ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

# Swin-HyConMamba: An Explainable Dual-Stream Fusion Framework with Cross-Attention for Kidney Pathology Classification

Sajid Ali[1]; Yihong Zhang[1,2*]; Sajad Ul Haq[3]; Ameer Hamza[1]; Ran Yao Yao[1]

[1]School of Information and Intelligent Sciences, Department of Electrical and Electronic Engineering, Donghua University, Shanghai 201620, China
[2]School of Information and Intelligent Sciences, Engineering Research Center of Digitized Textile & Fashion Technology, Ministry of Education, Donghua University, Shanghai 201620, China.
[3]School of Information and Intelligent Sciences, Department of Computer Science and Technology, Donghua University, Shanghai 201620, China

Correspondence Author: Yihong Zhang

Publication Date: 2026/04/08

**Abstract: Kidney disease is among the leading causes of morbidity worldwide, and early accurate diagnosis is critical for effective treatment. Medical imaging analysis has increasingly relied on convolutional neural networks (CNNs) and transformer-based models for disease identification. Transformers excel at global context representation but tend to lose fine-grained local detail, while CNNs are strong on local feature extraction but struggle with long-range dependencies. We propose Swin-HyConMamba, a dual-branch framework that combines the strengths of both. The Swin Transformer branch extracts hierarchical global contextual representations, while the HyConMamba branch handles local feature modeling and sequential dependency learning through convolutional and state-space operations. A cross-attention fusion module connects the two branches, enabling the model to attend to clinically relevant features while down-weighting background noise. We evaluate the model on the publicly available Kaggle kidney dataset, covering four classes: normal, cyst, stone, and tumour. The model achieves 99.84% classification accuracy, 99.9% micro-AUC-ROC, and 99.81% macro-average precision, recall, and F1-score, outperforming existing methods. Saliency maps and LIME are used to identify the image regions driving predictions, confirming that the model attends to pathologically relevant areas.**

***Keywords****: Kidney Disease Classification, Swin Transformer, HyConMamba architecture, Cross-Attention Fusion, Explainable AI, Medical Imaging.*

**How to Cite:** Sajid Ali; Yihong Zhang; Sajad Ul Haq; Ameer Hamza; Ran Yao Yao (2026) Swin-HyConMamba: An Explainable Dual-Stream Fusion Framework with Cross-Attention for Kidney Pathology Classification. *International Journal of Innovative Science and Research Technology*, 11(3), 3552-3569. https://doi.org/10.38124/ijisrt/26mar1555

## I. INTRODUCTION

Kidney failure has emerged as a major health concern worldwide, involving diverse pathological conditions that significantly affect patient mortality and morbidity[1]. For successful clinical care and treatment planning, accurate classification of normal tissues, cysts, tumors, and kidney stones is crucial. Computed tomography (CT) imaging is recognized as the definitive standard for identifying kidney disease, providing a broad anatomical view that aids in both preoperative evaluation and postoperative evaluation[2]. Automated technologies utilizing CT imaging for the evaluation of kidney pathology have significantly increased diagnostic accuracy and reduced the workload of radiologists recently[3-5]. These systems use machine learning (ML) and deep learning (DL) algorithms for multi-class classification of kidney-related conditions. Convolutional neural networks (CNNs) have been widely used as the primary framework for medical image classification. However, their ability to capture long-range spatial dependencies is limited[6], which reduces their usefulness in detecting complex anatomical patterns and subtle disease abnormalities in kidney CT images[7]. Furthermore, CNNs may struggle to focus on relevant features, as they may be distracted by background noise, image artifacts, or areas outside the specific region of interest[8]. This obstacle is particularly challenging when it comes to visually similar diseases, such as differentiating between complex cysts and cystic tumors or detecting small kidney stones in complex anatomical settings. In this context, vision transformers (ViTs) emerge as a viable solution to

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

these problems by using self-focusing mechanisms to capture global dependencies in complete images[9]. The Swain transformer extends this technique through hierarchical architecture and shifts the focus to a window-based one, achieving computational efficiency while maintaining robust modeling capability[10, 11]. However, pure transformer topologies have some drawbacks, such as high computational cost, insufficient inductive bias, and limited translational transformation, which limit their utility in medical imaging contexts characterized by a lack of annotated data. To address these challenges, recent study has been focused on hybrid approaches that amalgamate the mutually beneficial effects of CNNs and transformers with self-attentions mechanism at present[12, 13], approaches either use sequential integration or static fusion methods that cannot allow separate architectural streams to interact with each other in real time. These hybrid models often rely on basic concatenation or additive processes, neglecting the need for adaptive multiscale feature fusion and bidirectional information flow. It is seen most clearly when considering aspects such as different kidney pathologies that demand understanding not only fine local details (texture patterns, border characteristics) but also global context information such as (spatial relationships or anatomical context).

To overcome these challenges, our model utilizes a cross-attention fusion method that enables bidirectional feature recalibration between Swin-Tiny and HyConMamba streams. This method dynamically links local pathological features, such as how stones form and how cysts' textures change, to a more general anatomical context, including where the lesions are and how they relate to the surrounding tissue. This is crucial for accurately classifying different types of kidney disease. The cross-attention technique allows each stream to query relevant information from the other, which makes the model much better at classifying complicated and heterogeneous kidney diseases. The primary objective of this study is to combine multi-scale features from the Swin-Tiny transformer and the HyConMamba architecture to better capture local and global pathological traits. This will increase the accuracy of diagnosing kidney disease in four classes: normal, cystic, tumor, and stone. The proposed technique employs a dual-stream design with two parallel branches. The Swin-Tiny branch employs a hierarchical vision transformer that was trained from the ground up, with four steps that make it increasingly difficult to extract multi-scale local features via window-based self-attention approaches. The HyConMamba branch introduces a new hybrid architecture that combines depthwise convolutions with State Space Model-inspired transformations, as well as learnable state transition matrices, to effectively capture long-range relationships and global contextual information To highlight the pathology-related regions, we employ a foreground attention module that generates a spatial mask to focus on clinically important regions and then applies information from the multi-scale relevance module to enable the model to understand both fine details and broader structures. These features are further refined by residual convolutional blocks before being passed to classification layers for final diagnosis in four types of kidney disease.

Moreover, to improve model interpretability, we incorporate Explainable AI (XAI) methods such as saliency maps and Local Interpretable Model-agnostic Explanations (LIME) to better clarify, identify, and illustrate the model's decision-making process steps. These approaches allow us to better identify and understand the particular part of the CT scans that contributes heavily to model predictions, such as boundaries of lesions, internal features, and surrounding anatomical characteristics. The resulting saliency maps show which specific areas of our input CT scans are relevant for predicting outcomes from the Swin-Tiny and HyConMamba models, as well as from our hybrid model architecture. From the saliency maps, we infer that our proposed hybrid architecture method better directs more precise and targeted model attention at critically important regions of diseases than individual model pathways. Meanwhile, using visualizations from Explainable AI, we demonstrate that our proposed hybrid architecture method helps to better eliminate unnecessary background noise/artifacts and makes more prominent model features out of critically important regions like calcification of stones, fluid properties in cysts, irregular enhancements of tumors, and typical parenchymal properties. In addition to that, using Explainable AI, we improve model interpretability and clinical trust by informing medical professionals that the model is focused on diagnostically relevant regions, enabling clinical adoption and incorporation into diagnostic procedures.

➢ *The Main Contributions of this Study are as Follows:*

- In this paper, we proposed A dual-stream framework for classification in kidney pathology images is introduced, integrating both Swin-Tiny and HyConMamba through bidirectional cross-attention. This approach facilitates the adaptive fusion of global contextual information and accurate local features for the classification of Normal, Cyst, Tumor, and Stone objects.
- We introduce HyConMamba, a hybrid convolutional and SSM-inspired architecture enhanced with CBAM attention, engineered to effectively capture long-range spatial dependencies while improving emphasis on diagnostically pertinent regions and minimizing interference from extraneous background information.
- We applied explainable AI methods, such as saliency maps and LIME, to offer visual explanations of the decision-making processes for individual streams and the integrated model, illustrating enhanced localization of key pathological features, including tumor margins, cystic areas, calcifications, and normal kidney structures.

The rest of the paper is organized as follows: Section 2 discusses related studies on kidney disease diagnosis. Section 3 discusses the proposed system. Section 4 describes the experimental conditions and results. Section 5 represents the conclusion.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

## II. LITERATURE REVIEW

Recent rapid improvements in deep learning have led to a major surge in research interest in kidney disorders, kidney stones, and artificial intelligence-based automated diagnosis. Deep learning models, particularly Mamba-based state space models and Swin-Transformer architectures, are significantly increasing the diagnostic accuracy for glomerular analysis, eGFR progression prediction, and renal disease detection. These models exhibit great promise for early detection of kidney-related conditions, such as chronic kidney disease. Their efficacy stems from their capacity to simultaneously model global contextual information and capture fine spatial features, which is crucial for accurate kidney disease classification.

➢ *Transformer with CNN-Based Approaches*

Transformer-based models have demonstrated considerable promise in the diagnosis of kidney disease, especially when used with CT imaging, as demonstrated by Islam et al.[14] using a Kaggle dataset they gathered from several hospitals in Dhaka, Bangladesh. They compared six deep learning models, including transformer-based techniques like EANet, CCT, and swin-transformer, as well as traditional CNNs like ResNet, VGG16, and InceptionV3, and presented an AI-driven system for the automatic classification of kidney stones, cysts, and cancer. A high accuracy of 99.30% was attained by the Swin Transformer, and the explainability research showed that VGG16 performed better in capturing clinically relevant anatomical areas. At the same time, Shearen et al.[15] suggested MSKd_Net to classify four types of kidney disease from CT scans, with Swin Transformer as the foundational framework. It combines a convolutional module with Swin Transformer features and classification-based multi-head attention that collects local and global information. The model was evaluated using the kaggle dataset, with F1 scores of 0.97 for cysts, 0.99 for normal cases, 0.98 for stones, and 0.90 for tumors, indicating quick and reliable kidney disease diagnosis. Another study, Martin et.al [16] analyzed vision transformer models such as Swin Transformer and MaxVT for multi-organ tumor classification using CT scans and MRI images of the brain, lung, and kidney. The performance of the Swin Transformer model was quite good, with an average accuracy of 99.0% for individual tasks and 99.43% for the ensemble of all datasets. This experiment highlights the ability of transformer models to generalize across different organs and different images. Huang et al.[17] suggested CSTCM, a convolutional Swin Transformer-based model, for diagnosing kidney stones in CT images using the IoMT framework. The model was accurately tested on a dataset of 12,446 clinically diagnosed CT slices from multiple hospitals in Dhaka, Bangladesh, comprising both normal and kidney stone patients. CSTCM outperformed ViT, DeiT, ConvNeXt, and the typical Swin Transformer models, with 97.6% accuracy and 97.2% AUC. The Reciprocal domain adaptation network introduced by Iqbal et al.[18] aims to improve CKD diagnosis in different areas of the dataset domain. The RDAN models were tested in the kidney dataset of 12,446 labeled images and obtained 96.94% accuracy, and a 99.35% AUC by employing global and local pyramid pooling and adaptive models, respectively, to effectively learn from each domain and gain domain knowledge from each other's models. Rahman et al.[19] employed a hybrid technique that included Swin-VIT, DeepLabV3+, and numerous pre-trained CNN layers to diagnose and grade kidney cancer with multiple labels. They obtained an accuracy score of 0.992 using the kidney Kaggle dataset and used Grad-CAM for improved description. A dual-task network framework using a transformer architecture was proposed by Conze et al.[20] for segmentation of polycystic kidneys in patients with ADPKD using MRI images. Their method incorporates a common encoder with kidney-specific decoders and performs well on 112 patients from the Genkyst1 cohort with a DICE accuracy of 93.4% and a better threshold than single-task methods. To capture complex feature interactions and enhance picture contrast, Eliazer et al.[21] used Twins-SVT for feature extraction and an attention-guided BiTCN-BiLSTM model for classification. On the RCCGNet dataset of H&E-stained images, the system outperformed previous attempts with 98.26% classification accuracy, suggesting the possibility of automated RCC diagnosis.

➢ *Mamba State-Space with CNN Based Approaches*

In recent research, the combination of CNNs and Mamba has been shown to effectively learn long-range relationships in data, which is critical for the exact diagnosis of different illnesses utilizing medical imagery. Notably, the use of Mamba in CNNs has substantially enhanced the identification of kidney disease. In this context, Pan et al.[22] applied the CNN-Mamba-WOA technique to a large dialysis dataset of 758 patients and 98,015 sessions, encompassing 24 clinical characteristics. The approach includes a CNN structure, a Mamba state-space component, and the Whale Optimization Algorithm for parameter optimization, obtaining 0.823 AUC values and 0.902 accuracy, with SHAP values allowing clinical interpretation. At that time, Lu et al.[23] suggested an Efficient Vision Mamba–Transformer hybrid system for abdominal multi-organ segmentation. The model captures global context and channel-wise interactions with an Efficient Vision Mamba (EViM) module in a Transformer encoder. On Synapse and ACDC datasets, it achieved an average Dice score of 82.67% and HD95 of 16.36 mm, proving accuracy, computational efficiency, and generalizability across imaging modalities. However, PCU-SABENet, a Mamba framework for kidney lesion subtype classification on multi-phase CT images, has been presented by RMR S S et al.[24] Using the TCIA dataset, the model achieved an accuracy of 99.3% and a Dice score of 94.8%, demonstrating outstanding resilience and generalization capabilities by using phase-aware U-MAMBA and EfficientNet-B7 classification to fill in missing phases. In order to balance the accuracy and efficiency of medical imaging jobs, Qamar et al.[25] designed SAMA-UNet, a u-shaped model. They use a self-adaptive mamba-like aggregated attention block to combine local and global information, and they use a causal resonance multi-scale module to trace causal relationships. Using datasets from MRI, CT, and endoscopy, this method obtained 92.16% DSC and 96.54% NSD. The FM-Mamba CT imaging model by Zhang et al.[26] investigated long-term relationships using

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

frequency domain feature extraction and state space sequence models. For the CT-ORG dataset (140 patients and six organs), FM-Mamba had an average Dice coefficient of 93.86%, better than CNN and transformer models despite having fewer parameters. Su et al.[27] proposed the VMKLA-UNet framework, which integrates vision mamba and KAN linear and channel-spatial attention methods for medical images. The encoder utilised global background information, while the decoder highlighted the important regions. After testing on multiple datasets, including Kidney/ColonDB, it showed excellence with values of 73.53% and mIoU of 64.90%, higher than previous baselines.

Despite recent breakthroughs, research into the hybrid CNN-Mamba and Transformer models for kidney disease detection and classification remains active. The majority of existing research on kidney disease identification is focused on individual CNN models, transformer-based models, and hybrid models of the two and is limited to specific datasets. Despite their excellent accuracy and efficiency, these models nevertheless face hurdles in terms of dataset generalization, multiphase image processing, and computational costs.

## III. PROPOSED DUAL-STREAM ARCHITECTURE WITH CROSS-ATTENTION FUSION

In this study, we propose a dual-stream Swin–HyConMamba fusion model to improve kidney disease classification by combining local structural features and global context representation. As shown in Fig. 1, input CT images are first preprocessed and then processed in parallel by two branches. The Swin-Tiny branch employs hierarchical patch embedding and shifted window self-attention across four stages to capture multi-scale global context. In contrast, the HyConMamba branch utilizes four SSM-inspired blocks with convolutional downsampling to preserve fine-grained local details while modeling sequential dependencies. A dropout layer is applied to reduce overfitting after each branch. Spatial-attention is applied to Hyconmamba and Swin-tiny global features, respectively. Spatial-attention attention highlights the most informative kidney regions, while self-attention refines the kidney's local context and structure. Subsequently, a bidirectional cross-attention fusion module integrates complementary information from both branches. Finally, the fused representation is passed through global average pooling and fully connected layers to classify kidney CT images into Normal, Cyst, Stone, and Tumor categories.

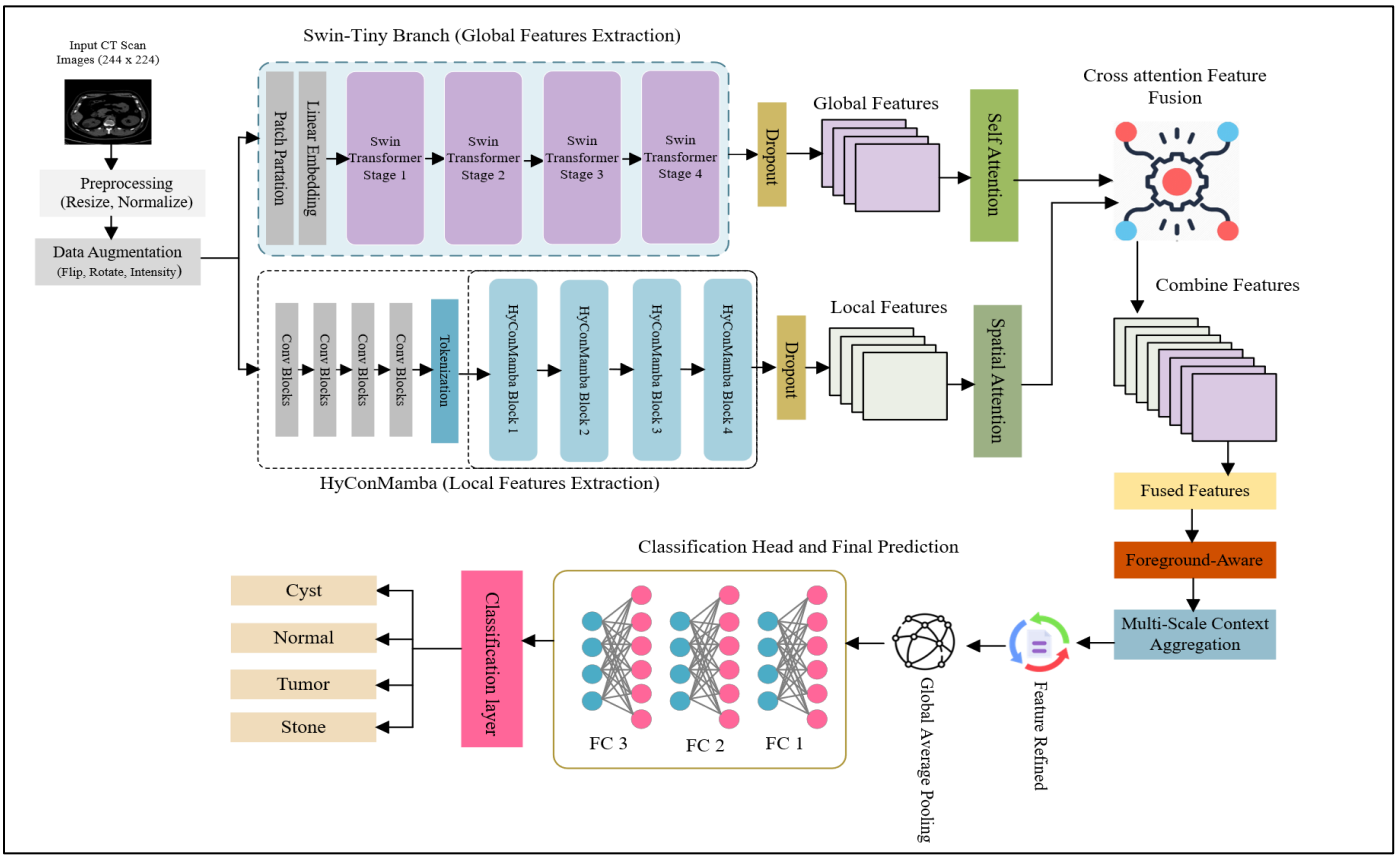



Fig 1 Schematic Overview of the Proposed Swin–HyConMamba Network, Illustrating Dual-Branch Feature Extraction and Fusion.

➢ *Preprocessing and Data Augmentation*

To enhance the quality and resilience of the input kidney CT images and the stability of the training, we implemented various preprocessing techniques. Initially, we resized the images to a final resolution of 224×224 pixels to align with the model input size. To match the majority, we

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

oversample the training class while maintaining an imbalance in the validation and test sets to represent real-world clinical situations. Next, employ data augmentation to increase diversity during training. To improve data variability, random cropping and horizontal flipping with a frequency of 0.5 were used. Lastly, ImageNet-based channel statistics were used to normalize the images once they were converted to tensors. The mean μ=[0.485,0.456,0.406] and an experimentally optimized standard deviation σ=[0.250,0.250,0.250] were chosen to better correspond with the intensity distribution of kidney CT images. The normalization operation is defined as follows as equation (1).

$$X' = \frac{X - \mu}{\sigma} \quad (1)$$

Where μ and σ are the mean and standard deviation vectors. This preprocessing improves data split distribution consistency, accelerating convergence and generalization.

➢ *Swin-Tiny Branch: Hierarchical Feature Extraction*

The proposed model's Swin-Tiny branch employs a hierarchical vision transformer architecture that was trained totally from the ground up, without the use of pretrained weights. This branch is designed to capture multi-scale global representations and spatial relationships necessary for understanding organ-level context in kidney pathology classification. As shown in Fig. 2, the Swin-Tiny branch consists of a patch embedding module followed by four progressive hierarchical stages with depths [2, 2, 6, 2], which use shifted window-based multi-head self-attention (SW-MSA) methods.

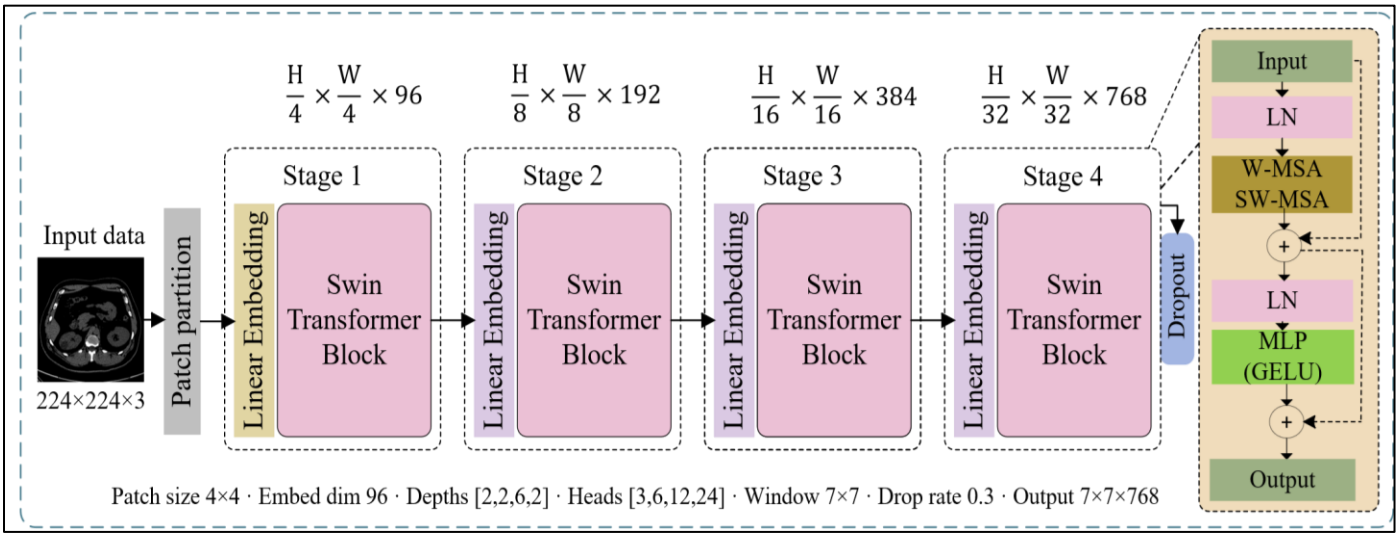



Fig 2 Swin-Tiny Transformer Branch's Architecture, Including Hierarchical Phases with Shifted Window Self-Attention and Progressive Multi-Scale Feature Extraction.

Given an input CT image $X \in \mathbb{R}^{H\times W\times C}$,, where $H = 224$, $W = 224$, and $C = 3$, the image is partitioned into non-overlapping patches of size $4 \times 4$. Unlike traditional linear patch embedding, the implementation employs a convolutional projection to enhance local feature extraction. Specifically, a convolutional layer with kernel size 4×4 and stride 4 is applied to map each patch into a 96-dimensional embedding space. The patch embedding operation is defined as equation (2).

$$P_{\text{embed}} = \text{LayerNorm}(\text{Conv}_{4\times4}(X)) \quad (2)$$

Where $P_{\text{embed}} \in \mathbb{R}^{3136\times96}$ represents the embedded patch tokens corresponding to a spatial resolution of 56×56. Window-based attention and relative position bias automatically capture spatial relationships; hence this design does not use absolute positional embeddings (APE = False). The embedded tokens undergo processing through four hierarchical stages of the Swin Transformer, during which spatial resolution is incrementally reduced but feature dimensionality is enhanced. Stage 1 uses three attention heads, an embedding dimension of 96, and a resolution of 56x56. This stage consists of 2 Swin Transformer blocks implementing window-based multi-head self-attention with a window size of 7×7, alternatively using regular and shifted window partitioning to facilitate cross-window interactions. Stage 2 employs a patch merging process that diminishes the spatial resolution to 28×28 while augmenting the embedding dimension to 192. This stage utilizes 2 Swin Transformer blocks featuring 6 attention heads. Stage 3 decreases spatial resolution to 14×14 and increases embedding dimension to 384. Six Swin Transformer blocks with 12 attention heads are used to extract deeper and more abstract representations. Stage 4 features 7×7 resolution, 768-dimensional embeddings, and 24 attention heads. Two Swin Transformer blocks are used to create high-level semantic characteristics. Patch merging layers between stages combine neighboring 2x2 tokens, reducing spatial resolution while boosting channel capacity. This enables efficient hierarchical representation learning. Each Swin Transformer block utilizes window-based multi-head self-attention (W-MSA), as described in equation (3).

$$\text{Attention}(Q,K,V) = \text{SoftMax}\left(\frac{QK^{\top}}{\sqrt{d_k}} + B\right)V \quad (3)$$

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

Here, $Q, K, V \in \mathbb{R}^{M^2 \times d}$ represent the query, key, and value matrices generated within each $M \times M$ window, $d_k$ denotes the dimensionality of each attention head and $B \in R^{M^2 \times M^2}$ signifies the learnable relative position bias matrix that encapsulates spatial correlations among tokens within a window. Shifted window attention is employed in alternating blocks to improve the exchange of information between adjacent windows, thereby improving the modeling of long-range spatial relationships. Finally, layer normalization is used to generate 49 spatial tokens on a 7x7 grid with 768-dimensional features. In addition, dropouts with a probability of 0.3 are used to prevent overfitting. This branch captures hierarchical global representations that encode spatial relationships, organ-level context, and coarse anatomical structures essential for understanding kidney pathology.

- *Hyconmamba Branch: Local Feature Extraction and Spatial Attention*

The HyConMamba branch captures local features, including textural variations, lesion boundaries, and fine anatomical structures, unlike standard self-attention mechanisms that have quadratic computational complexity, this branch uses a state space model-inspired architecture[28], as well as four progressive convolutional downsampling blocks and four HyConMamba blocks, which include projections multiscale depth-wise convolution, state space transformation, gating mechanism, and dropout layer as shown in Fig. 3 & 4.

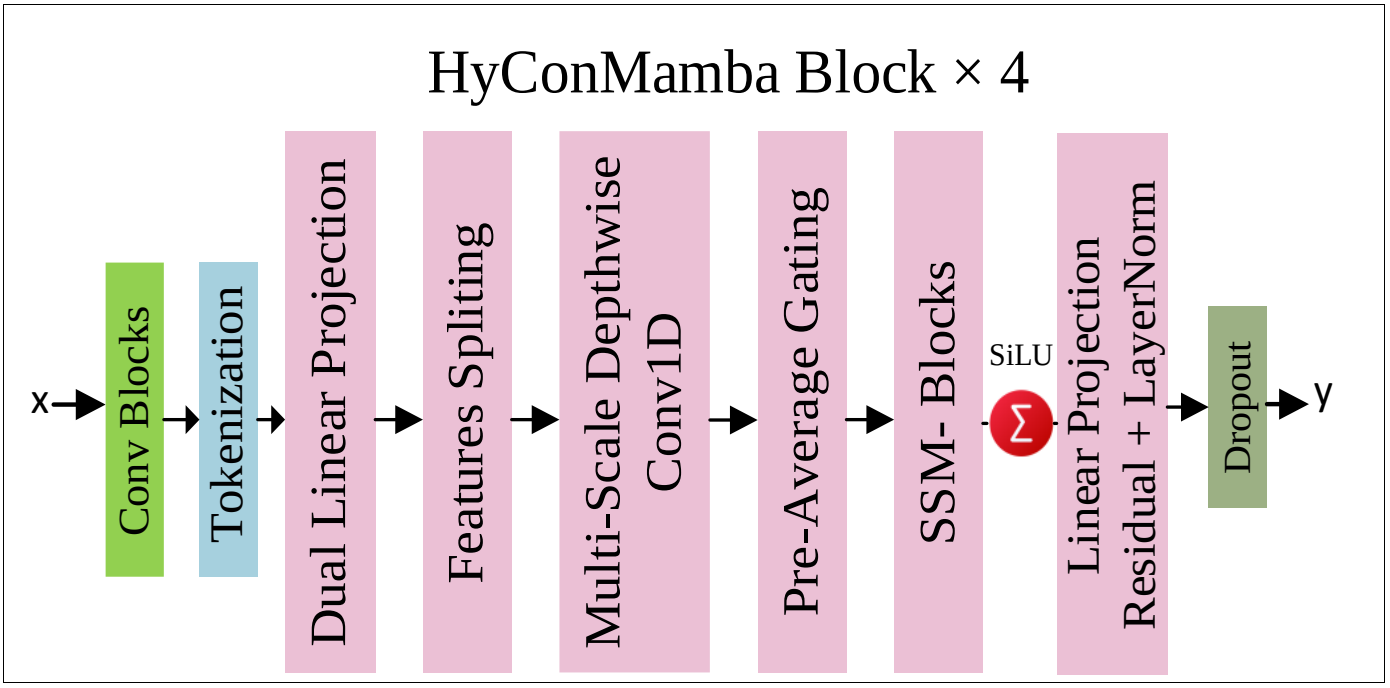



Fig 3 Architecture of the HyConMamba Branch for Local Feature Extraction.

The proposed encoder processes the input image $X \in \mathbb{R}^{B \times 3 \times 224 \times 224}$ and produces feature maps with spatial dimensions of $14 \times 14$ and 256 channels. To achieve low computing cost and high global context modeling, this model makes use of an expansion ratio of 2 with eight attention heads, depthwise convolutional kernels of sizes 3, 5, and 7, and state space model parameters A, B, and C.

The HyConMamba branch begins with four successive convolutional downsampling blocks, each consisting of a 3x3 convolutional layer with a stride of 2, batch normalization, and GELU activation, as illustrated in Fig. 4. This enables hierarchical feature extraction, spatial downscaling, and higher channel dimensionality. The first convolutional block applies 32 filters to reduce the input's spatial dimensions from 224x224x3 to 112x112x32. The second convolutional block reduces the spatial dimensions of the feature maps to 56x56x64 using 64 filters. The third block generates 28×28×128 feature maps using 128 filters. The fourth and final convolutional blocks are utilized to lower the spatial dimensions of the feature maps to 14x14x256, which is the specified spatial dimension for the HyConMamba encoder. Batch normalization is implemented to reduce internal coefficient shift, whereas GELU is used to address non-linearity and improve the acquisition of complicated pathology-specific features. The input image is denoted as $X \in \mathbb{R}^{H \times W \times C}$, where $H = 224$, $W = 224$, and $C = 3$, and is processed through the convolutional downsampling blocks defined in equation (4).

$$X_{down} = GELU\big(BatchNorm(Conv2D(X, W) + b)\big) \quad (4)$$

Where $X_{\text{down}}$ is the result of the convolutional downsampling, Conv2D is a 3×3 convolution with stride 2, $W$ is the learnable weight matrix, and $b$ is the bias. This progressive downsampling method reduces the computation cost from 224×224 to 14×14 while keeping the spatial structure needed for disease identification.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

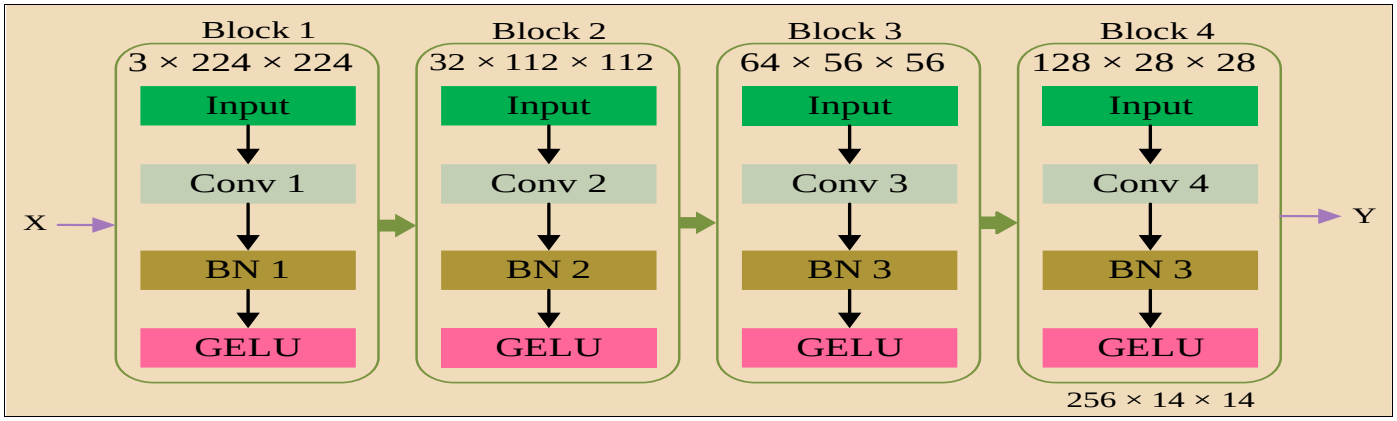



Fig 4 Progressive Downsampling Using Stride-2 Convolutions, Batch Normalisation, and GELU Activation is Carried Out via a Four-Stage Convolutional Encoder.

The spatially downsampled features are transformed from $X_{down} \in \mathbb{R}^{B\times256\times14\times14}$ to a $196 \times 256$ sequence representation, where $196 = 14 \times 14$ is the number of spatial tokens. These tokens are then processed through four sequential HyConMamba blocks. Each HyConMamba block begins with dual input projections that grow the channel dimension by a factor of 2 (expansion ratio = 2), resulting in hidden dimensions of 512 channels. The projected features are separated into feature and gate components as specified in equation (5).

$$\mathbf{Z} = \mathbf{X}\mathbf{W}_{\text{in}}, \mathbf{Z}_v, \mathbf{Z}_g = \text{Split}(\mathbf{Z}) \quad (5)$$

Where $\mathbf{W}_{\text{in}} \in \mathbb{R}^{256\times512}$ is the input projection matrix. Layer normalization is applied to the feature component before multi-scale processing. To capture multi-scale receptive field information, three depthwise convolutions with kernel sizes of 3, 5, and 7 are performed on normalised data. Depthwise convolution reduces processing costs while maintaining spatial pattern recognition by using a single filter for each channel. One-dimensional sequence convolution is facilitated by transforming features into (B, C, L) format, is a described at different scales by equations (6).

$$x_3 = \text{DWConv}_3(x),\ x_5 = \text{DWConv}_5(x),\ x_7 = \text{DWConv}_7(x) \quad (6)$$

Where $\text{DWConv}_k$ refers to depthwise 1D convolution with a kernel size of and padding to maintain sequence length. Kernels of size 3 capture fine local patterns, size 5 models mid-range dependencies, and size 7 captures context. A multi-scale representation is produced by concatenating outputs along the channel dimension using equation (7).

$$x_{\text{concat}} = \text{Concat}([x_3, x_5, x_7]) \quad (7)$$

The output features $x_{\text{concat}} \in \mathbb{R}^{B\times196\times1536}$ combine fine-grained and contextual information for effective pathology classification. Following that, a pre-average gating method is used to highlight informative features before scale aggregation. It keeps importance weights consistent across scales and geographic locations, reducing noise and allowing the model to focus on pathologically significant data. The weight gating is calculated using a linear projection followed by a sigmoid activation, as shown in equation (8).

$$g_{\text{pre}} = \sigma\big(\text{conv\_gate\_pre}(x_{\text{concat}})\big) \quad (8).$$

The concatenated features are further rearranged into the format $(B, L, 3, \text{hidden\_dim})$ to distinguish the three scales, and the expanded gating weights are applied elementwise, as illustrated in equation (9):

$$x_{\text{gated}} = x_{\text{concat\_reshaped}} \odot g_{\text{pre\_expanded}} \quad (9).$$

The gated features are subsequently averaged across the scale dimension by mean pooling to derive a consolidated representation, as expressed in equation (10).

$$x_{\text{avg}} = \text{Mean}\big(x_{\text{gated}}, \text{dim} = \text{scale}\big) \quad (10).$$

The pre-average gating mechanism enables the model adaptively to incorporate multi-scale features. Local contextual interactions that we modeled using learnable parameters that govern state transitions and output transformations. The SSM transformation consists of three successive operations that use learnable parameters $A, B, and\ C \in \mathbb{R}^{\text{hidden\_dim}}$, where $\text{hidden\_dim} = 512$. The average multi-scale features are first controlled using a learnable state transition parameter A, which regulates the temporal decay of state information. This operation is defined in equation (11).

$$x_{\text{ssm}} = x_{\text{avg}} \odot \sigma(A) \quad (11)$$

Here, $\sigma(\cdot)$ signifies the sigmoid activation function, while $\odot$ implies element-wise multiplication. The parameter $A$ is initialized at 0.5 and is adjusted throughout training, enabling the model to adaptively manage information flow throughout the sequence. Next, the state is modified using an input-dependent modulation utilizing parameter $B$, executed through a hierarchical gating mechanism to selectively enhance the hidden representation, as defined in equation (12).

$$x_{\text{ssm}} = x_{\text{ssm}} \odot \sigma(x_{\text{ssm}} \odot B) \quad (12)$$

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

While $B \in \mathbb{R}^{\text{hidden\_dim}}$ is initialized with tiny random values (mean 0, std 0.1), this operation allows the network to update the state selectively based on the input content.

Finally, the state is modulated by an output gating operation controlled by C to generate the final SSM-transformed features, as illustrated in equation (13).

$$x_{\text{ssm}} = x_{\text{ssm}} \odot \sigma(C) \tag{13}$$

Here, $C \in \mathbb{R}^{\text{hidden\_dim}}$ begins with small random values. The SSM output is subsequently connected to the original averaged features through a residual connection and subjected to a SiLU activation to generate nonlinearity. This demonstrates the ultimate representation described in equation (14).

$$x = \text{SiLU}\left(x_{\text{avg}} + x_{\text{ssm}}\right) \tag{14}$$

Afterwards, SSM transformation employed a post-gating technique to further improve the features and augment their discriminative efficacy. Analogous to the pre-average gating, a learned linear projection followed by sigmoid activation produces post-gating weights, which are applied elementwise to the SSM-transformed features, as described in equation (15).

$$x_{\text{gated}} = x_{\text{ssm}} \odot \sigma(\text{conv\_gate \_post } (x_{\text{ssm}})) \tag{15}$$

The gated features are projected back to the original channel dimension (256) via a linear output layer $W_{\text{out}} \in \mathbb{R}^{512\times256}$ and connected to the input through a residual connection reduced by a learnable parameter γ initialized to 0.1, as expressed in equation (16).

$$\text{output} = x_{\text{input}} + \gamma \cdot \text{out\_proj}(x_{\text{gated}}) \tag{16}$$

Post-gating and output projection improve SSM features and restore channel dimensions to enhance differentiation and ensure stable residual learning. After four stacked HyConMamba blocks, dropout layer and self-attention module with 8 heads is used, specifically designed to facilitate token exchange and further enhance global context connectivity. Assuming the token sequence is $\mathbf{T} \in \mathbb{R}^{B\times196\times256}$ the self-attention process is defined as shown in equation (17).

$$\mathbf{A}_{\text{out}} = \text{MHSA}(\mathbf{T}) \tag{17}$$

The attention output is combined with the input tokens through a residual connection, succeeded by layer normalization, resulting in the enhanced token representation as defined in equation (18).

$$\mathbf{T}' = \text{LayerNorm}(\mathbf{T} + \mathbf{A}_{\text{out}}) \tag{18}$$

These representations integrate multi-scale textural patterns via parallel depthwise convolutions, sequential dependencies through state-space modeling. This locally enriched representation complements the global context from the Swin-Tiny branch in the subsequent cross-attention fusion module.

➢ *Cross Attention Feature Fusion: Bridging Local and Global Representations*

Our proposed method employs cross-attention-based feature fusion to effectively combine the global context extracted by the Swin-Tiny branch with the local detail representations from the HyConMamba branch. Swin-Tiny utilizes hierarchical window-based self-attention to capture global spatial patterns, including organ-level structure and inter-region relationships, while HyConMamba leverages SSM-inspired transformations to model fine-grained local dependencies and textural details across the feature map. Rather than simple concatenation, a bidirectional cross-attention mechanism facilitates adaptive interaction between the two feature sets: Swin features attend to HyConMamba features to incorporate local details into the global representation, while HyConMamba features attend to Swin features to embed global context into local features, Fig. 5 depicts in detail the workflow of the proposed cross-attention fusion.

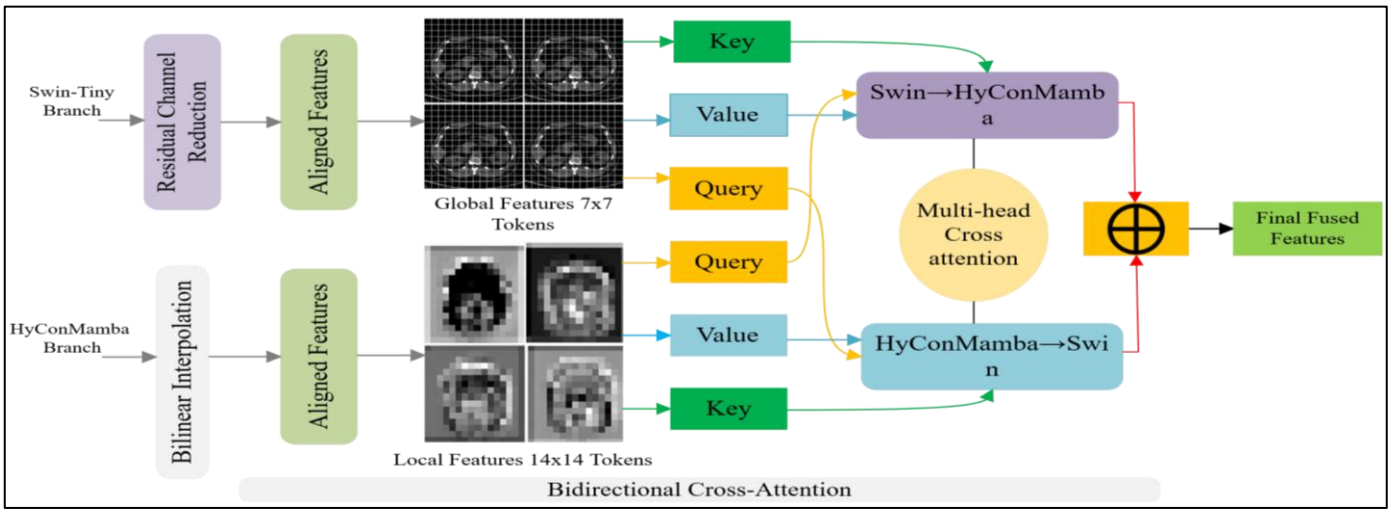



Fig 5 Cross-Attention Feature Fusion for Features Extracted Through Swin-Tiny and HyConMamba.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

The Swin-Tiny branch's output features are $\mathbf{F}_{\text{swin}} \in \mathbb{R}^{B\times768\times7\times7}$ where B is the batch size, 768 is the embedding dimension, and 7×7 signifies global context. In parallel, the HyConMamba branch output are $\mathbf{F}_{\text{hycon}} \in \mathbb{R}^{B\times256\times14\times14}$, which encodes higher spatial resolution. For efficient cross-attention, alignment of both channel dimensionality and spatial resolution is essential. First, A residual channel reduction module is utilized on the Swin features to project the embedding dimension from 768 to 256, expressed as equation (19).

$$F'_{\text{swin}} = F_{\text{swin}} + \gamma\, \phi(F_{\text{swin}} W_r), \tag{19}$$

Where $W_r \in \mathbb{R}^{768\times256}$, $\phi(\cdot)$ denotes convolution followed by batch normalization, and γ=0.1 is a trainable scaling parameter that ensures training stability. In parallel, bilinear interpolation[29] is employed to adjust the HyConMamba features to match the spatial resolution of the Swin features, allowing for the feature alignment described in equations (20).

$$F'_{\text{hycon}} = \text{Interp}\left(F_{\text{Hycon}}\right), \quad F'_{\text{hycon}} \in \mathbb{R}^{B\times49\times256} \tag{20}$$

Finally, both feature maps have been reshaped into token order for cross-attention and aligned as defined in equation (21).

$$F_{\text{swin\_aligned}}, F_{\text{hycon\_aligned}} \in \mathbb{R}^{B\times49\times256}, 49 = 7\times7 \tag{21}$$

These serve as compatible inputs for multimodal feature fusion. To enable bidirectional feature interaction, the aligned features are transformed into the query space (Q), key space (K), and value space (V) using a linear transformation, thus enabling bidirectional attention. In parallel, the Swin features interact with the HyConMamba features the projections and cross-attention are calculated as defined in equations (22) (23) and (24).

$$\begin{aligned} Q_{swin} &= W_Q^{swin} F_{swin\_aligned}, \\ K_{hycon} &= W_K^{hycon} F_{hycon\_aligned}, \\ V_{hycon} &= W_V^{hycon} F_{hycon\_aligned} \end{aligned} \tag{22}$$

$$F_{swin\to kycon} = MultiHeadAttention(Q_{swin}, K_{kycon}, V_{hycon}) \tag{23}$$

$$\text{Attention}(Q_{\text{swin}}, K_{\text{hycon}}, V_{\text{hycon}}) = \text{Softmax}\left(\frac{Q_{\text{swin}} K_{\text{hycon}}^{\top}}{\sqrt{d_k}}\right) V_{\text{hycon}} \tag{24}$$

Likewise, HyConMamba features attend to Swin features in cross attention modules as described by equation (25), (26) and (27).

$$\begin{aligned} Q_{\text{hycon}} &= W_Q^{\text{hycon}} F_{\text{hycon\_aligned}}, \\ K_{\text{swin}} &= W_K^{\text{swin}} F_{\text{swin\_aligned}}, \\ V_{\text{swin}} &= W_V^{\text{swin}} F_{\text{swin\_aligned}} \end{aligned} \tag{25}$$

$$F_{\text{hycon}\to\text{swin}} = \text{MultiHeadAttention}(Q_{\text{hycon}}, K_{\text{swin}}, V_{\text{swin}}) \tag{26}$$

$$\text{Attention}(Q_{\text{hycon}}, K_{\text{swin}}, V_{\text{swin}}) = \text{Softmax}\left(\frac{Q_{\text{hycon}} K_{\text{swin}}^{\top}}{\sqrt{d_k}}\right) V_{\text{swin}} \tag{27}$$

Where $Q_{\text{swin}}, K_{\text{swin}}$,and $V_{\text{swin}}$denote the query, key, and value embeddings derived from the Swin-Tiny features, respectively, and $Q_{\text{hycon}}, K_{\text{hycon}}$, and $V_{\text{hycon}}$ represent the corresponding query, key, and value embeddings of the HyConMamba features. The projection matrices $W_Q$, $W_K$, and $W_V$ are learnable parameters, and MultiHeadAttention(·) utilizes 8 attention heads. Subsequently, to maintain the original discriminative features of each branch while integrating cross-attended information, residual connections along with layer normalization are employed as shown in equation (28), (29).

$$\mathbf{F}_{\text{swin\_fused}} = \text{LayerNorm}(\mathbf{F}_{\text{swin}} + \mathbf{F}_{\text{swin}\to\text{hycon}}) \tag{28}$$

$$\mathbf{F}_{\text{hycon\_fused}} = \text{LayerNorm}(\mathbf{F}_{\text{hycon}} + \mathbf{F}_{\text{hycon}\to\text{swin}}) \tag{29}$$

The refined features from both branches are concatenated and processed through a fusion MLP to obtain a unified feature embedding as shown in equation (30).

$$\mathbf{F}_{\text{fusion}} = \text{MLP}(\text{Concat}[\mathbf{F}_{\text{swin\_fused}}, \mathbf{F}_{\text{hycon\_fused}}]) \tag{30}$$

The fused features are refined using a multilayer perceptron consisting of three linear layers. First, the feature dimension is expanded from 512 to 1024, then the GELU activation function and dropout layer are applied to reduce the dimension back to 512, and the same activation function and regularization are applied. The last linear layer projects the features to 256 dimensions, producing a compact representation for further tasks. Finally, the fused representation is normalized using layer normalization, as described in equation (31).

$$\mathbf{F}_{\text{fused}} = \text{LayerNorm}(\mathbf{F}_{\text{fusion}}) \tag{31}$$

The use of this cross-attention-based fusion is beneficial in that it allows for dynamic interactions between global and local features. Swin-Tiny features that representglobal spatial context can be complemented by HyConMamba's fine-grained local details, while HyConMamba features that capture local textural patterns can be enriched by Swin-Tiny's global structural information. The use of cross-attention enables adaptive feature selection, reduces representational redundancy, and yields more discriminative feature representations for accurate kidney pathology classification.

➢ *Classification Head and Final Prediction.*

After cross-attention fusion, the fused features are then transformed into a 7×7 spatial map for spatially aware processing. Firstly, a foreground attention [30] and a multi-scale context module[31] that uses parallel convolutions to gather both fine-grained details and more general contextual

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

data is added to the attended features. After that, a feature refinement module with residual connections improves discriminative representations and helps to stabilize training. Finally, global average pooling is used to produce a compact feature vector, which is then fed into a fully connected classification head to predict the four kidney pathology classes: normal, cyst, stone, and tumor.

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we describe the experimental setup, dataset, and evaluation criteria used to test the performance of the proposed Swin-HyConMamba fusion model.

➢ *Dataset*

For this study, we used the publicly accessible CT-Kidney Dataset by Islam et al. [14] , from multiple hospitals in Dhaka, Bangladesh. The dataset contains CT scan images across four classes: 5,077 normal kidneys, 3,709 cysts, 2,283 tumors, and 1,377 stones and 12,446 images in total. Data was split into training (60%), validation (25%), and testing (15%) sets using stratified sampling to preserve class distribution. The considerable class imbalance in the original dataset, with the Normal and Cyst classes having far more samples than the Tumor and Stone classes, needed the use of strategic data preparation techniques, which we applied.

➢ *Experimental Setup*

All experiments were conducted in a Kaggle environment using an NVIDIA Tesla P100 GPU to accelerate model training. We used PyTorch with CUDA Automatic Mixed Precision (AMP) to implement the model and train it efficiently. To balance memory constraints and convergence stability, the model was trained in 16-bit batches. We chose the AdamW optimizer because of its adaptive learning rate and weight decay capabilities, starting with an initial learning rate of $3 \times 10^{-5}$and weight decay of $1 \times 10^{-4}$ to prevent overfitting. A Cosine Annealing scheduler was applied to gradually reduce the learning rate to a minimum of $1 \times 10^{-6}$ (1% of the initial rate) over the training period. We used early stopping with an 8-epoch patience, which halts training if no improvement in validation accuracy is observed; the maximum epoch limit was set to 50, though early stopping terminated training at epoch 44. We applied the Cross-Entropy Loss function for the multiclass kidney disease classification task. Dropout layers and L2 regularization were incorporated to enhance generalization and prevent overfitting, and gradient clipping with a maximum norm of 1.0 was used to maintain training stability. To ensure reproducibility, all experiments were seeded at 42.

➢ *Model Performance Analysis*

To assess the effectiveness of our Swin-HyConMamba fusion model, we tracked both training and validation accuracy along with loss curves across epochs. The training loss vs. epochs and validation loss vs. epochs are depicted in Fig. 6, which shows a consistent decline in loss over time, suggesting stable convergence. The accuracy curves further confirm that the model consistently improved with training. Early stopping was applied to prevent overfitting, which terminated training at epoch 44 when no further improvements were observed in the validation accuracy.

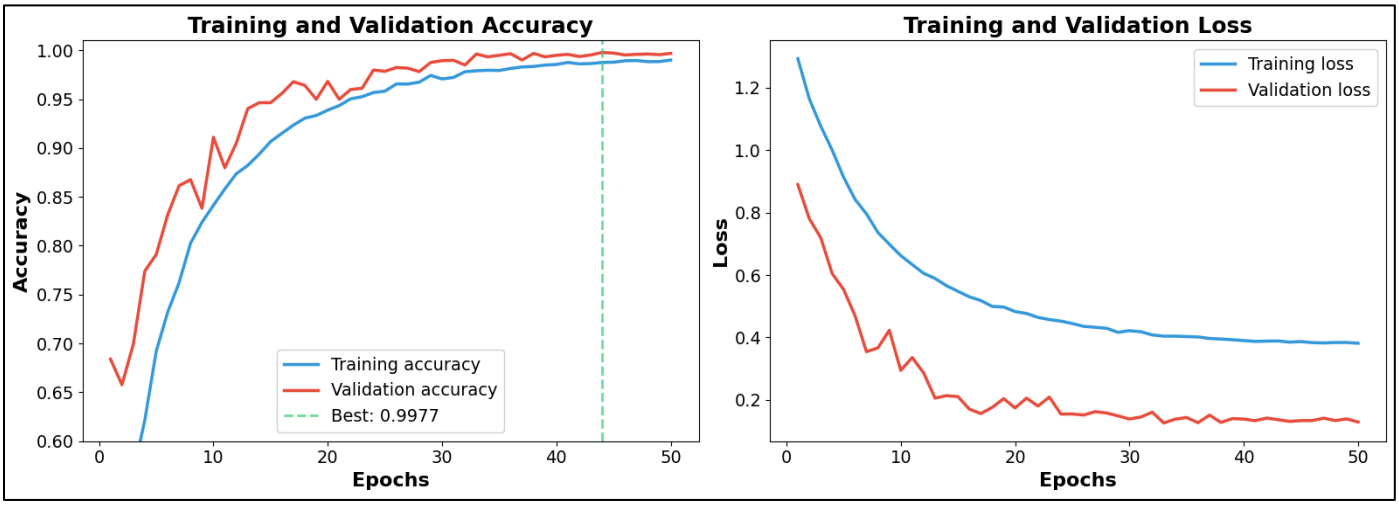



Fig 6 Depicts Train and Validation Accuracy and Loss Throughout Epochs.

Throughout training, the model showed no clear signs of overfitting, as both training and validation loss and accuracy followed consistent decreasing and increasing trends, respectively, with validation loss remaining slightly lower than training loss. The loss curves similarly confirm that underfitting was avoided, as validation loss decreased steadily without plateauing at elevated values.

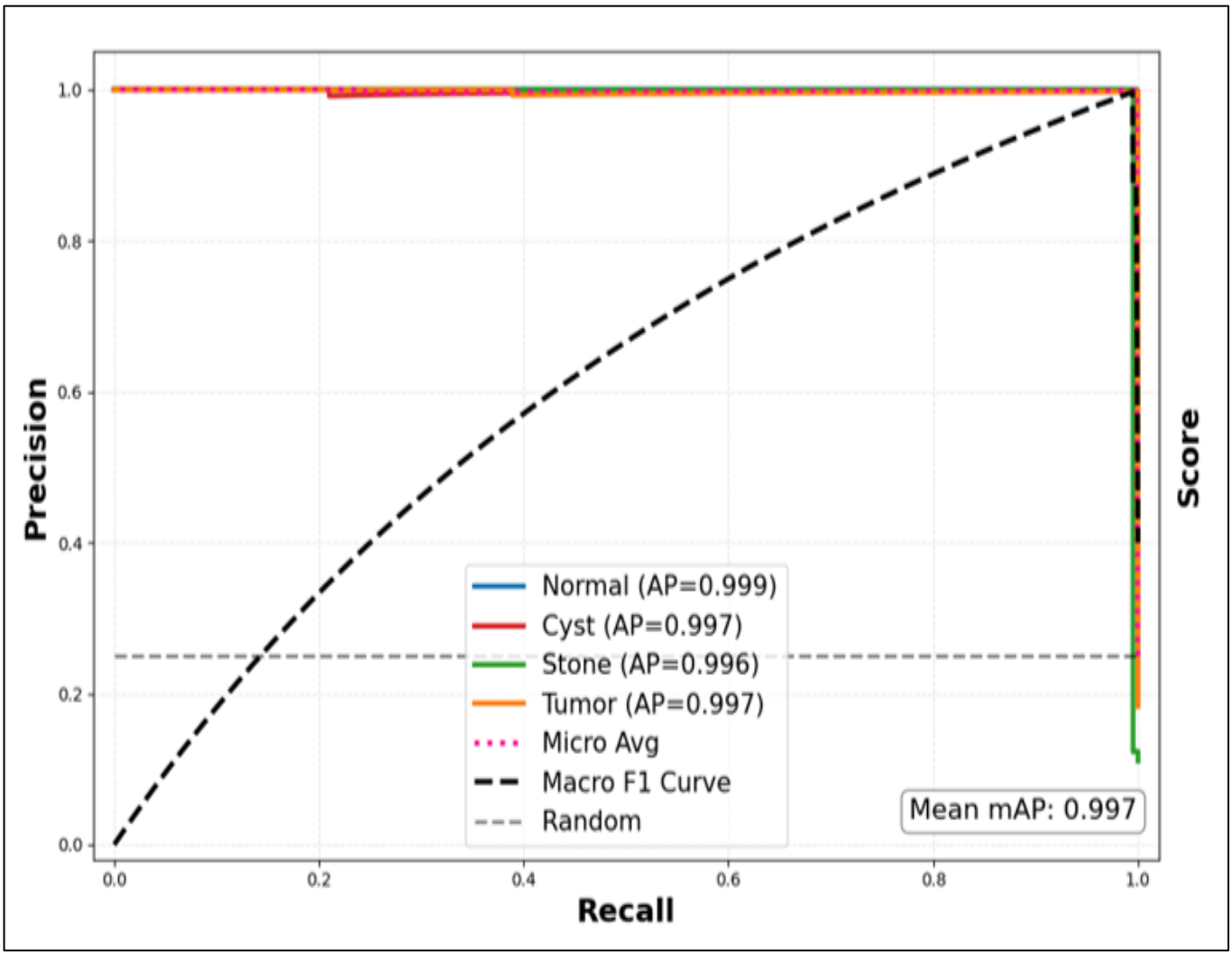
To comprehensively evaluate classification performance, as shown in Table 1, standard metrics such as accuracy, precision, recall, and F1-score were computed for each kidney disease class. On the kidney dataset, the proposed model achieved an overall classification accuracy of 99.84%. Furthermore, to its strong overall performance, the model demonstrates balanced results across all four classes, indicating a consistent and reliable distinction between Normal, Cyst, Stone, and Tumor samples.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

Table 1 Performance Matrices for Kidney Disease Classification

| Class | mAP (%) | Precision (%) | Recall (%) | F1-Score (%) | Accuracy(%) |
|---|---|---|---|---|---|
| Normal | 99.94 | 1.0000 | 99.74 | 99.87 | 99.74 |
| Cyst | 99.72 | 99.82 | 100.00 | 99.91 | 99.99 |
| Stone | 99.58 | 1.0000 | 99.52% | 99.76 | 99.52 |
| Tumor | 99.73 | 99.42 | 1.0000 | 99.71 | 99.99 |
| Macro Avg | 99.74 | 99.81 | 99.81 | 99.81 | 99.84 |

The precision and recall scores show that the classification performance is still balanced, and there are no major class imbalance problems after using the oversampling strategy. The mAP for all classes exceeded 99%, indicating robust performance across all kidney disease categories normal: 99.94%, cyst: 99.72%, stone: 99.58%, tumor: 99.73%, confirming that the model performs equally well across all classes without bias toward any pathology. The weighted mAP of 99.74% demonstrates strong performance when accounting for the test set's natural class distribution. A detailed visual analysis of the model's performance is presented in Fig. 7.




(a)

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

(b)

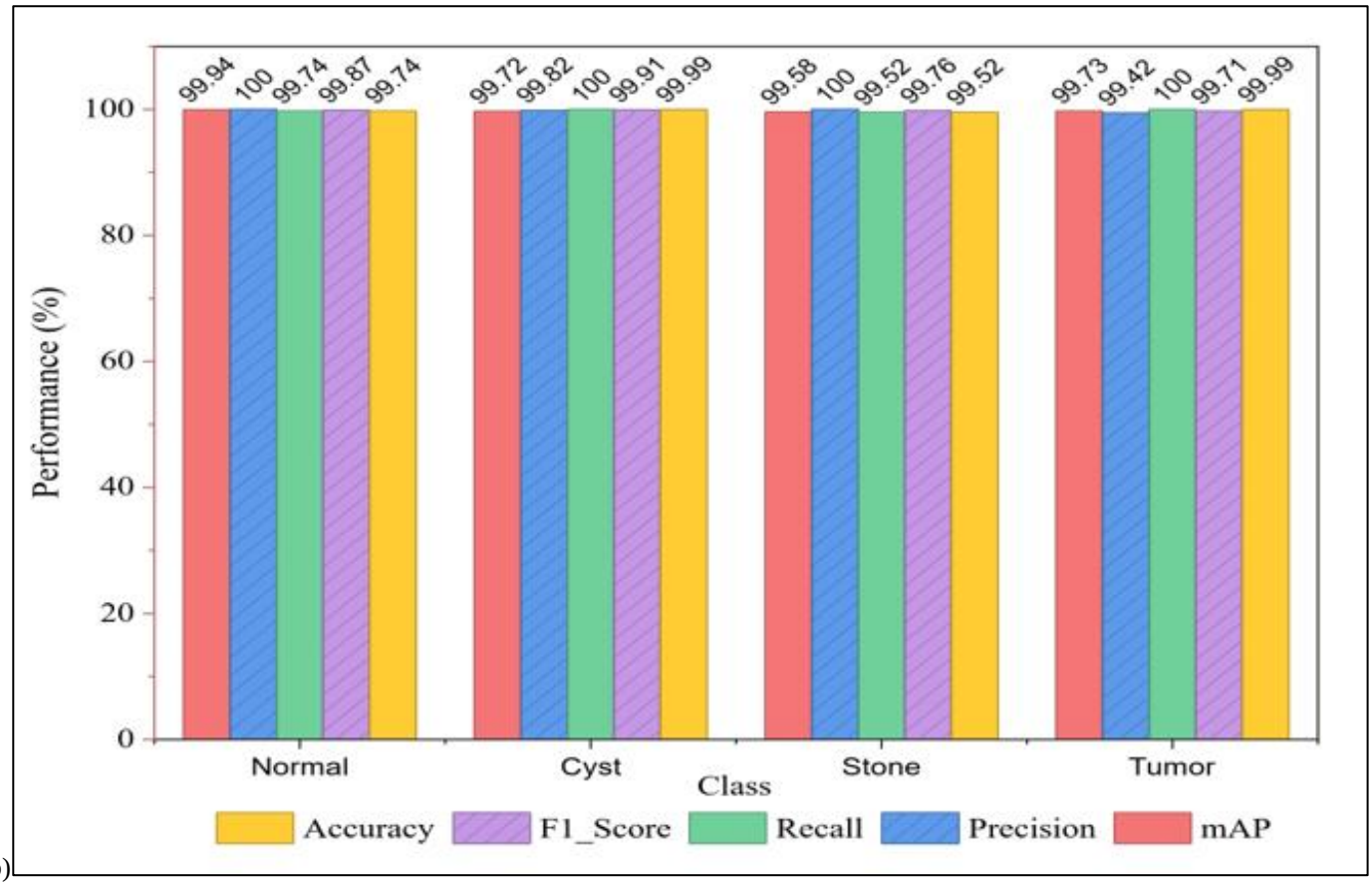



Fig. 7 Depicts the Performance Analysis of HybridSwinHyConMamba. (a) Precision-Recall Curves with Per-Class AP Scores and Mean mAP. (b) Per-Class Performance Analysis of Precision, Recall, F1-Score, and Accuracy.

The Precision-Recall curves in Fig. 7(a) demonstrate the classification performance for each pathological class, with corresponding average precision values. The mean mAP of 99.77% is highlighted in the bottom right corner, confirming strong overall performance. Fig. 7(b) provides detailed per-class metrics showing Precision, Recall, F1-Score, and Accuracy for all four classes.

The confusion matrix in Fig. 8(a) and the ROC-AUC curve in Fig. 8(b) show the performance of the Swin-HyConMamba model across different kidney disease classes. An extremely low rate of misclassification and good class-wise prediction accuracy are indicated by the confusion matrix's minimal off-diagonal values. Furthermore, with consistently high AUC values across all classes, confirming the model's strong ability to distinguish between normal, cyst, stone, and tumor categories.

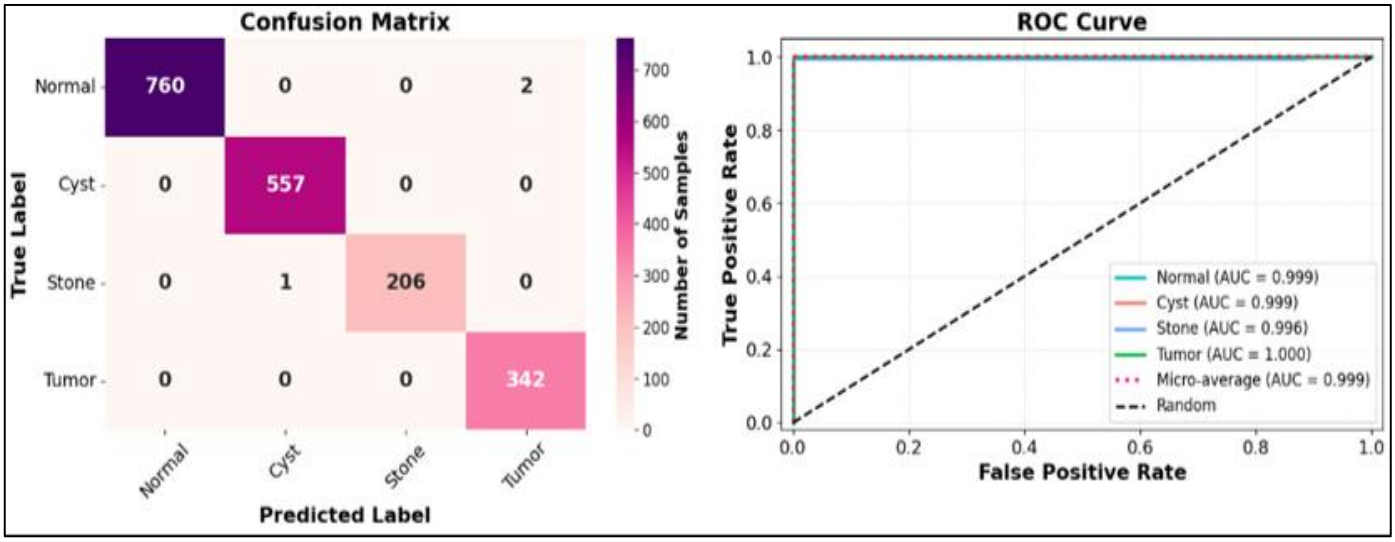



Fig 8 Performance Evaluation of Swin-HyConMamba. (a) Confusion Matrix Displaying Classification Results Across Four Classes (Normal, Cyst, Stone, Tumor) with Sample Counts. (b) ROC Curves with Per-Class AUC Values

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

Analysis of the confusion matrix demonstrates that the majority of samples are correctly classified, with only three misclassifications. For instance, two normal cases were misclassified as tumors, and one stone case was misclassified as cysts. These errors are caused by visual similarity and overlapping intensity patterns between certain kidney conditions. Despite this, the very low number of misclassifications indicates that the model effectively captures the discriminant features for reliable class differentiation.

➢ *Benchmarking Against State-of-the-Art Models*

In Table 2, we have presented the comparative evaluation of our proposed Swin-HyConMamba fusion model against several recent state-of-the-art methods, Although the majority of approaches are evaluated on the Kaggle multiclass kidney dataset, certain models have also been examined on other datasets, as described in the literature. Among the existing techniques, Islam et al.[14] utilized SwinTransformer, VGG16, and CCT to achieve 99.30% accuracy. Huang et al.[17] employed a CSTCM-SwinTransformer model and claimed 97.60% accuracy, whereas Iqbal et al.[18] applied the RDAN model and achieved 96.94% accuracy. Similarly, Eliazer et al.[21] presented TANNF-IDRCC and achieved an accuracy of 98.26 %. Other approaches, such as MSKD-Net by Shearen et al.[15] and CNN-Mamba-WOA by Pan et al.[22], have accuracy rates of 93.02% and 90.21%, respectively.

Table 2 Comparative Analysis of Our Proposed Swin-HyConMamba with the State-of-the-Art Technique on the Kaggle Multiclass Dataset[14].

| References | Dataset | Methods | Accuracy (%) | AUC / Precision / Recall / F1 (%) |
|---|---|---|---|---|
| Islam et al.[14] | Bangladeshi Hospitals (PACS) | SwinTransformer, VGG16, CCT | 99.30 | 99.15 / 99.15 / 99.15 / 99.98 |
| Shearen et al.[15] | Kaggle (PACS) | MSKD-Net | 93.02% | 93.08 / 98.03 / 98.08 / 98.13 |
| Huang et al.[17] | Kaggle (PACS) | CSTCM-SwinTransformer | 97.60% | 97.20 / 95.30 / 93.60 / 96.20 |
| Iqbal et al.[18] | Kaggle (PACS) | RDAN | 96.94% | 99.35 / 98.81 / 98.73 / 98.88 |
| Eliazer et al.[21] | RCCGNet | TANNF-IDRCC | 98.26% | 98.11 / 95.69 / 95.70 / 95.65 |
| Pan et al.[22] | Dialysis Database | CNN-Mamba-WOA | 90.21% | 82.30 / 73.20 / 89.40 / 88.10 |
| Ours | Kaggle (PACS) | Dual-branch SwinHyConMamba | 99.84 | 99.90 / 99.81 / 99.81 / 99.81 |

In comparison, the proposed model achieves the highest classification accuracy of 99.84%, outperforming all competing approaches. Beyond accuracy, it also leads across the remaining evaluation metrics 99.90% AUC, and 99.81% precision, recall, and F1-score, reflecting highly consistent and reliable classification across all kidney disease classes. The improved performance comes down to how well the model integrates global feature extraction from the Swin Transformer with the local context modeling of the HyConMamba architecture, connected through a cross-attention fusion mechanism. This design lets the model capture both fine-grained pathological features and broader anatomical context at the same time. Overall, the proposed Dual-branch Swin-HyConMamba model not only outperforms existing CNN- and transformer-based approaches but also sets a strong benchmark for kidney disease classification on the Kaggle dataset.

➢ *Ablation Studies*

To investigate the individual contributions of the Swin Transformer and HyConMamba backbones, as well as the contribution of the proposed feature fusion modules, we performed a thorough ablation analysis on the kaggle kidney CT dataset.

- *Comparison of Fusion Strategies and Individual Modules:*

As demonstrated in Table 3, we carried out an ablation study in the first experiment to assess the individual performances of the Swin Transformer and HyConMamba backbones as well as several architectural configurations (simple concatenation base fusion, multiscale, and feature refinement).

Table 3 Ablation Analysis for Swin-HyconMamba Model: Performance Comparison of Individual Backbones and Architectural Components

| Model Configuration | Accuracy (%) | Precision (%) | Recall (%) | F1 (Macro) (%) |
|---|---|---|---|---|
| Swin Transformer (baseline) | 95.13 | 93.47 | 94.08 | 93.73 |
| HyConMamba | 91.38 | 88.54 | 91.04 | 89.56 |
| Hybrid w/o Cross-Attention | 9599 | 9515 | 9508 | 9510 |
| Hybrid w/o Feature Refinement | 9872 | 9877 | 9817 | 9846 |
| Hybrid w/o Multiscale | 9877 | 9837 | 9848 | 9842 |
| Full Model (Proposed) | 99.84 | 99.81 | 99.81 | 99.81 |

The Swin Transformer backbone alone reached 95.13% accuracy, with precision, recall, and F1-score of 93.47%, 94.08%, and 93.73%. In contrast the HyConMamba encoder alone performed lower, at 91.38% accuracy and 88.54%, 91.04%, and 89.56% for the same metrics. Combining both backbones without cross-attention pushed accuracy to

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

95.99%, showing that feature integration alone offers a measurable gain. Adding feature refinement brought this further to 98.72%, and removing the multiscale component achieved 98.77% a marginal drop that still points to its role in capturing multi-resolution features. The full model, with cross-attention fusion and multiscale feature refinement included, achieved 99.84% accuracy and 99.81% precision, recall, and F1-score.

These results highlight the contribution of each component. Cross-attention fusion allows the model to combine global context from the Swin Transformer with local representations from HyConMamba, while feature refinement and multiscale modules help focus on diagnostically relevant regions. Each addition produces a clear improvement, and the complete architecture outperforms all ablated variants on kidney disease classification.

- *Impact of Dropout Regularization:*

In the second experiment, we evaluated the impact of dropout regularization on model performance and generalization capabilities. We examined model performance with various dropout configurations to find the best regularization approach, as shown in Fig. 9.

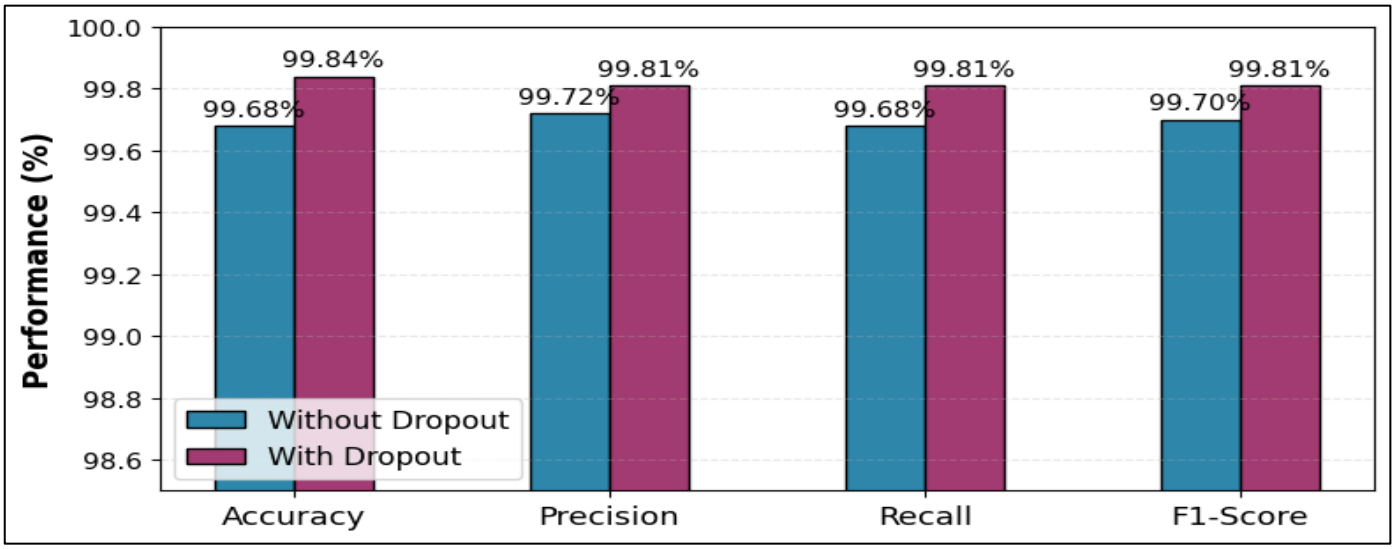



Fig 9 Dropout Performance Comparison of the Swin-HyConMamba Model with and Without Dropout Layers Applied After the Swin Transformer and HyConMamba

Adding dropout at a rate of 0.3 after both backbones and 0.4 in the classifier head produced the best results, with validation accuracy improved from 99.74% to 98.89% while also reducing the training-validation gap, and testing accuracy improved from 99.68% to 99.84%. The precision, recall, and F1-macro also improved with this configuration, increasing from 99.70%, 99.72%, and 99.68% to 99.81%, 99.81%, and 99.81%, respectively. The high dropout rate in the classifier head makes sense given its fully connected structure, which is more susceptible to overfitting. Maintaining a moderate rate in the backbone layers preserves rich feature representation while strong regularization at the classification stage prevents the model from remembering training samples. This separation of rates also reduced the training-validation gap, reflecting better generalization.

Overall, the results confirm that dropout placement and rate matter so much that if dropouts are used perfectly evenly, it will likely disorganize the classification while unnecessarily constraining the feature extractor.

➢ *Visual Explanations and Interpretability*

To verify that the model's predictions are grounded in clinically relevant image regions, two explainability techniques were applied: gradient-based saliency maps and Local Interpretable Model-Agnostic Explanations (LIME).

- *Saliency Map Analysis*

In this study, gradient-based saliency maps were used to compare the attentional regions of three models: Swin Transformer, HyConMamba, and the proposed Hybrid Swin-HyConMamba for kidney disease classification. The saliency maps highlight the image regions that most influence the prediction, where higher values correspond to features that the model weighs more heavily during classification. For each test image, the saliency map was computed by backpropagating the gradient of the predicted class score with respect to the input pixels. The absolute gradient values across the color channels were averaged and normalized to [0, 1] to produce a single-channel representation.

✓ *Swin-Transformer:*

The saliency map of the Swin model shows a scattered attention pattern on the image, consistent with its window-based self-attention mechanism. This mechanism captures the global structural background by simultaneously focusing on multiple regions. This broad coverage helps in understanding overall anatomical structures but may reduce attention to local lesion areas. The Swin saliency map is shown in Figure 10.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

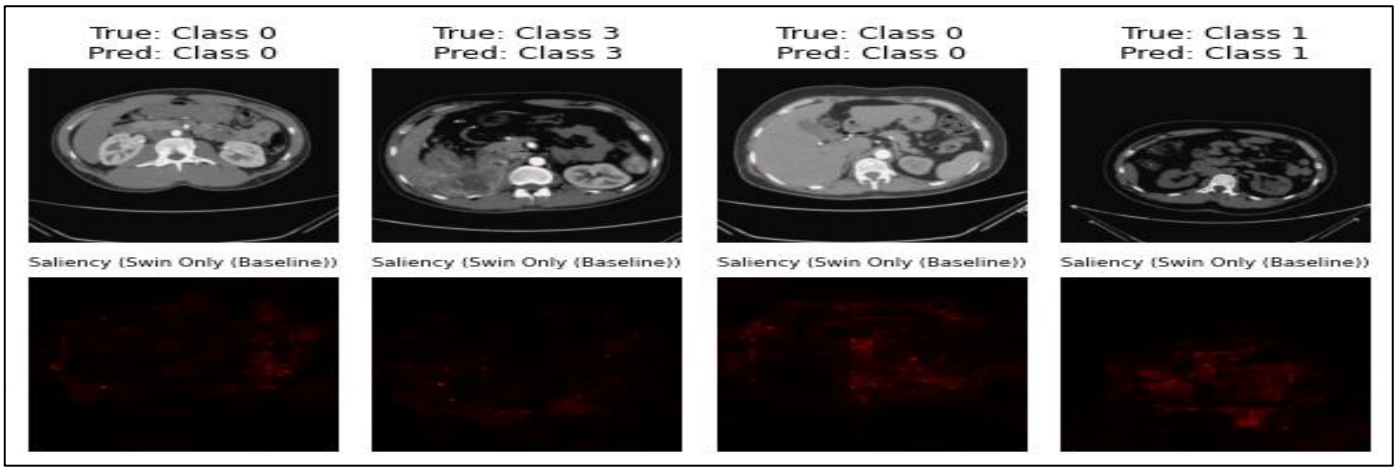



Fig 10 Gradient-based saliency map visualisation of the Swin Transformer model for classification of kidney diseases

✓ *HyConMamba:*

HyConMamba saliency maps show large, more diffuse regions of interest, demonstrating the ability to extract local state space features. This captures fine structural details throughout the image, although spatial accuracy may vary for small pathological areas. The HyConMamba saliency visualization is shown in Figure 11.

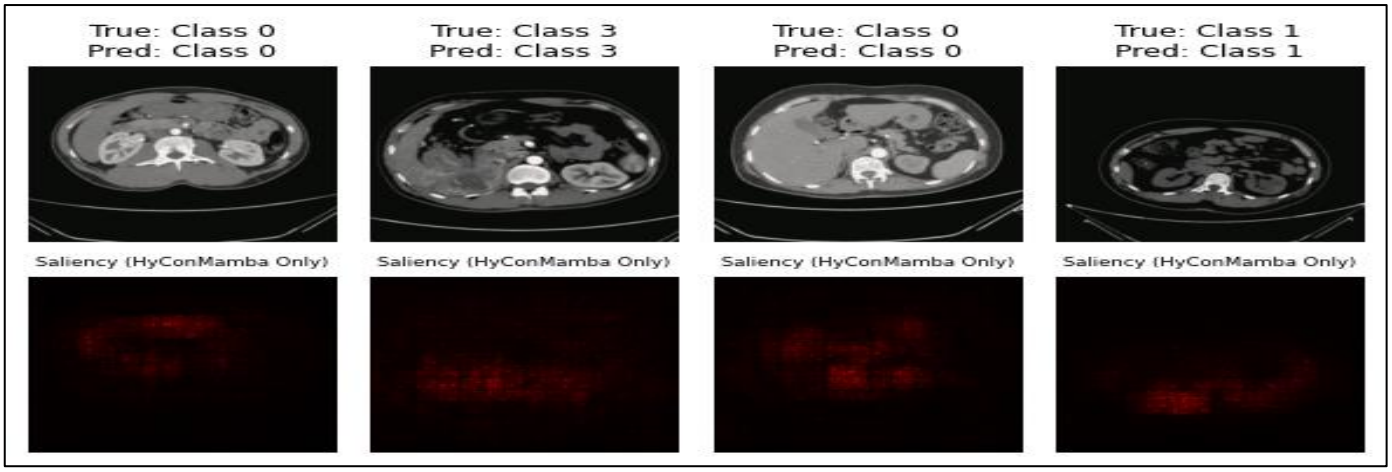



Fig 11 Gradient-Based Saliency Map Visualization of the HyConMamba Model for Kidney Disease Classification, Illustrating Broader Contextual Attention Over Kidney Structures.

✓ *Swin-HyConMamba:*

The proposed model generates a more focused and clinically consistent attention pattern. The cross-attention fusion mechanism combines global contextual representations from the Swin branch with hierarchical local features from HyConMamba, naturally suppressing background regions; simultaneously, feature refinement and multi-scale context aggregation modules capture lesions at different scales. As shown in Figure 12, these modules collectively generate a saliency map, focusing on key lesion regions such as cysts, calcified stones, tumor masses, and normal parenchyma.

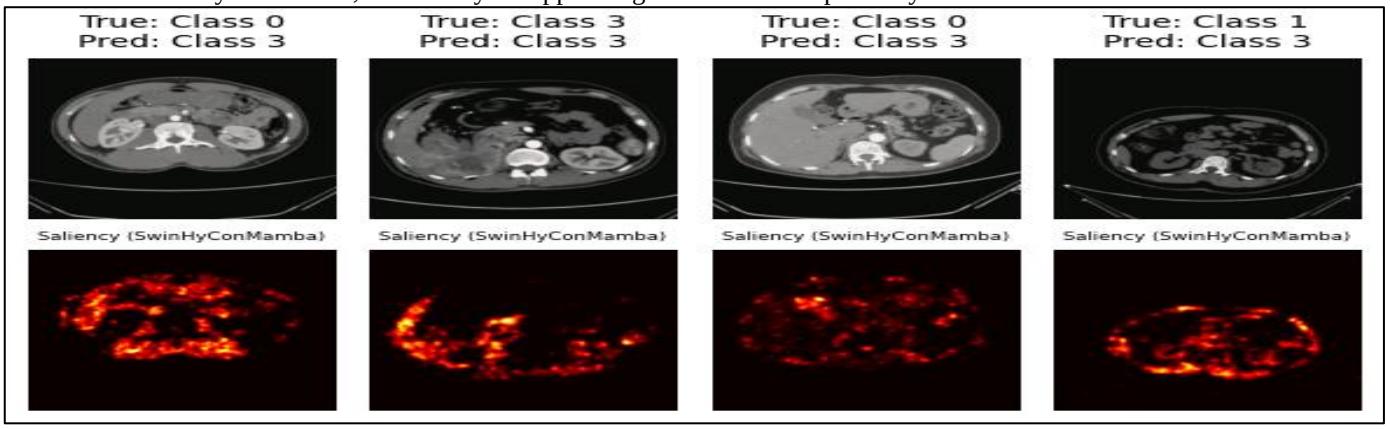



Fig 12 Gradient-Based Saliency Map Visualization of the Proposed Swin-HyConMamba Model.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

- *LIME Explanation*

In addition to saliency maps, LIME (Local Interpretable Model-Agnostic Explanations) was used to further analyze the decision-making process of the Hybrid Swin-HyConMamba model. LIME provides local explanations for individual predictions by perturbing the input image and observing changes in the model's output probabilities, enabling identification of the most influential image regions. For this analysis, the number of superpixels was set to 1, 2, or 3 to focus on clinically relevant structures such as kidney parenchyma, cystic regions, stones, and tumor masses. The number of perturbed samples was set to 1,000 to ensure stable explanations while maintaining reasonable computational cost. LIME explanations show that the model consistently focuses on pathologically meaningful regions, including cystic cores in Cyst cases, calcified deposits in Stone cases, and tumor masses with surrounding tissue context in Tumor cases, while Normal cases exhibit distributed attention across kidney parenchyma. To illustrate the model's decision-making behavior, three representative samples from each class (Normal, Cyst, Stone, Tumor) were visualized using LIME, as shown in Fig. 13. In some cases, additional superpixels appear in nearby kidney tissue when three superpixels are used, capturing both lesion regions and surrounding anatomical context.

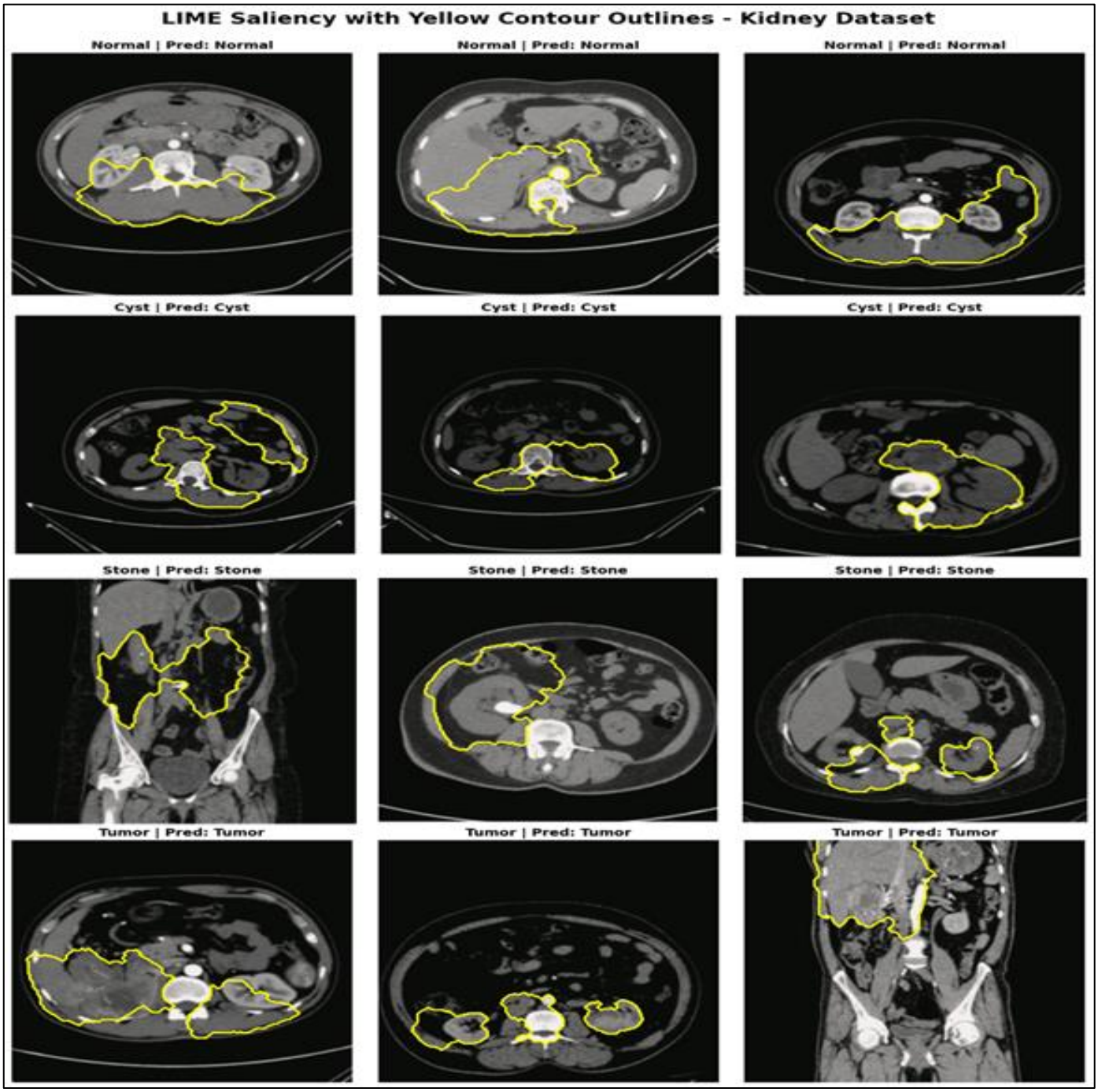



Fig 13 LIME Explanation Visualization for the Swin-HyConMamba Model Prediction: Highlighting Critical Superpixels in the CT Images from Each Kidney Disease.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

Overall, the LIME analysis confirms that the model relies on anatomically relevant features rather than background artifacts, supporting the interpretability and robustness of the proposed framework. Overall, the experimental evaluations and interpretability analyses confirm the effectiveness and robustness of the proposed Hybrid Swin-HyConMamba framework for automated kidney disease classification.

## V. CONCLUSION

In this study, a novel dual-branch deep learning architecture, termed Swin–HyConMamba, was proposed for automated kidney disease classification from medical images. The framework integrates the hierarchical feature extraction capability of the Swin Transformer with the long-range dependency modeling ability of the HyConMamba encoder. Through a cross-attention fusion mechanism, the model effectively combines complementary information from both branches, enabling the extraction of both global structural context and fine-grained pathological features. This design allows the proposed framework to capture complex patterns associated with different kidney disease categories, including normal tissue, cysts, stones, and Tumors. In addition to improving feature representation, the proposed approach incorporates explainable artificial intelligence techniques to enhance the interpretability of the model. Visualization methods such as gradient-based saliency maps help identify diagnostically relevant regions in the images, allowing clinicians to better understand the model's decision-making process. This interpretability is particularly important for increasing trust in artificial intelligence systems in clinical environments. Despite these advantages, some challenges remain when dealing with highly complex pathological cases where multiple abnormalities coexist. Future work will focus on evaluating the model on larger and more diverse datasets to further assess its clinical applicability, as well as exploring additional regularization strategies to improve robustness across varying imaging conditions.

## REFERENCES

[1]. KOVESDY C P. Epidemiology of chronic kidney disease: an update 2022 [J]. Kidney international supplements, 2022, 12(1): 7-11.
[2]. KUMAR P, BHATIA M. Role of CT in the pre-and postoperative assessment of conotruncal anomalies [J]. Radiology: Cardiothoracic Imaging, 2022, 4(3): e210089.
[3]. FRIEBE M. AI in radiology and interventions: a structured narrative review of workflow automation, accuracy, and efficiency gains of today and what's coming [J]. International Journal of Computer Assisted Radiology and Surgery, 2025: 1-10.
[4]. KAUR R, JUNEJA M, MANDAL A K. Computer-aided diagnosis of renal lesions in CT images: a comprehensive survey and future prospects [J]. Computers & Electrical Engineering, 2019, 77: 423-34.
[5]. ZHANG M, YE Z, YUAN E, et al. Imaging-based deep learning in kidney diseases: recent progress and future prospects [J]. Insights into imaging, 2024, 15(1): 50.
[6]. THOMAS N R, ANITHA J, POPIRLAN C, et al. Next-Generation Deep Learning Approaches for Kidney Tumor Image Analysis: Challenges, Clinical Applications, and Future Perspectives [J]. Computers, Materials, & Continua, 2025, 85(3): 4407.
[7]. LIANG Y. Application of multi-scale dynamic enhancement based on deep neural network and CT urinary tract secretory phase image fusion in the diagnosis of urinary system diseases [J]. BMC Medical Imaging, 2025, 25(1): 1-21.
[8]. ZHANG K, WANG W, LV Z, et al. Computer vision detection of foreign objects in coal processing using attention CNN [J]. Engineering Applications of Artificial Intelligence, 2021, 102: 104242.
[9]. YIN Y, TANG Z, WENG H. Application of visual transformer in renal image analysis [J]. BioMedical Engineering OnLine, 2024, 23(1): 27.
[10]. SINGH D P, BANERJEE T, DURAI C A D, et al. A Comprehensive Study of Various Hybrid Deep Learning Models for Automated and Explainable Pneumonia Detection in the Pulmonary Alveolar Region: Current Insights and Future Directions [J]. Archives of Computational Methods in Engineering, 2025: 1-33.
[11]. ASIRI A A, SHAF A, ALI T, et al. Advancing brain tumor detection: harnessing the Swin Transformer's power for accurate classification and performance analysis [J]. PeerJ Computer Science, 2024, 10: e1867.
[12]. WANG Y, MEI S, MA M, et al. HTACPE: A hybrid transformer with adaptive content and position embedding for sample learning efficiency of hyperspectral tracker [J]. IEEE Transactions on Multimedia, 2025, 27: 2384-98.
[13]. WU X, CAO Z-H, HUANG T-Z, et al. Fully-connected transformer for multi-source image fusion [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(3): 2071-88.
[14]. ISLAM M N, HASAN M, HOSSAIN M K, et al. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography [J]. Scientific Reports, 2022, 12(1): 11440.
[15]. SHAREN H, NARENDRA M, ANBARASI L J. MSKd_Net: Multi-Head Attention-Based Swin Transformer for Kidney Diseases Classification [J]. IEEE Access, 2024, 12: 181975-86.
[16]. MARTÍN Ó A, SÁNCHEZ J. Evaluation of Vision Transformers for Multi-Organ Tumor Classification Using MRI and CT Imaging [J]. Electronics, 2025, 14(15): 2976.
[17]. HUANG H, HUANG Y, DU X, et al. Contrastive Swin Transformer-Based Classification Model for Internet of Medical Things-Driven Kidney Stone Diagnosis [J]. IEEE Internet of Things Journal, 2025.
[18]. IQBAL S, QURESHI A N, ALHUSSEIN M, et al. A novel reciprocal domain adaptation neural network for enhanced diagnosis of chronic kidney disease [J]. Expert Systems, 2025, 42(2): e13825.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26mar1555

[19]. REHMAN A, MAHMOOD T, SABA T. Robust kidney carcinoma prognosis and characterization using Swin-ViT and DeepLabV3+ with multi-model transfer learning [J]. Applied Soft Computing, 2025, 170: 112518.
[20]. CONZE P-H, ANDRADE-MIRANDA G, LE MEUR Y, et al. Dual-task kidney MR segmentation with transformers in autosomal-dominant polycystic kidney disease [J]. Computerized Medical Imaging and Graphics, 2024, 113: 102349.
[21]. ELIAZER M, KUMAR G M, AMARAN S, et al. Advanced transformer with attention-based neural network framework for precise renal cell carcinoma detection using histological kidney images [J]. Scientific Reports, 2025, 15(1): 35345.
[22]. PAN W, LIU Y. CNN–Mamba–WOA: an efficient and explainable state-space fusion framework with multi-objective optimization for large-scale hemodialysis time-series prediction [J]. The Journal of Supercomputing, 2026, 82(1): 23.
[23]. LU F, XU J, SUN Q, et al. An Efficient Vision Mamba–Transformer Hybrid Architecture for Abdominal Multi-Organ Image Segmentation [J]. Sensors, 2025, 25(21): 6785.
[24]. RMR S S, MB S, R D, et al. A phase-aware Cross-Scale U-MAMba with uncertainty-aware segmentation and Switch Atrous Bifovea EfficientNetB7 classification of kidney lesion subtype [J]. Lasers in Medical Science, 2025, 40(1): 398.
[25]. QAMAR S, FAZIL M, AHMAD P, et al. UNet with self-adaptive Mamba-like attention and causal-resonance learning for medical image segmentation [J]. Scientific Reports, 2025.
[26]. ZHANG X, WANG X, NIU T. CT Image segmentation using frequency domain feature-assisted selective long memory state space model [J]. Sensing and Imaging, 2025, 26(1): 74.
[27]. SU C, LUO X, LI S, et al. VMKLA-UNet: vision Mamba with KAN linear attention U-Net [J]. Scientific Reports, 2025, 15(1): 13258.
[28]. SUN K, ZHOU J, WANG M, et al. S 2 Mamba: An Efficient Mamba Accelerator With Word-Importance SSM Sparsity [J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2026.
[29]. ZHOU R-G, HU W, FAN P, et al. Quantum realization of the bilinear interpolation method for NEQR [J]. Scientific Reports, 2017, 7(1): 2511.
[30]. DONG Y, YUE X, XU Z, et al. Correlation and Foreground Attention to Improve Object Detection; proceedings of the 2023 IEEE International Conference on Image Processing (ICIP), F, 2023 [C]. IEEE.
[31]. LIU Y, LI H, HU C, et al. Learning to aggregate multi-scale context for instance segmentation in remote sensing images [J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 36(1): 595-609.