Explainable Knowledge Distillation via Capsule Vision Transformers for Automated Kidney Disease Categorization

Sachin Dattatraya Shingade[1], Midhun Chakkaravarthy[2], Dimitrios A. Karras[3] , Sachin S[4], Komal M Masal[5]

[1,2]LUCM PetaLing Jaya Malaysia; [3] LUCM, NKUA Athens Greece and EUT Albania; [4] PICT Pune, IITP India, [5] PICT , DOT SPPU Pune India

[1]pdf.sachin@lincoln.edu.my ; [2] midhun@lincoln.edu.my; [3] dimitrios.karras@gmail.com; [4]sachin_pa2503mth305@iitp.ac.in; [5]kmmasal@pict.edu

Abstract: Kidney disease, characterized by the progressive decline in the renal system's ability to filter metabolic waste and excess fluids, poses a significant global health risk. When these physiological impairments persist beyond a three-month threshold, the condition is classified as Chronic Kidney Disease (CKD). Current diagnostic frameworks often struggle with high computational overhead, suboptimal precision, and a lack of architectural efficiency for deployment on lightweight devices. To resolve these challenges, this study introduces CapViT-DKD , an advanced kidney detection and classification framework utilizing a Capsule Vision Transformer integrated with Self-Supervised Diverse Knowledge Distillation. The methodology utilizes the CT-kidney dataset, initiated by a robust preprocessing pipeline comprising image resizing, normalization, and strategic data augmentation. The core architecture employs a teacher-student paradigm: a high-capacity Perceptive Capsule Transformer Network (PCapTN) serves as the teacher, transferring complex feature representations to a Lightweight Capsule Transformer Network (LCapTN) student model. This diverse knowledge distillation (DKD) approach significantly boosts the student model's performance while maintaining a small computational footprint. To address the "black box" nature of deep learning, we incorporate Principal Component of Gradient-Class Activation Mapping (PCG-CAM), which provides visual explanations by highlighting the specific anatomical regions influencing the diagnostic output. Empirical results demonstrate that the proposed system achieves superior performance metrics, including an accuracy of 99.75%, precision of 99.50%, Recall of 99.15 % and an F1-score of 99.10%, validating its efficacy for clinical decision support

Keywords: CT-Kidney Dataset; CapViT-DKD (Capsule Vision Transformer with Diverse Knowledge Distillation); PCG-CAM (Explainable AI); Teacher-Student Learning Architecture; Medical Image Classification; Self-Supervised Learning.

Introduction

The Renal pathologies represent a critical global health crisis, with epidemiological data from a recent cross-sectional survey in China indicating a prevalence of approximately 8.7%, affecting nearly 82 million people [1]. A fuzzy-enhanced detection strategy with Machine Learning Operations (MLOps) to facilitate a fusion of twin transferable networks and weighted ensemble classifiers, optimizes feature learning and classification robustness through an automated, scalable pipeline designed for high-precision renal tumor identification [2].The diagnostic landscape is further complicated by a wide array of abnormalities, including renal calculi, cysts, and both benign and malignant tumors, alongside systemic functional impairments such as acute kidney injury (AKI) and chronic kidney disease (CKD) [3]. Traditional diagnostic pathways specifically urinalysis, blood panels, and invasive renal punctures require specialized clinical

environments and intensive manual oversight. These conventional methods are often hindered by their time-consuming nature and labor requirements, creating significant barriers for patients who need accessible and efficient monitoring [4]. Existing literature has comprehensively reviewed the role of state-of-the-art Artificial Intelligence, specifically deep learning and machine learning, in optimizing the precise detection and categorization of renal tumors within CT imagery [5].

By integrating ML and DL methodologies, the healthcare sector has achieved significant improvements in diagnostic precision, reduced latency in clinical workflows, and overall operational efficiency [6]. The emergence of Explainable Artificial Intelligence (XAI) has introduced a transformative layer of transparency to clinical decision-making, providing the necessary interpretability to demystify the "black box" nature of algorithmic predictions [7]. In the current landscape of AI and machine learning, researchers are increasingly leveraging these diagnostic frameworks to identify chronic kidney disease, moving beyond traditional clinical observation toward data-driven modeling [8]. By translating histopathological image data into quantifiable metrics such as cellular morphology, density, and spatial distribution AI-driven tools serve as vital diagnostic adjuncts that mitigate the global shortage of pathologists while simultaneously enhancing the objectivity and precision of tumor classification [9].

This research aims to bridge the gap in renal diagnostic technologies by introducing a high-fidelity classification framework that overcomes the precision deficits, interpretive opacity, and computational inefficiencies inherent in current methodologies. Central to this work is the development of the CapViT-DKD (Capsule Vision Transformer with Self-Supervised Diverse Knowledge Distillation) architecture, which leverages a teacher-student paradigm to reduce model complexity while maintaining high performance.

Related work

Advancements in deep learning have significantly transformed the landscape of automated kidney abnormality detection within CT imaging, offering a viable solution to the global shortage of specialized nephrologists and the increasing demand for computer-aided diagnosis (CAD) in clinical radiology. Current research predominantly focuses on the application of state-of-the-art convolutional neural networks (CNNs) and transfer learning frameworks to categorize renal conditions such as cysts, tumors, and calculi. These studies explore various optimization vectors, including the development of novel architectures, region-of-interest (ROI) localization, sophisticated image preprocessing, and model interpretability, all aimed at enhancing diagnostic precision and clinical integration.

The research introduced by Midhun et al. [10] utilizes a Guided Triple Gaussian Filter for noise reduction and a Pyramid Residual Autoencoder for multi-scale feature extraction. The architecture integrates a specialized Squeeze-and-Excitation classifier with depthwise separability, achieving a high-precision diagnostic accuracy of 98.81% and an F1-Score of 98.81. To address the diagnostic challenges of renal cell carcinoma, Midhun et al. [11] introduces an efficient Multi-Scale Hybrid Transformer Network optimized for high-precision malignancy identification. The architecture utilizes a specialized transformer backbone and a lightweight channel-wise attention module to capture long-range spatial dependencies while suppressing anatomical noise.

Hossain et al. [12] optimized kidney abnormality detection by utilizing segmentation-based ROI extraction and bicubic interpolation for pixel reduction, significantly lowering computational overhead. Their EfficientNet B7 and Random Forest ensemble achieved a peak accuracy of 99.75%, demonstrating high clinical potential for robust and efficient diagnostic integration. Almuayqil et al. [13] developed Kidney-Net, a specialized CNN architecture featuring eight convolutional layers and Grad-CAM integration to ensure diagnostic transparency. This interpretable framework outperformed standard models in identifying stones, cysts, and tumors, providing a stable solution for early-stage chronic kidney disease detection.

Pimpalkar et al. [14] utilized fine-tuned transfer learning architectures, including VGG16 and ResNet50, combined with hyperparameter optimization to classify renal tumors with a remarkable 99.96% accuracy. Their approach underscores the efficacy of advanced image processing and neural network refinement in establishing high-performance benchmarks for automated CT classification. Grover et al. [15] engineered a CNN-based multiclass diagnostic tool trained on a large-scale hospital dataset to categorize renal scans into four distinct clinical classes. By employing strategic data augmentation and max-pooling layers, the model achieved exceptional F1-scores and precision, validating its reliability as an automated decision-support system for radiologists.

Motivation

The global escalation of renal malignancies necessitates high-precision diagnostic tools, as the asymptomatic nature of early-stage tumors often leads to late-stage detection and poor survival rates. This research is driven by the requirement for automated CT analysis frameworks that can identify subtle renal abnormalities with clinical accuracy. By facilitating earlier intervention and targeted therapeutic planning, these intelligent systems serve as a critical bridge between radiological imaging and improved patient outcomes and this is the motivation behind research work.

Proposed Methodology

The proposed diagnostic framework introduces an explainable deep learning pipeline for automated kidney disease categorization, utilizing a comprehensive dataset of 12,446 CT scans classified into normal, cyst, tumor, and stone categories. The methodology begins with extensive image preprocessing, including standardization to a fixed resolution, pixel intensity normalization to stabilize learning, and geometric data augmentations like rotation and random cropping to mitigate overfitting and class imbalance. At the core of the system is the Capsule Vision Transformer (CapViT) architecture, which employs a teacher-student paradigm where a high-capacity Perceptive Capsule Transformer Network (PCapTN) transfers hierarchical spatial knowledge to a Lightweight Capsule Transformer Network (LCapTN). This diverse knowledge distillation process allows the student model to maintain high diagnostic precision while reducing computational complexity for efficient clinical deployment. To ensure interpretability, the framework integrates Principal Component of Gradient-Class Activation Mapping (PCG-CAM), which visualizes the critical anatomical regions influencing the model's decisions. This integrated approach culminates in a multi-class prediction system that provides radiologists with both quantitative diagnostic labels and qualitative visual evidence to support targeted therapeutic strategies..

.

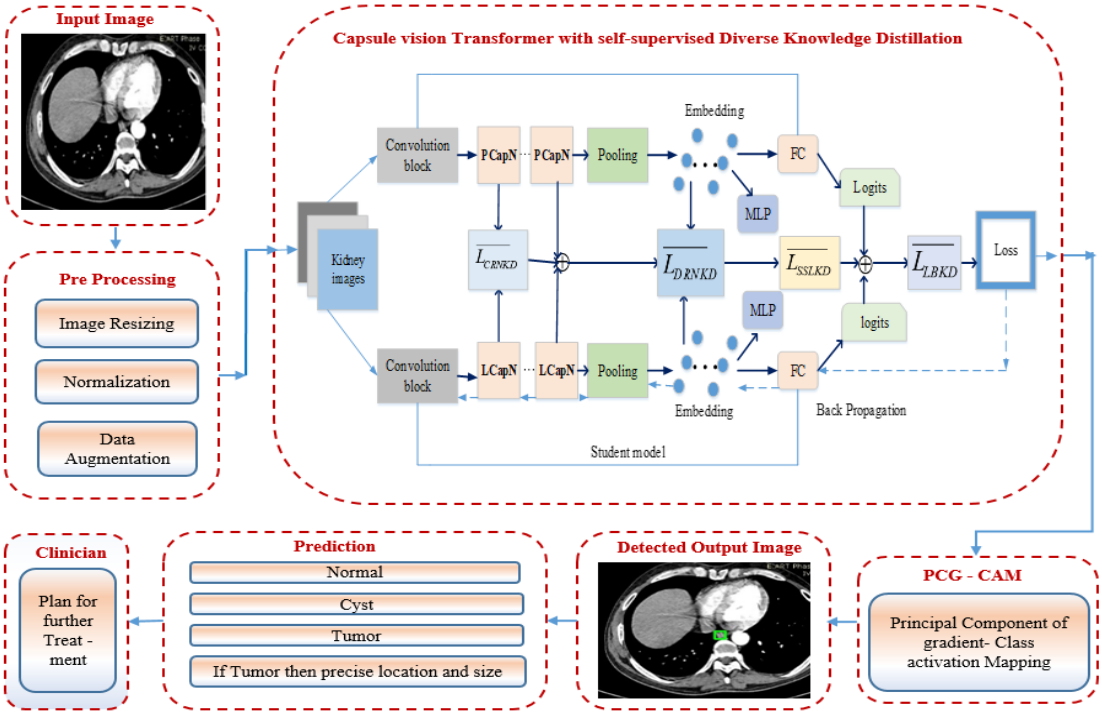

Figure 1. Architecture of the Explainable Knowledge Distillation via Capsule Vision Transformers for Automated Kidney Disease Categorization .

Data Acquisition and Image Pre-processing

The system utilizes a comprehensive CT kidney dataset comprising four clinical categories: normal, cyst, tumor, and stone. The dataset contains 12,446 distinct data points, including 3,709 cysts, 2,283 tumors, 1,377 stones, and 5,077 normal scans. Data was sourced from both coronal and axial slices of urograms, covering contrast and non-contrast studies. DICOM images were converted to a lossless JPEG format using the Sante DICOM editor, while the Philips Intellispace portal was employed for expert annotation and visualization.

The set of input kidney images is represented as: $I = \{ I_1, I_2, I_3, \dots , I_n \}$, where I_i denotes the i-th CT kidney image and n represents the total number of images in the dataset. The mathematical representation of the entire pre-processing sequence is expressed as: $\tilde{I} = A (N (R (I)))$, In this formulation, $\tilde{I}$ represents the raw input image, while the operators R , N , and A denote the sequential application of resizing, normalization, and data augmentation, respectively. To align with the architectural constraints of the implemented system, all images undergo a resizing operation to a fixed resolution of 224 x 224 pixels .Normalization is performed to stabilize the learning process by ensuring that all input features reside within a comparable numeric range . This step handles outliers and facilitates faster convergence during backpropagation by centering the pixel intensities .The normalization for each pixel is calculated using the following equation:

$$X\_norm = (X - \mu) / \sigma \qquad (1)$$

Where X_norm is the resulting standardized pixel value, X is the original input pixel intensity, mu (μ)is the mean intensity value calculated across the image dataset, sigma (σ) is the standard deviation of the image dataset. To resolve class imbalance issues and prevent the model from overfitting on limited clinical samples, a suite of geometric transformations is applied. This includes 15-degree rotations, vertical/horizontal flipping, and Random Image Cropping and Patching (RICAP). These operations artificially expand the diversity of the training set, forcing the network to learn rotation-invariant and occlusion-resistant features.

Capsule Vision Transformer (PCapN & LCapN)

The core of the methodology is the CapViT-DKD architecture, which employs a teacher-student learning paradigm. The Perceptive Capsule Transformer Network (PCapN) serves as the teacher, while the Lightweight Capsule Transformer Network (LCapN) acts as the student. Both models utilize a Local Feature Network (LFN) and a Global Relation Network (GRN) to capture intricate spatial relationships. Equation for the Squash Activation Function: is

$$v_j = (||s_j||^2 / (1 + ||s_j||^2)) \times (s_j / ||s_j||) \quad (2)$$

Where v_j is the output vector of capsule j, s_j is the total input vector to capsule j, $|| s_j ||$ is the magnitude of the input vector, ensuring the output length is squashed between 0 and 1.

To capture extensive spatial dependencies across diverse renal regions, the system processes capsule embeddings through a Capsule Vision Transformer (CapViT) module. While the capsule layers handle local hierarchies, this transformer-based block addresses global contextual relationships that may exist between distant anatomical markers in the kidney. By utilizing a self-attention mechanism, the architecture allows every individual capsule to interact with and attend to every other capsule, ensuring that the structural integrity of the organ is maintained within the model's internal representation. The relationship and rules between these internal representations are extracted by correlating data across different locations using the following attention operation:

$$Attention(Q, K, V) = Softmax ((Q \times K^T) / \sqrt{d}) \times V \quad (3)$$

Where Q (Query) is the map of query vectors used to identify relevant feature relationships, K (Key) is the set of key values that the queries are compared against to determine attention scores, V (Value) is the actual feature information that is weighted and aggregated based on the attention scores, d Denotes the embedding dimension, where √d serves as a scaling factor to prevent gradient vanishing during training. Softmax is an activation function used to normalize the scores into a probability distribution.

Diverse Knowledge Distillation (DKD)

The core of This block facilitates knowledge transfer from the teacher (PCapN) to the student (LCapN) through various knowledge types: logit-based, intra-instance, and inter-instance relational knowledge. This allows the student model to inherit the high performance of the teacher while remaining computationally efficient for lightweight implementation. Equation for the Total Distillation Loss is :

$$L\_total = \alpha \cdot L\_LBDK + \beta \cdot L\_DRKD + \gamma \cdot L\_CRKD + \delta \cdot L\_SSLKD \quad (4)$$

Where L_total is the combined multi-faceted loss function, L_LBDK is the Logit-based Dark Knowledge loss. L_DRKD is the data Relational (intra-instance) Knowledge Distillation loss, L_CRKD is the channel-wise Relational (inter-instance) Knowledge Distillation loss, L_SSLKD is the Self-Supervised Learning Knowledge Distillation loss, α(alphe), β (beta) , γ (gamma), δ (delta) are the hyperparameters used to balance the weight of each loss term.

Explainability via PCG-CAM

To ensure clinical accountability, the framework incorporates Principal Component of Gradient-Class Activation Mapping (PCG-CAM). This method visualizes the specific input regions that are most influential to the model's final prediction by extracting principal components from class-specific gradients. Equation for Gradient Centralization is:

$$G'_i = G_i - \mu(G) \quad (5)$$

Where G'_i is the centralized gradient matrix for feature map I , G_i is the original class-specific gradient matrix, G_i is the mean of the gradient matrix, used to prevent any single feature from dominating the visualization.

Output Classification and Prediction

In the final stage, the model categorizes the input CT scan into one of four classes: Normal, Cyst, Tumor, or Stone. The terminal phase of the proposed architecture functions as a comprehensive clinical decision support system. The framework integrates high-fidelity classification results with transparent visual explanation maps generated by the PCG-CAM module. This integration allows healthcare professionals to not only identify the presence of pathologies such as cysts, stones, or tumors but also to verify their precise anatomical location and spatial extent within the kidney. By aligning automated categorization with clinical interpretability, the system facilitates, by detecting subtle structural abnormalities that may be asymptomatic in early stages, it also provides localized data to assist in surgical or therapeutic strategy formulation. It Enhances the speed and accuracy of the diagnostic workflow, which is occurred due to the global shortage of specialized pathologists. The final prediction is rendered through a multiclass output layer, categorizing the input scan into one of four distinct states: Normal, Cyst, Tumor, or Stone. This framework serves as a reliable secondary opinion for radiologists, ensuring that the final clinical path is backed by both quantitative data and qualitative visual evidence.

Algorithm I: Explainable Knowledge Distillation via Capsule Vision Transformers for Automated Kidney Disease Categorization

Input : Contrast-enhanced CT kidney image set, $I = \{ I_1, I_2, I_3, \dots , I_n \}$
Output: Detected renal region with class label, $C \in \{$ Normal, Cyst, Tumor, Stone $\}$ and explainable localization map

1: Load contrast and non-contrast CT kidney images from the dataset
2: Convert DICOM images into lossless JPEG format..
3: Assign expert-validated labels for Normal, Cyst, Tumor, and Stone classes.
4: **for each** CT image $I_i \in I$ **do**
5: Resize the image to 224 × 224 pixels...
6: Normalize pixel intensity values to a standard range.
7: **end for**

8: **for each** normalized image $\tilde{I}_i$ **do**
9: Apply vertical and horizontal flipping.
10: Apply 15-degree rotation and Random Image Cropping and Patching (RICAP)
11:**end for**
12 **for each** augmented image **do**
13: Extract low-level features using convolutional blocks.
14: Generate capsule features using squash activation: $v_j = (||s_j||^2 / (1 + ||s_j||^2)) \times (s_j / ||s_j||)$
15:**end for**
16: Construct multi-scale feature pyramid representations from capsule features
17**: for** each pyramid scale **do**
18: Apply capsule vision transformer self-attention: $\text{Attention}(Q, K, V) = \text{Softmax}((Q \times K^T) / \sqrt{d}) \times V$
19 Encode global contextual and hierarchical relationships.
20:**end for**
21: Transfer knowledge from teacher (PCapN) to student (LCapN).
22: Compute the combined distillation loss: $L_{total} = \alpha L_{LBDK} + \beta L_{DRKD} + \gamma L_{CRKD} + \delta L_{SSLKD}$.
23: Update student model parameters using backpropagation.
24: Fuse transformer-encoded features across all pyramid levels.
25: Enhance fused features using the feature fusion head.
26: **for each** enhanced representation **do**
27: Predict class label $C \in \{$ Normal, Cyst, Tumor, Stone $\}$.
28: **if** C = Tumor **then**
29 : Estimate tumor location and size and generate bounding box.
30: **else** assign corresponding non-tumorous class label.
31: **end for**
32:Generate PCG-CAM heatmaps using centralized gradients: $G'_i = G_i - \mu (G)$,
33: Display predictions with explainability maps and evaluate performance using Accuracy, Precision, Recall, and F1-score.
34 **end algorithm**

Overall Prediction-Process Pipeline Framework

The proposed algorithm facilitates automated kidney disease classification through a systematic pipeline, beginning with the ingestion of contrast-enhanced CT images standardized via lossless JPEG conversion and expert labeling1. Initial image refinement involves resizing data to 224 x 224 pixels and applying statistical normalization to ensure uniform intensity ranges . Robustness is further enhanced through augmentation techniques, including 15-degree rotations, flipping, and Random Image Cropping and Patching (RICAP) . Low-level and capsule features are then extracted using squash activation , to preserve critical spatial hierarchies . Global contextual relationships are subsequently modeled by applying a capsule vision transformer's self-attention mechanism across multi-scale feature pyramids . Knowledge is transferred from a teacher (PCapN) to a student (LCapN) model by minimizing a multi-faceted distillation loss. Final predictions categorize scans into Normal, Cyst, Tumor, or Stone classes, with tumor detections triggering precise location and size estimation. To ensure transparency, PCG-CAM heatmaps are generated using centralized gradients, $G'_i = G_i - \mu (G)$, providing clinicians with explainable visual evidence.

Results and Discussion

This section provides a rigorous empirical assessment and performance analysis of the proposed diagnostic architecture. Specifically engineered for high-precision detection of kidney tumors (KT) within CT imagery, the framework addresses the need for efficient clinical classification. The implementation was conducted using the Python programming language, utilizing the Spyder integrated development

environment alongside advanced deep learning libraries for model synthesis. To facilitate rapid training and reliable performance quantification, all computational tasks were executed on high-performance Graphics Processing Units (GPUs).

Dataset and Description

The performance of the proposed classification framework was evaluated using the CT Kidney Dataset, a comprehensive repository sourced from Kaggle containing 12,446 high-resolution clinical images. The data distribution includes 3,709 cyst instances, 2,283 tumor samples, 1,377 stone cases, and 5,077 normal renal scans, providing a robust foundation for multiclass diagnostic modeling. To ensure effective learning and unbiased performance quantification, the dataset was partitioned using an 80/20 split for training and independent testing, respectively. By utilizing these pre-processed images, the system successfully distinguishes between various renal pathologies with high clinical precision. This balanced approach to data utilization directly supports the model's ability to generalize across diverse diagnostic scenarios, mitigating risks associated with overfitting.

Implementation Details of the Explainable Knowledge Distillation Framework

The proposed architecture facilitates automated kidney disease categorization through a structured, end-to-end pipeline that integrates advanced feature extraction with clinical interpretability , the hyperparameters of the model are set as in table 1:

Table 1. Implementation Details

Parameter	Details
Input Image Size	224 x 224
Kernel Size	3 × 3
Activation Function	Rectified Linear Unit (ReLU)
Pooling Operation	Max pooling
Knowledge Distillation	Teacher–student learning with PCapN & LCapN model.
Loss Function	Cross-Entropy Loss + Localization Loss (Bounding Box).
IoU Threshold	0.5 .
Optimizer	Adam optimizer .
Initial Learning Rate	Set to 0.001 .
Learning Rate Policy	learning rate updated every 10 epochs.
Batch Size	8,
Weight Decay	0.0005,
Dropout Rate	Ranges from 0.2 to 0.5
Weight Initialization	Xavier initialization

Performance Metrics

The diagnostic efficacy of the proposed model is rigorously validated through industry-standard metrics, including Accuracy, Precision, Recall, IoU and the F1-Score [16,17]. These benchmarks are fundamental for assessing the discriminative and localization capabilities of sophisticated computer vision architectures [18,19]. The research utilizes a suite of multi-faceted metrics, including Precision to minimize false positives and Recall to ensure no malignant lesions are overlooked. The F1-Score is implemented as a

harmonic mean to reflect overall classification integrity, while IoU and mAP are employed to quantify spatial localization accuracy and detection performance across varied tumor scales. These benchmarks provide a robust foundation for comparing the architecture's predictive "purity" and sensitivity against existing state-of-the-art methodologies [20,21].

Performance Evaluation

The performance comparative analysis evaluates the proposed model's diagnostic efficacy by benchmarking it against established state-of-the-art architectures, including InceptionV3, IncResNetV2, MobileNet, Xception, and Stacked EnsembleNet.

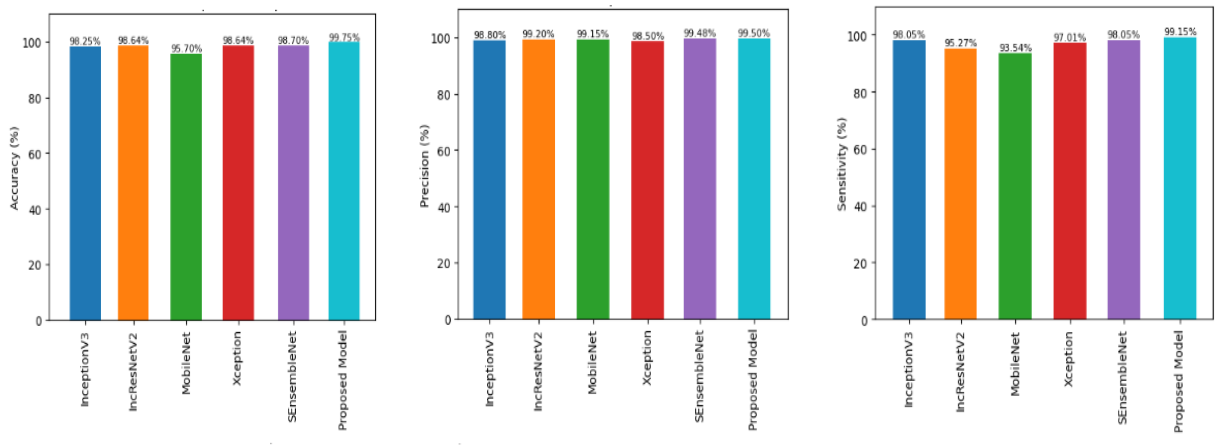

Figure 2. Performance Comparative Analysis: a) Accuracy, b) Precision, c) Sensitivity

Figure 2 (a) illustrates accuracy performance comparison of proposed model with state-of-the-art architectures, including InceptionV3, IncResNetV2, MobileNet, Xception, and Stacked EnsembleNet. The proposed model achieves a peak accuracy of 99.75%, outperforming all compared baseline architectures. In contrast, standard models such as InceptionV3 and Xception yield lower accuracies of 98.25% and 98.64%, respectively, while MobileNet records the lowest performance in this category at 95.70%. The high accuracy of the system is attributed to the integration of the Capsule Vision Transformer, which effectively captures long-range spatial dependencies that traditional scalar-based CNNs often overlook.

Figure 2 (b) illustrates precision performance comparison . Regarding precision, the proposed framework reaches a leading value of 99.50%, ensuring a high degree of reliability in identifying true positive cases of cysts, stones, and tumors. This exceeds the precision recorded for SEnsembleNet (99.48%) and IncResNetV2 (99.20%). The superior precision is primarily driven by the diverse knowledge distillation (DKD) process, which enables the student model (LCapTN) to inherit refined discriminative features from the teacher model (PCapTN) while minimizing false positive detections.

Figure 2 (c) illustrates Sensitivity (Recall) performance comparison . The sensitivity of the proposed model is quantified at 99.15%, which is significantly higher than the 93.54% achieved by MobileNet and the 95.27% achieved by IncResNetV2. This high recall rate is critical in a clinical setting, as it minimizes the risk of missing actual kidney abnormalities. The inclusion of the Principal Component of Gradient-Class

Activation Mapping (PCG-CAM) supports this metric by highlighting critical diagnostic regions, allowing the network to focus on the most salient features of the CT-kidney dataset

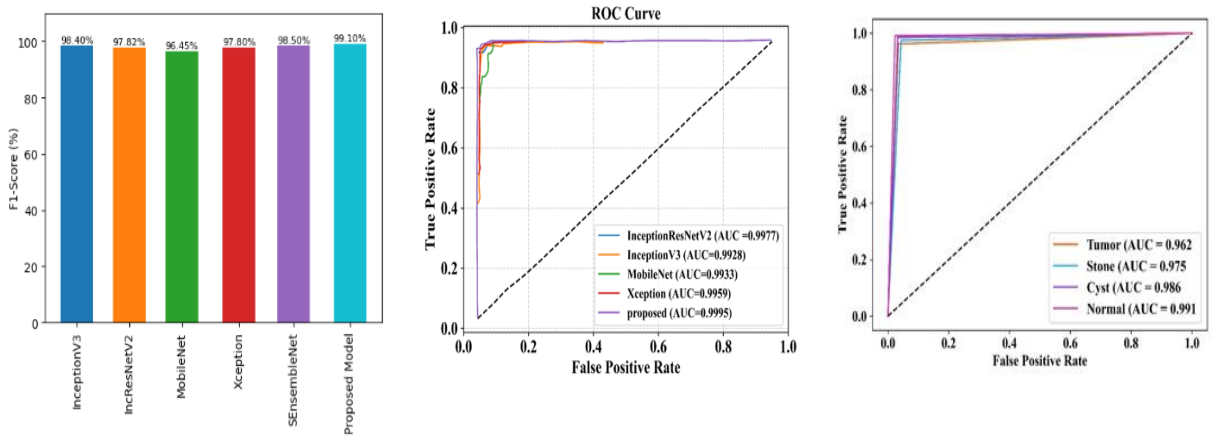

Figure 3. Performance Comparative Analysis: d) F1 Score, e) RoC Curve-Model-wise, f) RoC Curve Class-wise

Figure 3 (d) illustrates F1-score performance comparison of proposed model. In the comparative assessment visualized in Plot (d), F1-score values of InceptionV3 is 98.40%, whereas InceptionResNetV2 has 97.82%, MobileNet has 96.45 % , Xception has 97.80%, Stacked Ensemble Network has 98.50% and the proposed model attains a definitive peak F1-score of 99.10%. This superior metric validates the efficacy of the Capsule Vision Transformer and knowledge distillation framework in achieving nearly perfect diagnostic consistency across all kidney pathology classifications.

Figure 3 (e) illustrates Analytical Evaluation of Model-wise ROC Curves. The proposed model attains the most authoritative performance with a near-perfect AUC of 0.9995. The geometric path of its ROC curve adheres most closely to the upper-left axis, mathematically confirming its superior ability to maximize true positive detections while virtually eliminating false alarms across all renal diagnostic categories.

Figure 3 (e) illustrates Analytical Evaluation of class-wise ROC Curves. This class-specific visualization is essential for understanding the model's reliability in distinguishing between healthy tissue and various pathological states. The Normal class achieves AUC of 0.991, signifying near-perfect discrimination between healthy kidneys and those with abnormalities. Similarly, the Cyst category demonstrates high sensitivity and specificity with an AUC of 0.986, proving that the model can effectively capture subtle morphological variations characteristic of fluid-filled sacs. For the Stone classification, the architecture records an AUC of 0.975, reflecting its proficiency in identifying dense, calcified structures with minimal inter-class confusion. Finally, the Tumor class which typically presents the highest degree of diagnostic complexity due to varying shapes and intensities attains a robust AUC of 0.962. Collectively, these metrics confirm that the proposed Capsule Vision Transformer framework maintains exceptional performance across all diagnostic categories, ensuring reliable clinical support for both benign and malignant conditions.

Conclusion

The research presents an explainable diagnostic framework for automated kidney disease classification utilizing a novel Capsule Vision Transformer integrated with Self-Supervised Diverse Knowledge Distillation . The methodology successfully standardizes CT-kidney data through a multi-stage

preprocessing pipeline involving resizing to 224 x 224 pixels, statistical normalization, and geometric data augmentation . By employing a teacher-student paradigm, the high-capacity Perceptive Capsule Transformer Network (PCapTN) effectively transfers intricate spatial and relational knowledge to the Lightweight Capsule Transformer Network (LCapTN), significantly reducing computational complexity without sacrificing performance . The inclusion of Principal Component of Gradient-Class Activation Mapping (PCG-CAM) further ensures clinical accountability by visualizing the decisive anatomical regions within the CT scans . While the system demonstrated exceptional empirical results achieving an accuracy of 99.75%, precision of 99.50%, Recall of 99.15 % and an F1-score of 99.10% . The study is currently constrained by its reliance on a single dataset. Future research will aim to overcome this limitation by integrating the model into real-time clinical workflows and exploring multi-modal data fusion, combining CT imaging with longitudinal clinical records to further refine diagnostic precision in real-world environments.

References

1. Liang, Q., Lin, H., Li, J., Luo, P., Qi, R., Chen, Q., Meng, F., et al., 2025. Combining multifrequency magnetic resonance elastography with automatic segmentation to assess renal function in patients with chronic kidney disease. *Journal of Magnetic Resonance Imaging*. Doi: 10.1002/jmri.29719
2. Ghosh, A., and Chaki, J., 2025. Fuzzy enhanced kidney tumor detection: Integrating Machine Learning Operations for a fusion of twin transferable network and weighted ensemble machine learning classifier. *IEEE Access*. Doi : 10.1109/ACCESS.2025.3526272
3. Khan, S. U. R., 2025. Multi-level feature fusion network for kidney disease detection. *Computers in Biology and Medicine*, *191*, 110214. Doi: 10.1016/j.compbiomed.2025.110214
4. Zhang, X., Hu, Y., Li, H., Chen, J., Lv, C., Yang, X., Liu, F., Chen, X., and Dong, H., 2025. Artificial intelligence-assisted wearable porous eutectogel with high-performance NH3 enrichment and visual sensing enables non-invasive monitoring of chronic kidney disease. *Chemical Engineering Journal*, *507*, 160678. Doi: 10.1016/j.cej.2025.160678
5. Chakkaravarthy, M., Karras, D. A., and Masal, K. M., 2025. Advancing Deep Learning Techniques for Early Detection and Classification of Renal Cell Carcinoma: A review. *SGS-Engineering & Sciences*, *1*(1). https://spast.org/techrep/article/view/5265
6. Ficili, I., Giacobbe, M., Tricomi, G., and Puliafito, A., 2025. From sensors to data intelligence: Leveraging IoT, cloud, and edge computing with AI. *Sensors*, *25*(6), 1763. Doi : 10.3390/s25061763
7. Moreno-Sánchez, P. A., 2023. Data-driven early diagnosis of chronic kidney disease: development and evaluation of an explainable AI model. *IEEE Access*, *11*, pp.38359-38369. Doi: 10.1109/ACCESS.2023.3264270.
8. Arifuzzaman, M., Ahmed, I., Chowdhury, M. J. U., Sakib, S., Rahman, M. S., Hossain, M. E., and Absar, S., 2024. A Novel Ensemble-Based Deep Learning Model with Explainable AI for Accurate Kidney Disease Diagnosis. Doi: 10.48550/arXiv.2412.09472.
9. Moon, S. W., Kim, J., Kim, Y. J., Kim, S. H., An, C. S., Kim, K. G., and Jung, C. K., 2025. Leveraging explainable AI and large-scale datasets for comprehensive classification of renal histologic types. *Scientific Reports*, *15*(1), 1745. Doi: 10.1038/s41598-025-85857-8

10. Shingade, S. D., Chakkaravarthy, M., Karras, D., and Masal, K. M., 2025. Meta-Heuristic Optimized Dual-Attention Deep Network with Depth-wise Separability for High-Precision Renal Tumor Diagnosis. *Acta Scientiae*, *26*(3), pp.231–243. https://www.periodicos.ulbra.org/index.php/acta/article/view/551.
11. Shingade, S. D., Chakkaravarthy, M., Karras, D., and Masal, K. M., 2025. An Efficient Multi-Scale Hybrid Transformer for High-Precision Renal Malignancy Identification. *Acta Scientiae*, *26*(3), pp.244–257. https://www.periodicos.ulbra.org/index.php/acta/article/view/552.
12. Hossain, M. J., Monir, M. F., and Ahmed, T., 2025. Optimized ROI Extraction and Pixel Reduction Methods for Kidney Abnormality Detection. In *SoutheastCon 2025*, pp. 1091-1096. IEEE. Doi: 10.1109/SoutheastCon56624.2025.10971551.
13. Almuayqil, S. N., El-Ghany, S. A., El-Aziz, A. A. A., and Elmogy, M., 2024. KidneyNet: A Novel CNN-Based Technique for the Automated Diagnosis of Chronic Kidney Diseases from CT Scans. *Electronics*, *13*(24), 4981. Doi: 10.3390/electronics13244981
14. Pimpalkar, A., Saini, D. K. J. B., Shelke, N., Balodi, A., Rapate, G., and Tolani, M., 2025. Fine-tuned deep learning models for early detection and classification of kidney conditions in CT imaging. *Scientific Reports*, *15*(1), 10741. Doi: 10.1038/s41598-025-94905-2
15. Fuladi, S., Chaturvedi, H., Nallakaruppan, M. K., Grover, V., Alshahrani, H., and Baza, M., 2024. Efficient Approach for Kidney Stone Treatment Using Convolutional Neural Network. *Traitement du Signal*, *41*(2), 929. Doi: 10.18280/ts.410233
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. Doi:10.1109/ICECA58529.2023.10394723
17. 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. Doi: 10.1109/ICECA58529.2023.10394801
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. Doi: 10.1109/IICCCS61609.2024.10763601 .
19. Sachin, S., and Mudhalwadkar, R. P., 2024. Development of crop prediction model using data analytics and machine learning techniques. PhD thesis. Savitribai Phule Pune University, Department of Technology. Available at: http://hdl.handle.net/10603/671678.
20. 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. Doi: 10.1063/5.0230840
21. 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). https://spast.org/techrep/article/view/5467.