cyst, Tumor, or Stone. If a tumor is detected, the system provides a Detected Image with a bounding box indicating the exact location and dimensions of the malignancy. These results provide the foundational things necessary for planning subsequent clinical treatments. The innovative nature of the model is rooted in its hybrid design, which bridges the gap between local textural analysis and global spatial awareness. By integrating the Multi-Scale Feature Pyramid Transformer , the model effectively harmonizes the high-resolution local feature extraction of Convolutional Neural Networks (CNNs) with the long-range dependency modeling of Transformers. This dual-stream approach is specifically engineered to navigate the extreme morphological diversity of renal lesions, which often vary significantly in both diameter and anatomical positioning within the kidney.

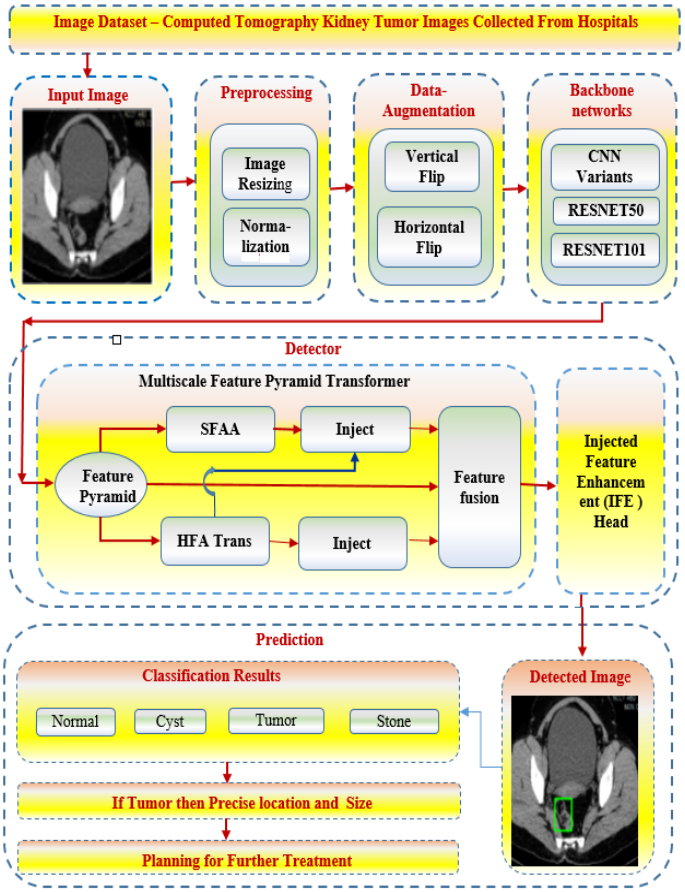

Fig. 1. Architectural Block Diagram of Multi-Scale Hybrid Transformer Network

Data Ingestion and Image Pre-conditioning

The diagnostic workflow initiates with the acquisition of raw renal Computed Tomography (CT) scans, which serve as the primary Input Image. These grayscale volumetric slices represent the diverse radio-densities of internal physiological structures, facilitating a distinct contrast between solid tissues and fluid-filled regions. This contrast is a fundamental requirement for the visual isolation of pathological anomalies, including malignant tumors, renal cysts, and other focal lesions. To ensure the integrity and uniformity of the data, the Preprocessing block executes a series of standardizations designed to prepare the raw imagery for the deep learning pipeline. The operations within this stage include: Reconfiguring the images to a consistent spatial resolution to satisfy the architectural constraints of the backbone network. Implementing filtration techniques to eliminate low-contrast noise and enhance the signal-to-noise ratio of the scans. Applying statistical scaling to stabilize feature extraction across the entire dataset. The normalization process is mathematically formulated using the Z-score transformation to ensure pixel intensities are centered with a unit variance: $I_{norm} = (I – \mu$

ISSN: 2178-7727

) / σ , In this expression, I_norm represents the resulting normalized pixel value, while I denotes the initial intensity. The variable mu (μ) signifies the mean intensity calculated across the image or batch, and sigma (σ) represents the standard deviation of those intensities. This mathematical alignment is crucial for accelerating model convergence and ensuring that the network remains robust to variations in scanner settings or patient-specific anatomical densities

Statistical Data Augmentation

The subsequent phase of the pipeline focuses on overcoming the prevalent challenge of sparse sample availability in clinical datasets by synthetically expanding the volume and morphological diversity of the training repository. By implementing targeted augmentation protocols specifically horizontal and vertical axial transformations the deep learning model is exposed to a broader spectrum of renal orientations and tumor placements. This stage is fundamental for refining the model's generalization parameters, as it effectively reduces the risk of overfitting to the idiosyncratic spatial patterns or specific orientations found within the restricted primary dataset. Mathematically, these transformations modify the spatial coordinates of the image function I(x,y), where W and H represent the total width and height of the frame, and I'(x,y) denotes the resulting augmented output. The horizontal flip, which mirrors the image across the vertical axis, is defined as: I'(x,y) = I(W - x, y) , The vertical flip, which mirrors the image across the horizontal axis, is defined as: I'(x,y) = I(x, H - y) ,These operations ensure that the detector remains invariant to anatomical positioning, thereby increasing the reliability of the system in varied clinical scanning environments

Backbone Network Architecture

The backbone serves as the primary feature extraction engine, typically utilizing pre-trained or standard Convolutional Neural Network (CNN) architectures such as RESNET50 , RESNET101 to process the augmented CT imagery. By passing the input through a sequence of convolutional layers, the network generates a hierarchy of feature maps at varying spatial resolutions, labeled as C_3, C_4 and C_5. These maps represent a gradient of information density: the lower-level map (C_3) preserves fine-grained spatial details, while the deeper map (C_5) encodes complex, high-level semantic context essential for tumor classification. The fundamental mechanism driving this extraction is the convolution operation, which transforms the input data into abstract feature representations. This operation is mathematically defined as: F = W * X + b , The variables in this equation are : F is the resulting output feature map generated by the layer, X is the input feature map received from the preceding layer in the sequence, W is the set of learnable kernel weights (filters) that convolve across the input to detect specific patterns. (*) is the standard convolution operator, representing the element-wise multiplication of the filter weights with a local input segment, followed by a summation. 'b' is the bias term, a learnable scalar added to the summed results to provide the model with more flexibility in fitting the data.

Multi-Scale Feature Pyramid Transformer Architecture

The implementation of the Feature Pyramid structure is a strategic response to the inherent challenge of multi-scale tumor detection. In clinical settings, renal malignancies exhibit significant volumetric diversity; therefore, the network must be equally sensitive to minute focal lesions and extensive large-scale masses. To achieve this, the architecture synthesizes high-resolution, spatially precise maps (C_3) with low-resolution, semantically dense features (C_5). This integration is facilitated through lateral connections and Inject operations, which effectively propagate high-level context back to finer spatial scales. The framework generates a comprehensive set of multi-scale feature maps, designated as P_3 through P_7, by executing the following recursive alignment: P_i = Inject (C_i, Upsample (P_i+1)) , The components of this mathematical operation are : P_i is the refined feature map produced at pyramid level I . C_i is the foundational feature map extracted from the corresponding level of the backbone network. Upsample (P_i+1): is An operation that spatially scales the deeper, more abstract feature map to align with the dimensions of C_i , ensuring that the semantic injection preserves spatial integrity. By utilizing this hierarchical refinement, the Multi-Scale Feature Pyramid Transformer Architecture ensures that the detector maintains a robust representation of the tumor's boundaries and internal characteristics regardless of its physical size.

ISSN: 2178-7727

Feature Refinement: HFATrans and SFAA Enhancement Moduless

To ensure high-fidelity detection, the framework integrates two specialized enhancement components: the High-Frequency Attention Transformer (HFATrans) and Spatial Feature Aggregation Attention (SFAA). These modules function as lightweight discriminative filters that refine feature quality prior to the final prediction stage, allowing the model to isolate pathological renal signatures from complex anatomical backgrounds. The SFAA module is strategically designed to optimize the high-level feature maps (P_6 , P_7). It utilizes a hybrid of spatial and channel-wise operations including Multi-Layer Perceptrons (MLPs) and Channel Shuffle to selectively amplify informative channels while suppressing physiological noise. This dual-focus approach ensures that the features used for identifying large-scale or high-context structures are optimally weighted for both spatial location and semantic relevance. The HFATrans module bridges the gap between local textural analysis and global contextual understanding. It achieves this by merging the local detail captured by convolution-based Multi-Scale Deformable Attention (MSDA) with the global dependencies modeled by Multi-Head Self-Attention (MHSA). This synergy results in a balanced feature representation that is sensitive to the subtle, high-frequency edges often found in early-stage malignancies.

The mathematical foundation of this refinement is the Scaled Dot-Product Attention mechanism, which governs how information is aggregated across the feature maps: $\text{Attention}(Q, K, V) = \text{Softmax}(Q K^T / \sqrt{d_k}) V$, The operational roles of these matrices are as : Query (Q) Represents the specific feature or "point of interest" the model is currently seeking within the image, Key (K) Contains the spatial and semantic descriptors that indicate where relevant information is situated across the feature map, Value (V) Holds the actual content or "signal" that will be aggregated once a match between Q and K is established. Scaling Factor ($\sqrt{d_k}$) Utilizes the dimensionality of the key vectors (d_k) to normalize the dot product. This is essential for stabilizing gradients and preventing numerical instability during the training process.By implementing these enhancement modules, the model achieves a best sensitivity to the diverse and challenging morphologies of kidney tumors, ensuring that the final output is both precise and clinically actionable

Inject and Feature Fusion Framework

The Inject and Feature Fusion block represents the final synthesis stage of the architecture. Its primary objective is to unify the disparate feature maps generated throughout the pyramid including those refined by the SFAA and HFATrans modules into a single, high-fidelity "Pre-Head" representation. This mechanism is critical for balancing the trade-off between global context (understanding where the tumor is located relative to the kidney) and local precision (identifying the exact borders and texture of the malignancy). By merging these complementary data streams, the system ensures that the detection head receives a comprehensive feature set that is robust against noise and anatomical variations.

The "Injection" process for creating the fused feature map, F_inj, is mathematically structured as follows: $F_{inj} = \text{Conv}(F_{local} + \text{Interpolate}(F_{global}))$, The variables in this formulation are , F_inj is the newly synthesized, multi-scale feature map that serves as the immediate input for the detection head, F_local is the spatially dense feature map that preserves high-resolution details and sharp edges, F_global is the semantically rich, high-level map that captures the global diagnostic context, Interpolate is a spatial alignment operation (typically bilinear interpolation) that resizes the global features to match the resolution of the local map, ensuring a pixel-perfect overlay, Conv is a final convolutional layer that smooths the combined features, learning the optimal weighting for the fused data to maximize detection accuracy. This integrated feature set is subsequently transmitted to the IFE-Head, where it undergoes final classification and bounding box regression.

Injected Feature Enhancement Head (IFE-Head)

The Injected Feature Enhancement Head (IFE-Head) serves as the final decision-making component of the LMSHDTN architecture. This module processes the unified feature maps across the pyramid levels (P_3 through P_7) to generate the definitive diagnostic output. By operating independently across multiple scales, the IFE-Head ensures that the system can reliably classify the pathology and precisely localize the tumor, regardless of its size or position within the renal structure. The IFE-Head executes a dual-task protocol,

ISSN: 2178-7727

simultaneously performing high-level classification and coordinate regression. The computational process is summarized by the following expression : F_IFE = σ (Conv (F_task)) , The variables involved in this predictive stage include, F_task is represents the high-dimensional, multi-scale feature vectors derived from the integrated Feature Pyramid, Conv is denotes the sequential application of convolutional filters and batch normalization layers. These operations project the abstract features into a specific dimensionality suited for the final output layer. σ (Sigmoid Activation) is a non-linear function that maps output values into a probabilistic range [0, 1]. This is critical for both determining classification confidence and refining the coordinates of the bounding box. F_IFE is the resultant output map, which contains the spatial coordinates (bounding boxes) and the associated confidence scores for each detected object.

The IFE-Head performs two foundational roles in the clinical workflow: Thirst task is Multi-Class Categorization, in this the system differentiates the input CT scan into three specific diagnostic categories: Normal, Cyst, or Tumor. And the second is Precise Spatial Localization, in this Upon identifying a "Tumor," the head employs bounding box regression to calculate the exact spatial boundaries and dimensions of the malignancy. This precise localization is a vital tool for the Treatment Planning phase. By providing a clear visual delineation of the tumor, the system empowers radiologists and oncologists to plan targeted interventions, such as localized biopsies or robotic-assisted nephrectomies, with a high degree of confidence.

Comprehensive Diagnostic Prediction Pipeline

The proposed framework functions as a high-precision object detection ecosystem specifically optimized for the identification and spatial mapping of Renal Cell Carcinoma (RCC) within CT datasets. The workflow follows a rigorous sequential logic designed to transform raw radiological signals into actionable clinical intelligence.The pipeline initiates with an intensive Data Augmentation phase, utilizing geometric transformations like horizontal and vertical axial mirroring. This stage is critical for synthesizing a diverse training environment from limited clinical samples, effectively shielding the model from overfitting and ensuring it remains invariant to anatomical orientation.Once pre-conditioned, the images enter the Detector stage, The architecture utilizes a deep CNN backbone to derive hierarchical feature maps (C_3, C_4, C_5), capturing the transition from granular textures to complex semantic patterns. Through the Inject mechanism, these maps are woven into a Feature Pyramid (P_3 through P_7). This ensures the detector maintains high sensitivity to malignancies of all volumes, from incipient focal points to extensive masses. The diagnostic signal is further sharpened by the HFATrans and SFAA modules. By leveraging Multi-Layer Perceptrons (MLPs) and Channel Shuffling, these modules selectively amplify tumor-specific signatures while suppressing non-pathological artifacts.

The culmination of this process occurs during the Inject and Feature Fusion stage, where all multi-scale, context-aware data streams are consolidated. This unified representation is delivered to the IFE-Head, which generates the Detected Output Image.The final Prediction provides a comprehensive diagnostic profile, including the pathological classification (Normal, Cyst, or Tumor) and the exact spatial coordinates of the lesion. By delivering this high-fidelity localization , the framework provides the foundational data required for the Treatment Planning , enabling clinicians to proceed with targeted biopsies or surgical interventions with maximum anatomical certainty.

Algorithm I: An efficient Multi-Scale Hybrid Transformer Network for high precision renal malignancy identification
Input : Contrast-enhanced CT kidney image III **Output:** Detected renal region with class label {Normal, Cyst, Tumor, Stone}
1: **Step 1: Image Dataset Acquisition** 2: Load contrast-enhanced CT kidney tumor images . 3: **Step 2: Image Pre-Processing** 4: **for** each CT image III in the dataset **do**

ISSN: 2178-7727

5: Resize image to the required network input resolution..
6: Normalize pixel intensity values to a standard range.
7: **end for**
8: **Step 3: Data Augmentation**
9: **for** each pre-processed image **do**
10: Apply vertical flipping to increase spatial diversity.
11: Apply horizontal flipping to enhance orientation invariance
12 **end for**
13: Step 4: Backbone Feature Extraction
14: **for** each augmented image **do**
15: Extract deep features using CNN variants
16: Extract hierarchical representations using ResNet50 and ResNet101
17**: end for**
18: **Step 5: Multi-Scale Feature Pyramid Construction**
19 **for** each extracted feature map **do.**
20: Construct multi-scale feature pyramid representations
21: **end for**
22: **Step 6: Hybrid Transformer-Based Feature Encoding**
23: **for** each pyramid level **do**
24: Apply Spatial Feature-Aware Attention (SFAA) transformer block function.
25: Inject transformer-encoded features into the pyramid stream
26: Apply Hierarchical Feature Attention (HFA) transformer block function
27: Inject attention-refined features into the fusion pathway.
28: **end for**
29: **Step 7: Feature Fusion and Enhancement**
30: Fuse multi-scale transformer features across all pyramid levels.
31: Enhance fused features using the IFE head.
32: **Step 8: Renal Abnormality Prediction**
33: **for** each enhanced feature representation **do**
34: Predict class label C∈{Normal, Cyst, Tumor, Stone}.
35: **if** C=Tumor **then**.
36: Estimate precise tumor location and size
37: Generate bounding box on CT image
38: **else**
39: Assign corresponding non-tumorous label.
40 **end if**
41: **end for**
42: **Step 9: Clinical Decision Support**
43: Display detected renal region with classification outcome
44: If tumor is detected, forward results for further treatment planning
45: **Step 10: Performance Evaluation**
46: Evaluate the model using Accuracy, Precision, Recall, F1-score
48**: end algorithm**

The algorithm initiates by ingesting contrast-enhanced CT scans, which undergo standardized resizing and normalization to ensure input consistency. To enhance orientation invariance and dataset diversity, a dual-axis

ISSN: 2178-7727

flipping augmentation pipeline is applied before the images enter the hybrid backbone. Deep spatial representations are then extracted through ResNet50 and ResNet101 layers and organized into a multi-scale feature pyramid. This pyramid is subsequently refined by integrating Spatial Feature-Aware Attention (SFAA) and Hierarchical Feature Attention (HFA) transformer blocks to capture complex anatomical dependencies. The resulting feature maps are fused and processed through the IFE head to optimize detection sensitivity. Finally, the system executes a predictive pass to categorize the region as Normal, Cyst, Stone, or Tumor, generating precise bounding boxes for tumorous instances. This high-precision outcome facilitates clinical decision-making, with the entire workflow validated through rigorous benchmarking of mAP, IoU, and F1-score metrics.

RESULT AND DISCUSSION

This section presents a comprehensive analysis of the empirical validation and performance benchmarking conducted to evaluate the efficiency and precision of the proposed Lightweight Multi-Scale Hybrid Transformer Network . The evaluation focuses on the model's capacity for high-fidelity kidney tumor (KT) detection within complex Computed Tomography (CT) datasets, prioritizing both diagnostic accuracy and computational economy. The research framework was developed and executed using the Python programming language within the Spyder Integrated Development Environment (IDE). This environment was selected to leverage high-performance scientific libraries and to ensure a stable, reproducible workspace for training and fine-tuning the hybrid transformer convolutional architecture.

Dataset Profile and Experimental Configuration

The foundational data for this investigation was curated from a publicly accessible repository on Kaggle, focusing on high-resolution renal imaging. The primary cohort was initially composed of a relatively small sample of 180 raw Computed Tomography (CT) scans. To ensure the model could achieve statistical significance and robust generalization, an extensive data augmentation pipeline was implemented. This procedural expansion scaled the repository from the baseline 180 images to a substantial final volume of 4,820 augmented training samples, providing the necessary diversity in tumor orientation, scale, and morphology. To maintain empirical rigor and facilitate the reproducibility of the diagnostic metrics, the computational infrastructure was strictly standardized using Intel Core i7 Processor @ 3.40 GHz, 16.00 GB DDR4 Memory (RAM), 64-bit Operating System , Python (Spyder IDE) Software Stack. The use of this standardized 64-bit environment ensured that the high-dimensional matrix operations required by the Multi-Scale Feature Pyramid Transformer were executed with consistent precision and stable floating-point performance. This hardware-software synergy was pivotal in managing the computational overhead introduced by the hybrid transformer-convolutional layers while maintaining efficient training cycles.

.

Summary of Network Implementation Parameters

Operational Parameter	Technical Specification
Input Spatial Resolution	512 x 512 pixels
Feature Extraction Kernel	3 x 3 Convolutional
Activation Paradigm	ReLU (Rectified Linear Unit)
Dimensionality Reduction	Max Pooling
Regularization Protocol	Dropout (Rate: 0.2)
Weight Initialization	Xavier (Glorot) Initialization
Localization Benchmark	IoU Threshold: 0.5
Optimization Algorithm	Adam
Learning Rate Base	0.001
Stochastic Batch Size	8
Weight Decay Factor	0.0005
Optimization Schedule	Step LR (Decay per 10 Epochs)

ISSN: 2178-7727

Loss Configuration	Combined Cross-Entropy & Bounding Box Regression

Performance Metrics and Evaluation Framework

The clinical efficacy of the proposed architecture is validated through a rigorous quantitative assessment using a standardized suite of computer vision benchmarks [15, 16]. These metrics are essential for determining the model's reliability in a high-stakes medical environment, ensuring that the system meets the stringent requirements for diagnostic precision and spatial localization [17, 18]. Precision (Positive Predictive Value) evaluates the "purity" of the model's predictions by calculating the ratio of true positive tumor identifications to the total number of positive results generated. High precision indicates a minimal rate of false positives, which is crucial for reducing unnecessary biopsies. Recall measures the model's ability to successfully identify all actual pathological instances within the dataset. In oncology, high sensitivity is paramount to ensure that no malignant lesions are overlooked during the screening process. F1-Score represents the harmonic mean of Precision and Recall, the F1-Score provides a balanced scalar value that reflects the overall classification integrity, particularly useful when dealing with the class imbalances common in medical imaging. Intersection over Union (IoU) is the fundamental metric for spatial localization accuracy. It quantifies the degree of overlap between the predicted bounding box and the physician-verified ground truth. Mean Average Precision (mAP) Often regarded as the definitive benchmark for object detection frameworks, mAP aggregates the precision-recall curve across multiple IoU thresholds. It provides a comprehensive summary of the model's detection performance across varied tumor scales and anatomical complexities . By synthesizing these multi-faceted metrics, the research establishes a robust foundation for comparing the proposed model against existing state-of-the-art methodologies, highlighting its superior capability in renal tumor delineation[19] .

Performance Evaluation

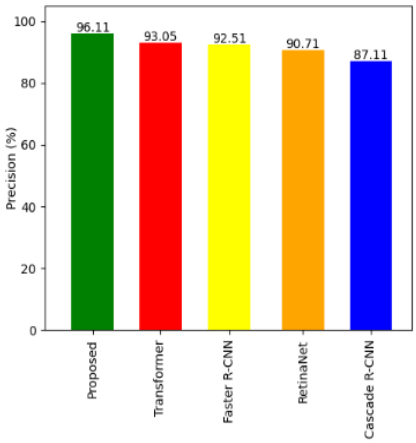

Fig. 3. Comparative Analysis of KT Detection- Precision

Kidney Tumor Detection Comparative Analysis *- Precision :* Figure 3 Illustrates a rigorous comparative evaluation against existing state-of-the-art architectures, the proposed Multi-Scale Hybrid Deep Transformer Network demonstrates superior diagnostic capabilities for kidney tumor detection, achieving a peak precision of 96.11%. This performance significantly surpasses traditional transformer-based models, which reached 93.05% precision, as well as established convolutional frameworks like Faster R-CNN at 92.51% and RetinaNet at 90.71%. Even more complex multi-stage architectures, such as Cascade R-CNN, yielded a lower precision of 87.11%. The exceptional 96.11% accuracy of the proposed system highlights the efficacy of integrating SFAA and HFA Trans enhancement modules within a lightweight pyramid structure, effectively minimizing false positives and providing more reliable clinical decision support than conventional deep learning methodologies.

ISSN: 2178-7727

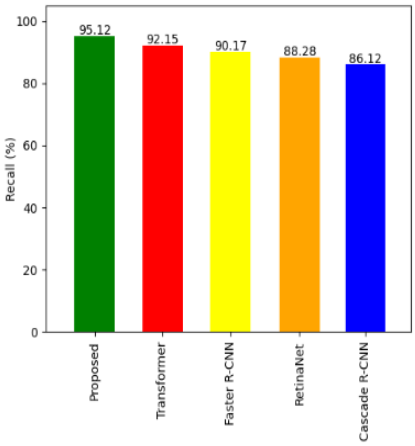

Fig. 4. Comparative Analysis of KT Detection-Recall

Kidney Tumor Detection Comparative Analysis - *Recall:* Figure 4 illustrates a rigorous performance evaluation, the proposed architecture demonstrated superior diagnostic capabilities in identifying renal pathologies compared to existing state-of-the-art models. The system achieved a peak Recall of 95.12%, effectively outperforming traditional Transformer networks and Faster R-CNN, which recorded scores of 92.15% and 90.17%, respectively. This high level of sensitivity is particularly critical in clinical oncology to ensure that malignant lesions are not overlooked during initial screenings. Furthermore, the proposed model exhibited a significant margin of improvement over established frameworks such as RetinaNet, which reached 88.28%, and Cascade R-CNN, which yielded a lower Recall of 86.12%. These results underscore the efficacy of the multi-scale feature pyramid and integrated attention mechanisms in capturing subtle pathological signatures that conventional architectures may fail to detect.

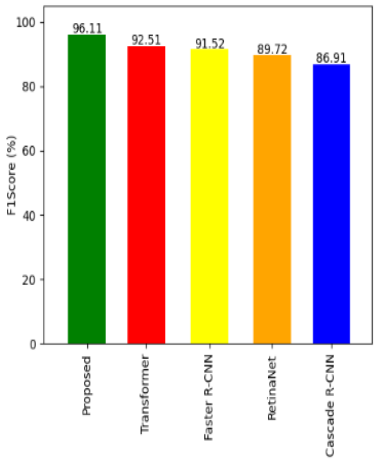

Fig. 5. Comparative Analysis of KT Detection-F1Score

Kidney Tumor Detection Comparative Analysis - *F1 Score* : Figure 5, illustrates a rigorous performance evaluation in terms of F1 Score, the proposed architecture demonstrated superior overall F1 Score classification, for renal diagnostics compared to established deep learning models. The system achieved a peak F1-Score of 96.11%, significantly outperforming traditional Transformer networks and Faster R-CNN, which recorded scores of 92.51% and 91.52%, respectively. This high harmonic mean indicates that the model maintains an exceptional balance between precision and recall, effectively minimizing both false positives and missed detections. Furthermore, the proposed framework exhibited a substantial improvement over other standard methodologies, such as RetinaNet at 89.72% and Cascade R-CNN at 86.91%. These results underscore the efficacy of integrating the Multi-Scale Feature Pyramid with specialized attention modules to capture the complex morphologies of kidney tumors.

ISSN: 2178-7727

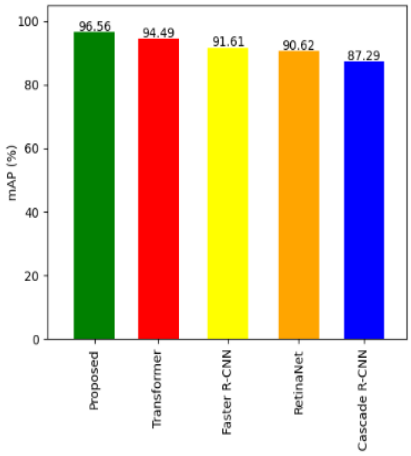

Fig. 6. Comparative Analysis of KT Detection- mAP

Kidney Tumor Detection Comparative Analysis - *mAP* :Figure 6, illustrates a rigorous performance evaluation in terms of mAP, the proposed architecture demonstrated superior overall mAP classification for renal diagnostics compared to established deep learning models.. The proposed system achieved Mean Average Precision (mAP) of 96.56%. This high-performance benchmark indicates that the model maintains exceptional consistency and accuracy across various detection thresholds compared to conventional architectures. In a head-to-head comparison, the proposed model significantly outperformed standard Transformer-based detectors, which recorded an mAP of 94.49%, as well as the Faster R-CNN and RetinaNet frameworks, which achieved scores of 91.61% and 90.62% respectively. Furthermore, the system demonstrated a substantial margin of improvement over the Cascade R-CNN, which yielded a lower mAP of 87.29%. These results underscore the efficacy of combining Multi-Scale Feature Pyramid Transformers with specialized attention modules to capture the complex spatial features required for precise kidney tumor identification.

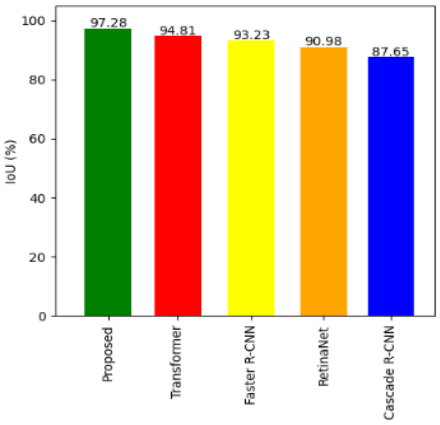

Fig. 7. Comparative Analysis of KT Detection- IoU

Kidney Tumor Detection Comparative Analysis - *IoU :* Figure 7 illustrates a rigorous evaluation of spatial localization accuracy in terms of IoU, the proposed architecture achieved Intersection over Union (IoU) of 97.28%, demonstrating exceptional performance in delineating the boundaries of renal malignancies. This high-fidelity localization significantly outpaces existing methodologies, surpassing the performance of standard Transformer networks, which attained an IoU of 94.81%, and established convolutional frameworks like Faster R-CNN at 93.23%. Furthermore, the system exhibited a substantial improvement over other benchmark models such as RetinaNet, which recorded 90.98%, and Cascade R-CNN, which yielded a lower localization score of 87.65%. These results underscore the efficacy of the integrated feature fusion and attention mechanisms in minimizing the geometric offset between predicted bounding boxes and medical ground truths, providing radiologists with highly accurate spatial data for therapeutic planning..

ISSN: 2178-7727

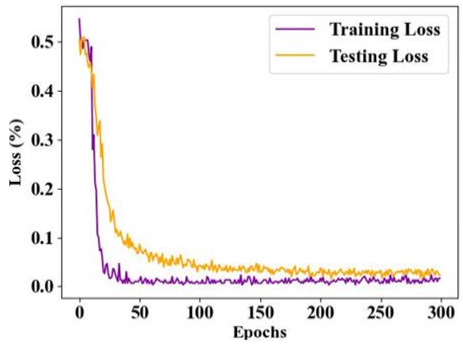

Fig. 8. Train- Test Loss curve

Train Test Loss Curve: Figure 8, illustrates Train-Test Loss curve of the architecture demonstrating exceptional convergence characteristics. During the initial 50 epochs, both training and testing loss exhibited a sharp exponential decay, plummeting from a baseline above 0.5% to approximately 0.1%, indicating rapid model stabilization. As the training progressed toward 300 epochs, the training loss reached a near-zero steady state, while the testing loss consistently mirrored this trajectory, stabilizing just below the 0.1% threshold. This minimal generalization gap between the two curves signifies that the model successfully learned complex renal features without succumbing to overfitting, a result bolstered by the calibrated 0.001 initial learning rate and 0.2 dropout rate specified in the implementation details.

CONCLUSION

In this research,implemented "An efficient Multi-Scale Hybrid Transformer Network for high precision renal malignancy identification". The model successfully engineered and validated as a high-performance solution for the automated detection of kidney tumors within CT imaging. By synthesizing a lightweight transformer-based backbone with specialized spatial attention mechanisms, the study effectively addressed the traditional trade-off between computational efficiency and diagnostic precision. To mitigate the constraints of a limited primary dataset, an extensive data augmentation protocol was utilized to ensure the model achieved robust generalization across diverse clinical scenarios. The integration of the SFAA and HFATrans modules proved pivotal, as they refined post-neck feature outputs to isolate pathological signatures with extreme accuracy. Empirical testing confirmed the framework's clinical viability, yielding a state-of-the-art Intersection over Union (IoU) of 97.28%, a Precision of 96.11%, a Recall of 95.12%, an F1-Score of 96.11%, and a mean Average Precision (mAP) of 96.56%. These results, supported by a stable Train-Test Loss curve, substantiate the effectiveness of hybrid transformer architectures in early-stage oncological screening, with future efforts aimed at evolving the system into a multi-class classifier for the histological differentiation of RCC subtypes.

REFERENCES

1. Alzu'bi, D., Abdullah, M., Hmeidi, I., AlAzab, R., Gharaibeh, M., El-Heis, M., Almotairi, K. H., Forestiero, A., Hussein, A. M. A., and Abualigah, L., 2022. Kidney tumor detection and classification based on deep learning approaches: a new dataset in CT scans. *Journal of Healthcare Engineering*, *2022*(1), 3861161.
2. Abdelrahman, A., and Viriri, S., 2022. Kidney tumor semantic segmentation using deep learning: A survey of state-of-the-art. *Journal of Imaging*, *8*(3), 55.
3. Mahmud, S., Abbas, T. O., Mushtak, A., Prithula, J., and Chowdhury, M. E. H., 2023. Kidney cancer diagnosis and surgery selection by machine learning from CT scans combined with clinical metadata. *Cancers*, *15*(12), 3189.
4. Kumar, Y., Brar, T. P. S., Kaur, C., and Singh, C., 2024. A comprehensive study of deep learning methods for kidney tumor, cyst, and stone diagnostics and detection using CT images. *Archives of Computational Methods in Engineering*, *31*(7), pp.4163-4188.
5. Zhang, M., Ye, Z., Yuan, E., Lv, X., Zhang, Y., Tan, Y., Xia, C., Tang, J., Huang, J., and Li, Z., 2024. Imaging-based deep learning in kidney diseases: recent progress and future prospects. *Insights into Imaging*, *15*(1), 50.
6. Choudhari, K., Sharma, R., and Halarnkar, P., 2020. Kidney and tumor segmentation using U-Net deep

ISSN: 2178-7727

learning model. In *5th International Conference on Next Generation Computing Technologies (NGCT-2019)*, Vol. 3527410.

7. Patel, V. V., Yadav, A. R., Jain, P., and Cenkeramaddi, L. R., 2024. A Systematic Kidney Tumor Segmentation and Classification Framework using Adaptive and Attentive-based Deep Learning Networks with Improved Crayfish Optimization Algorithm. *IEEE Access*.
8. Kadhim, D. A., and Mohammed, M. A., 2025. Advanced Machine Learning Models for Accurate Kidney Cancer Classification Using CT Images. *Mesopotamian Journal of Big Data*, *2025*, pp.1-25.
9. Komal, K. M., Chakkaravarthy, M., and Karras, D. A., 2025. Advancing Deep Learning Techniques for Early Detection and Classification of Renal Cell Carcinoma: A review. *SGS-Engineering & Sciences*, *1*(1).
10. Taha, E. T. E. M., and Mustafa, T. E. K. E., 2024. Enhanced deep learning based decision support system for kidney tumour detection. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, *4*(2), 100174.
11. 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.
12. Masal, K., Chakkaravarthy, M., and Karras, D. A., 2025. Hybrid Deep Pyramid Convolutional Coordinate Attentional Residual Autoencoder Network for Kidney Tumor Diagnosis from CT Scans. *SGS-Engineering & Sciences*, *1*(2).
13. Habchi, Y., Kheddar, H., Himeur, Y., Boukabou, A., Atalla, S., Mansoor, W., and Al-Ahmad, H., 2024. Deep transfer learning for kidney cancer diagnosis. *arXiv preprint arXiv:2408.04318*.
14. Sachin, S., Midhun, C., Karras, D. A., and Masal, K. M., 2025. Enhanced Kidney Tumor Detection using hybrid Deep MSFPT Transformer in computed Tomography images. *SGS-Engineering & Sciences*, *1*(4).
15. Masal, K. M., Bhatlawande, S., and Shingade, S. D., 2023. Hybrid Deep Artificial Humming Bird Algorithm For Improved Real Time Blind Assistance with Advanced Jetson Nano GPU. In *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 99-104. IEEE.
16. Masal, K. M., Bhatlawande, S., and Shingade, S. D., 2023. Deep Learning Attentional Dense based Indoor Object Recognition for Visually Impaired People. In *2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 658-663. IEEE.
17. Shingade, S. D., Mudhalwadkar, R. P., and Masal, K. M., 2022. Random Forest, DT and SVM Machine Learning Classifiers for Seed with Advanced WSN Sensor Node. In *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pp. 321-326. IEEE.
18. Masal, K. M., Bhatlawande, S., and Shingade, S. D., 2024. Attention-based Deep Learning model for Indoor Object Recognition Framework for Visually Impaired Individuals. In *2024 IEEE International Conference on Communication, Computing and Signal Processing (IICCCS)*, pp. 1-6. IEEE.
19. Gavali, P., Chawda, P., Joshi, S., and Masal, K., 2024. Comprehensive analysis of variants in generative adversarial networks. In *AIP Conference Proceedings* (Vol. 3156, No. 1, p. 060008). AIP Publishing LLC.