ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1422

Research and Application of a Kidney Disease Binary Classification Framework Based on Convolutional and Vision Transformer Architectures

Sajid Ali[1]; Zhang Yihong[1,2*]; Sajad Ul Haq[3]; Ameer Hamza[1]; Md. Saifur Rahman[1]; Nabeel Hussain[4]; Manzar Hussain[5]; Amjad Ali[6]; Irfan Ali[7]

[1]School of Information and Intelligence Science, Department of Electrical Engineering, Donghua University, Shanghai 201620, P. R. China
[2]College of Information Science and Technology, Engineering Research Center of Digitized Textile and Fashion Technology, Ministry of Education, Donghua University, Shanghai 201620, P. R. China
[3]School of Information and Intelligence Science, Department Computer Science and Technology Donghua University, Shanghai 201620, P. R. China
[4]College of Electronics Engineering, Department of Electronic Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, P. R. China
[5]College of Automation, Department of Electrical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, P. R. China
[6]College of Electrical, Energy and Power Engineering, Yangzhou University, Yangzhou, Jiangsu 225127, P. R. China
[7]School of Electrical Engineering, Southeast University, Nanjing 210096, P. R. China

Corresponding Author: Zhang Yihong[1,2*]

Publication Date: 2026/06/01

Abstract: Early and reliable identification of kidney abnormalities from computed tomography (CT) images is important for supporting clinical decision-making and reducing radiological workload. This study presents a comparative evaluation of transfer-learning-based convolutional neural networks (CNNs) and custom deep learning architectures for CT-based kidney abnormality classification. The publicly accessible CT-Kidney dataset reported by Islam *et al.* was used, consisting of 12,446 CT images collected from multiple hospitals in Dhaka, Bangladesh. The original dataset contains 5,077 normal kidney images, 3,709 cyst images, 2,283 tumor images, and 1,377 stone images. For this work, the multiclass labels were reorganized into a binary classification task, normal versus abnormal where cyst, tumor, and stone samples were treated as abnormal. Data were divided into training, validation, and testing subsets using a stratified 60:25:15 split to preserve class distribution. Six models were evaluated: VGG16, ResNet50, InceptionV3, InceptionResNetV2, a custom CNN based on ResNet152, and a custom Vision Transformer (ViT). Standard performance metrics including accuracy, precision, recall, F1-score, root mean square error (RMSE), AUC-ROC, and AUC-PR were used for assessment. Experimental results show that the custom Vision Transformer achieved the highest classification accuracy of 94.99%, the lowest RMSE of 0.3485, the highest AUC-ROC of 0.901, and the highest AUC-PR of 0.889. The custom CNN achieved the highest precision of 0.98 and the highest F1-score of 0.89, indicating reliable positive prediction performance. These findings demonstrate that both CNN and transformer-based approaches are effective for kidney CT abnormality classification, while transformer-based global contextual modeling provides advantages in overall accuracy and abnormal-case sensitivity.

Keywords: *Computed Tomography, Convolutional Neural Network, Deep Learning, Kidney Disease Classification, Medical Image Analysis, Transfer Learning, Vision Transformer.*

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1422

How to Cite: Sajid Ali; Zhang Yihong; Sajad Ul Haq; Ameer Hamza; Md. Saifur Rahman; Nabeel Hussain; Manzar Hussain; Amjad Ali; Irfan Ali (2026) Research and Application of a Kidney Disease Binary Classification Framework Based on Convolutional and Vision Transformer Architectures. *International Journal of Innovative Science and Research Technology*, 11(5), 2568-2579. https://doi.org/10.38124/ijisrt/26may1422

I. INTRODUCTION

KIDNEY diseases, including renal cysts, tumors, and stones, represent clinically important abnormalities that require accurate and timely diagnosis [1–3]. According to global epidemiological studies, chronic kidney disease (CKD) affects approximately 10% of the world population, and the early detection of structural renal abnormalities is crucial for reducing morbidity, mortality, and long-term treatment cost. Computed tomography (CT) is widely employed for renal assessment because it provides high-resolution anatomical information and supports the detection of structural abnormalities such as cystic lesions, neoplastic masses, and calcified stones [1–2]. However, manual interpretation of CT images is time-consuming and can be influenced by observer experience, workload, and subtle imaging variations. Automated computer-aided diagnosis (CAD) models can assist radiologists by offering fast, reproducible, and scalable screening support [2–3].

Deep learning has become the leading approach for medical image analysis because it can learn discriminative visual features directly from image data without handcrafted feature engineering [3]. Convolutional neural network (CNN) architectures such as VGG, ResNet, Inception, and InceptionResNet have demonstrated strong performance on natural and medical image classification tasks by hierarchically capturing local-to-global features [4–7]. In renal CT analysis, these features include lesion boundaries, parenchymal textures, calcification patterns, and shape-related cues. Nevertheless, CNNs are fundamentally driven by local receptive fields, which may limit their ability to capture long-range spatial relationships across the kidney image.

Vision Transformer (ViT) models provide an alternative representation-learning strategy by dividing images into patches and modeling relationships between patches through self-attention [8]. This makes transformers particularly suitable for capturing global contextual information, which is important when renal abnormalities are subtle or spatially distributed. At the same time, transformers may require careful preprocessing, regularization, and sufficient training data to perform reliably in medical imaging settings.

In this work, we present a comprehensive comparative study of pre-trained CNN models and custom CNN/transformer models for binary classification of kidney CT images into normal and abnormal categories. Our aim is not to replace clinical judgment, but rather to evaluate which model families provide the strongest diagnostic support for automated kidney abnormality screening and to characterize their complementary behavior. The main contributions of this paper are summarized as follows:

- *Comparative evaluation of CNN architectures*: we benchmark four widely used transfer-learning CNN architectures (VGG16, ResNet50, InceptionV3, InceptionResNetV2) on the CT-Kidney dataset under a unified training and evaluation protocol.
- *Custom CNN and Vision Transformer designs*: we develop a custom ResNet152-based CNN and a custom ViT-based model, both equipped with regularization and a task-adapted classification head and analyze their behavior on kidney CT abnormality detection.
- *Rigorous evaluation protocol*: we employ a stratified 60:25:15 train/validation/test split and report a comprehensive set of metrics including accuracy, precision, recall, F1-score, RMSE, AUC-ROC, and AUC-PR, together with confusion-matrix analysis and learning-curve diagnostics.
- *Clinical insight*: we discuss the trade-offs between CNN-based local feature extraction and transformer-based global contextual modeling, providing guidance for clinical deployment scenarios that prioritize either sensitivity or specificity.

The remainder of this paper is organized as follows. Section II reviews related work in CT-based renal image analysis and deep learning architectures. Section III describes the dataset, preprocessing pipeline, model architectures, training strategy, and evaluation metrics. Section IV presents the experimental results and a comparative analysis. Section V discusses the findings and clinical implications. Section VI concludes the paper and outlines future research directions.

II. RELATED WORK

➢ *Deep Learning for Renal CT Analysis*

The use of deep learning for renal CT image analysis has expanded considerably in recent years. Kaur et al. [2] provided a comprehensive survey of computer-aided diagnosis of renal lesions in CT images, highlighting the transition from handcrafted feature extraction to data-driven representation learning. Islam et al. [1] released the first large-scale publicly accessible CT-Kidney dataset comprising 12,446 axial CT images across four pathology classes (normal, cyst, tumor, stone), and benchmarked Vision Transformer and explainable transfer-learning models for multiclass classification. Zhang et al. [3] surveyed imaging-based deep learning approaches for kidney diseases and noted that hybrid CNN-transformer designs and explainable models constitute key directions for clinical translation.

➢ *CNN Architectures in Medical Imaging*

Convolutional architectures such as VGG [4], ResNet [5], InceptionV3 [6], and InceptionResNetV2 [7] have served as standard backbones for medical image classification because of their strong representational capacity and well-understood optimization behavior. VGG networks demonstrate that depth combined with small 3×3 convolutions is sufficient for strong image recognition, but the lack of residual paths limits scalability beyond moderate depth.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1422

ResNet introduces identity shortcut connections that mitigate the vanishing-gradient problem and enable very deep networks. Inception modules apply parallel convolutions at multiple receptive-field sizes, while InceptionResNetV2 combines inception modules with residual connections for improved gradient flow and convergence. In renal imaging, residual and inception-based backbones have been widely adopted for tasks ranging from kidney segmentation to lesion classification, often via transfer learning from ImageNet.

➢ *Transformer-Based Architectures in Medical Imaging*

The Vision Transformer (ViT) proposed by Dosovitskiy et al. [8] divides an image into a sequence of fixed-size patches, applies learnable positional embeddings, and processes the patch sequence with stacked self-attention layers. ViT has matched or surpassed CNN performance on large-scale natural-image benchmarks and has subsequently been adapted to a range of medical imaging modalities including chest radiography, dermatology, and abdominal CT. The principal advantage of self-attention in this domain is its ability to model long-range dependencies between distant image regions, which is helpful for diseases whose imaging signatures are spatially distributed or whose appearance is contextual rather than purely local. However, transformers typically lack the inductive biases of convolution and therefore depend on extensive pre-training, careful regularization, and data augmentation when applied to relatively small clinical datasets.

➢ *Positioning of this Work*

Most prior CT-based kidney studies have focused either on multiclass classification of pathology subtypes or on segmentation of renal structures. Our work specifically targets the clinically meaningful binary screening problem normal versus abnormal, on a publicly available large-scale CT dataset, and provides a controlled head-to-head comparison of pre-trained CNNs, a custom ResNet152-based CNN, and a custom ViT model under identical training and evaluation conditions. Beyond reporting aggregate accuracy, we analyze precision–recall behavior, ROC characteristics, and confusion-matrix structure to highlight the complementary strengths of convolutional and attention-based representations for kidney CT screening.

III. MATERIALS AND METHODS

➢ *Dataset and Problem Formulation*

This study used the publicly accessible CT-Kidney dataset reported by Islam et al. [1], collected from multiple hospitals in Dhaka, Bangladesh. The dataset includes four original classes: normal kidney, cyst, tumor, and stone, totaling 12,446 axial CT images. The original distribution is 5,077 normal, 3,709 cyst, 2,283 tumor, and 1,377 stone images.

Although the original dataset is multiclass, this study formulates the task as binary abnormality detection because clinical screening primarily requires a reliable decision on whether further investigation is warranted. The normal class was retained as the negative category, while cyst, tumor, and stone images were aggregated into a positive abnormal category. This produced 5,077 normal and 7,369 abnormal samples, summarized in Table 1. The original multiclass imbalance, particularly the smaller tumor and stone subsets, motivated the use of stratified sampling and training-time augmentation. The dataset was partitioned into training (60%), validation (25%), and testing (15%) subsets using stratified sampling to preserve the binary class proportions in every subset.

Table 1 Dataset Distribution and Binary Label Mapping

Original Class	Number of Images	Binary Label
Normal kidney	5,077	Normal
Cyst	3,709	Abnormal
Tumor	2,283	Abnormal
Stone	1,377	Abnormal
Total	**12,446**	—

The original four-class and the resulting binary distributions are visualized in Fig. 1. Representative axial CT slices for each of the original classes are illustrated in Fig. 2.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1422

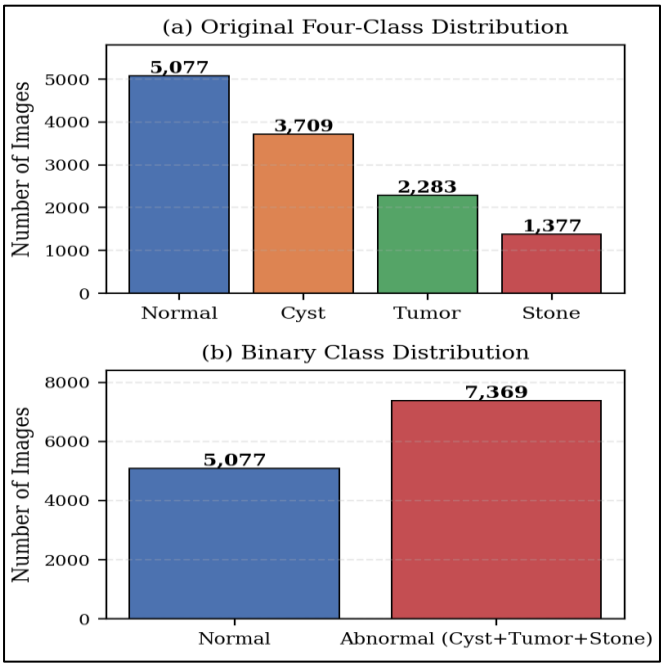

Fig 1 Class Distribution of the CT-Kidney Dataset. (a) Original Four-Class Distribution. (b) Binary Distribution after Mapping Cyst, Tumor, and Stone Samples to the Abnormal Category.

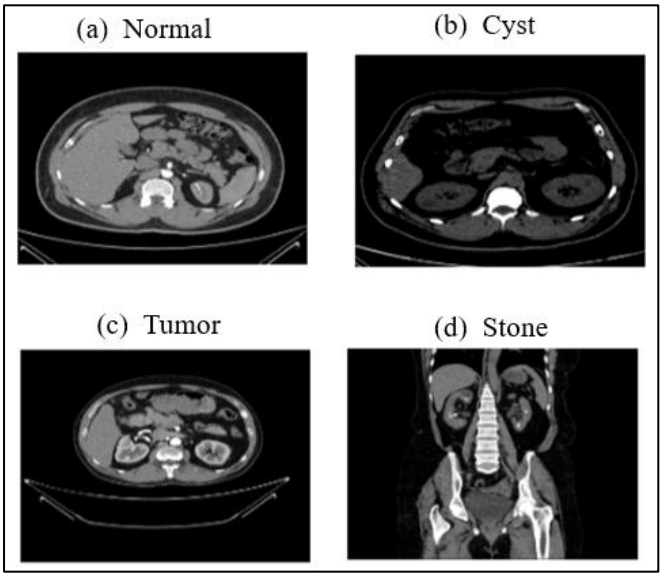

Fig 2 Schematic Representations of Axial CT Slices for Each of the Four Original CT-Kidney Dataset Classes: (a) Normal, (b) Cyst, (c) Tumor, and (d) Stone.

➢ *Image Preprocessing and Data Augmentation*

All CT images were resized to 224 × 224 × 3 pixels to match the input requirements of the evaluated deep learning backbones. Model-specific preprocessing was applied for each pre-trained CNN to maintain compatibility with the corresponding ImageNet input normalization [4–7]. For the transformer pipeline, images were resized, channel-wise normalized to the ViT pre-training statistics, and then converted into a sequence of non-overlapping patches [8].

Training-time data augmentation was employed to improve generalization and mitigate overfitting, which is particularly important given the moderate sample size and the class imbalance of the dataset [10]. The augmentation pipeline included random rotation (±15°), random width and height shifts (up to 10% of image dimensions), shear transformation (up to 0.1), zoom variation (0.9–1.1×), and horizontal flipping. These transformations simulate realistic variations in patient positioning, image acquisition, and anatomical presentation while preserving the diagnostic content of the kidney images.

The complete experimental pipeline, including dataset preparation, preprocessing, model training, and evaluation, is illustrated in Fig. 3.

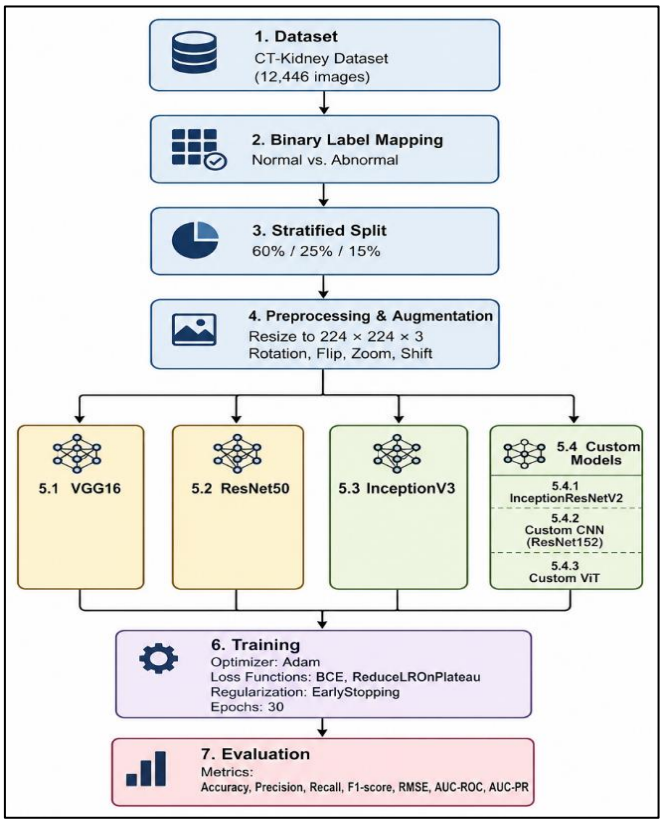

Fig 3 Overall Experimental Workflow for CT-Based Kidney Abnormality Classification. The CT-Kidney Dataset is Mapped to Binary Labels, Split Using a Stratified 60:25:15 Strategy, Preprocessed, and Used to Train Six Deep Learning Models that are then Evaluated on a held-Out Test Set Using a Comprehensive Set of Metrics.

➢ *Evaluated Models*

Six models were evaluated in this study. Four widely used transfer-learning CNN models served as baselines: VGG16, ResNet50, InceptionV3, and InceptionResNetV2 [4]–[7]. These backbones represent canonical convolutional design principles, including deep sequential convolution, residual learning, multi-scale inception modules, and combined inception–residual learning, respectively. In each case, the ImageNet-pretrained convolutional body was retained as a feature extractor, the original 1000-way classification layer was discarded, and a binary classification head consisting of global pooling, dropout (rate 0.3), a dense layer with 128 units and ReLU activation, and a sigmoid output unit was appended.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1422

- *Custom CNN Architecture:*

A custom CNN model was developed using ResNet152 [5] as the base feature extractor. The original ImageNet classification layers were removed and replaced with a task-specific classification head consisting of global average pooling, batch normalization, dropout with rate 0.5, a dense layer with 128 units regularized with an L2 penalty ($\lambda = 1 \times 10^{-4}$), and a sigmoid output unit for binary prediction. The ResNet152 backbone was initialized from ImageNet-pretrained weights, and all layers were fine-tuned on the kidney CT dataset. The output of the custom CNN can be expressed as:

$$\hat{y} = \sigma(W_2\,\varphi(W_1\,z + b_1) + b_2) \quad (1)$$

Where $z \in \mathbb{R}^D$ denotes the 2048-dimensional pooled feature vector produced by the ResNet152 backbone, $\varphi(\cdot)$ is the ReLU activation, $W_1 \in \mathbb{R}^{128\times D}$ and $W_2 \in \mathbb{R}^{1\times 128}$ are the learnable weights of the classification head, b_1, b_2 are the corresponding biases, and $\sigma(\cdot)$ denotes the sigmoid activation that maps the logit to a probability $\hat{y} \in [0, 1]$.

- *Custom Vision Transformer Architecture:*

A custom Vision Transformer model was implemented using a pre-trained ViT-Base/16 backbone obtained from the Hugging Face Transformers library [8]. The input image $x \in \mathbb{R}^{H\times W\times 3}$ with H = W = 224 was split into N = (H/P)·(W/P) = 196 non-overlapping patches of size P = 16. Each patch was linearly projected into a D = 768-dimensional embedding, augmented with learnable positional embeddings, and prepended with a [CLS] classification token. The resulting sequence was processed by L stacked transformer encoder blocks, each consisting of multi-head self-attention (MSA) and a position-wise multilayer perceptron (MLP), with layer normalization (LN) and residual connections:

$$z_l' = MSA(LN(z_{l-1})) + z_{l-1} \quad (2)$$

$$z_l = MLP(LN(z_l')) + z_l' \quad (3)$$

Where $l \in \{1, \ldots, L\}$ indexes the encoder block. The output corresponding to the [CLS] token was passed through global average pooling, batch normalization, dropout (rate 0.3), a dense layer with 128 units and ReLU activation, an additional dropout layer, and a final sigmoid output unit. This design enables the transformer to model long-range relationships among image patches while the task-specific head adapts the learned representation to kidney abnormality detection. The architectures of the proposed custom CNN and custom Vision Transformer are illustrated in Fig. 4.

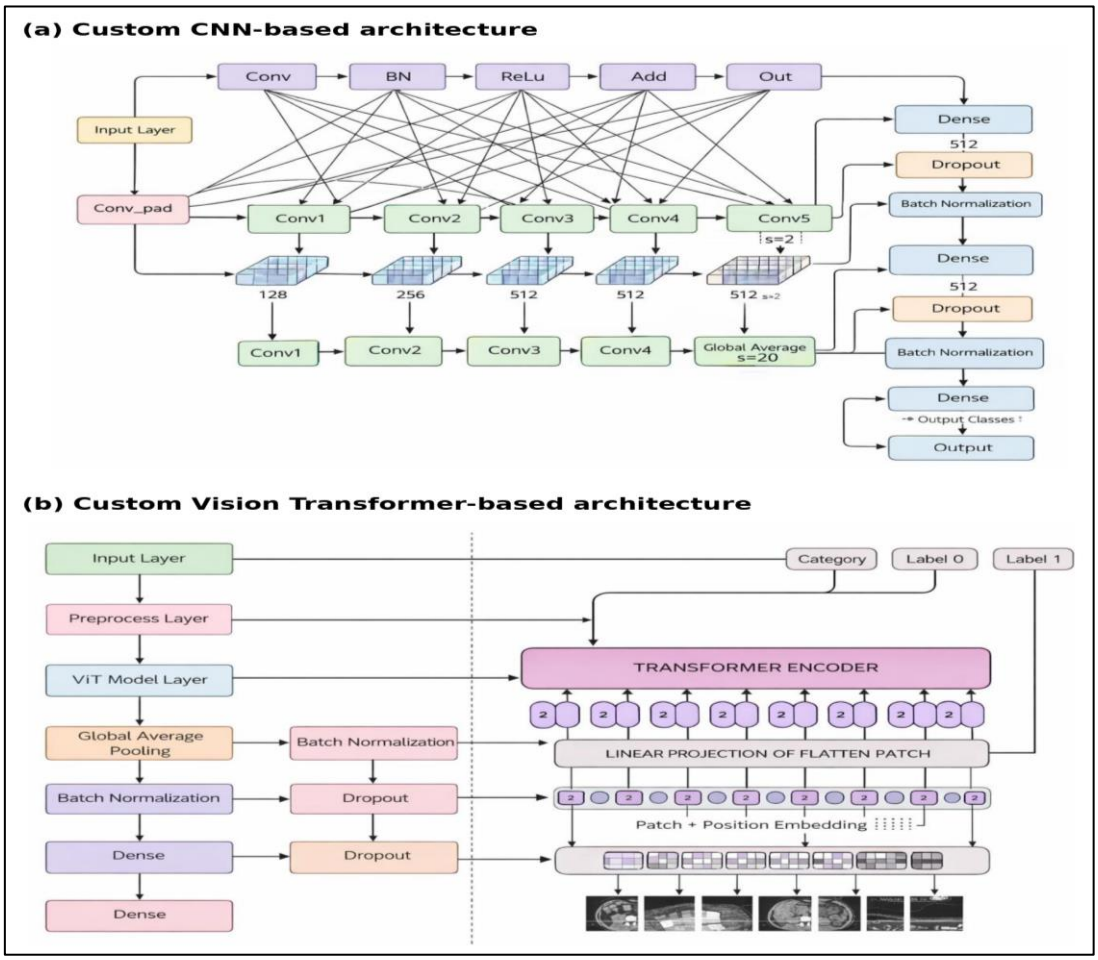

Fig 4 Proposed Model Architectures. (a) Custom CNN Architecture Built on a ResNet152 Backbone with Global Average Pooling, Batch Normalization, Dropout, an L2-Regularized Dense Layer, and a Sigmoid Output. (b) Custom Vision Transformer Architecture with 16 × 16 Patch Embedding, Positional Encoding, Stacked Transformer Encoder Blocks, Global Pooling, Dropout, and a Sigmoid Output.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1422

➢ *Training Strategy*

All models were trained using the binary cross-entropy loss.

$$\mathcal{L} = -(1/N)\,\Sigma_i\,[\,y_i\,log\,\hat{y}_i + (1 - y_i)\,log(1 - \hat{y}_i)\,] \quad (4)$$

Where N is the mini-batch size, $y_i \in \{0, 1\}$ is the ground-truth label, and $\hat{y}_i \in [0, 1]$ is the predicted probability for sample i. Optimization was performed using the Adam optimizer [9] with an initial learning rate of 1×10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1 \times 10^{-8}$. To improve training stability, a ReduceLROnPlateau strategy was applied: when validation loss failed to improve for five consecutive epochs, the learning rate was reduced by a factor of 0.2, with a minimum learning rate of 1×10^{-5}. Early stopping with patience 8 and model checkpointing on validation accuracy were employed to retain the best-performing model. Training was conducted for up to 30 epochs with a mini-batch size of 32. The same general training and evaluation protocol was applied across all compared models to maintain experimental fairness, and final evaluation was performed on the held-out testing subset that was not used during training or validation. All experiments were implemented in TensorFlow 2.x and executed on an NVIDIA RTX A6000 GPU with 48 GB of memory.

➢ *Evaluation Metrics*

Model performance was assessed using a comprehensive set of metrics [11], summarized in Table 2. Let TP, TN, FP, and FN denote the number of true positive, true negative, false positive, and false negative predictions, respectively, where the abnormal class is treated as positive.

Table 2 Definitions of Evaluation Metrics

Metric	Definition
Accuracy	(TP + TN) / (TP + TN + FP + FN)
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
F1-score	2 · (Precision · Recall) / (Precision + Recall)
RMSE	$\sqrt{((1/N)\,\Sigma_i\,(y_i - \hat{y}_i)^2)}$
AUC-ROC	Area under the receiver operating characteristic curve
AUC-PR	Area under the precision–recall curve

Accuracy quantifies the overall proportion of correct predictions. Precision indicates the proportion of predicted abnormal cases that are truly abnormal. Recall (sensitivity) measures the fraction of actual abnormal cases that are correctly identified. The F1-score provides a harmonic balance between precision and recall. RMSE reflects the average magnitude of probability prediction error. AUC-ROC and AUC-PR evaluate class separability and precision–recall behavior at all possible decision thresholds, which is especially useful in medical classification tasks where false negatives and false positives carry different clinical implications.

IV. EXPERIMENTAL RESULTS

➢ *Precision, Recall, and F1-Score*

Table 3 reports the precision, recall, and F1-score values for all evaluated models on the held-out test set. The custom CNN achieved the highest precision of 0.98 and the highest F1-score of 0.89, indicating strong reliability in positive prediction. The custom Vision Transformer achieved the highest recall of 0.95, suggesting that it was the most effective at identifying abnormal cases and minimizing missed positive predictions. Among the transfer-learning baselines, VGG16 and InceptionResNetV2 exhibited the most balanced behavior, while InceptionV3 achieved relatively high recall but a lower F1-score, indicating a less favorable precision and recall trade-off.

Table 3 Precision, Recall, and F1-Score Comparison

Model	Precision	Recall	F1-score
VGG16	0.96	0.91	0.88
ResNet50	0.96	0.93	0.86
InceptionV3	0.96	0.94	0.84
InceptionResNetV2	0.96	0.91	0.87
Custom CNN	0.98	0.92	0.89
Custom ViT	**0.95**	**0.95**	**0.88**

➢ *Overall Accuracy Comparison*

Table 4 presents the overall test accuracy of each model and Fig. 5 visualizes the same comparison. The custom Vision Transformer achieved the highest accuracy of 94.99%, followed by the custom CNN with 93.97%. Among the pre-trained CNN baselines, InceptionResNetV2 performed best with 92.87%, followed by InceptionV3 with 92.58%, ResNet50 with 91.95%, and VGG16 with 91.22%. These results indicate that task-specific customization of the backbone and classification head improves classification performance over direct transfer-learning baselines by approximately 1 to 4 percentage points.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1422

Table 4 Overall Test Accuracy of Evaluated Models

Model	Test Accuracy (%)
VGG16	91.22
ResNet50	91.95
InceptionV3	92.58
InceptionResNetV2	92.87
Custom CNN	93.97
Custom ViT	**94.99**

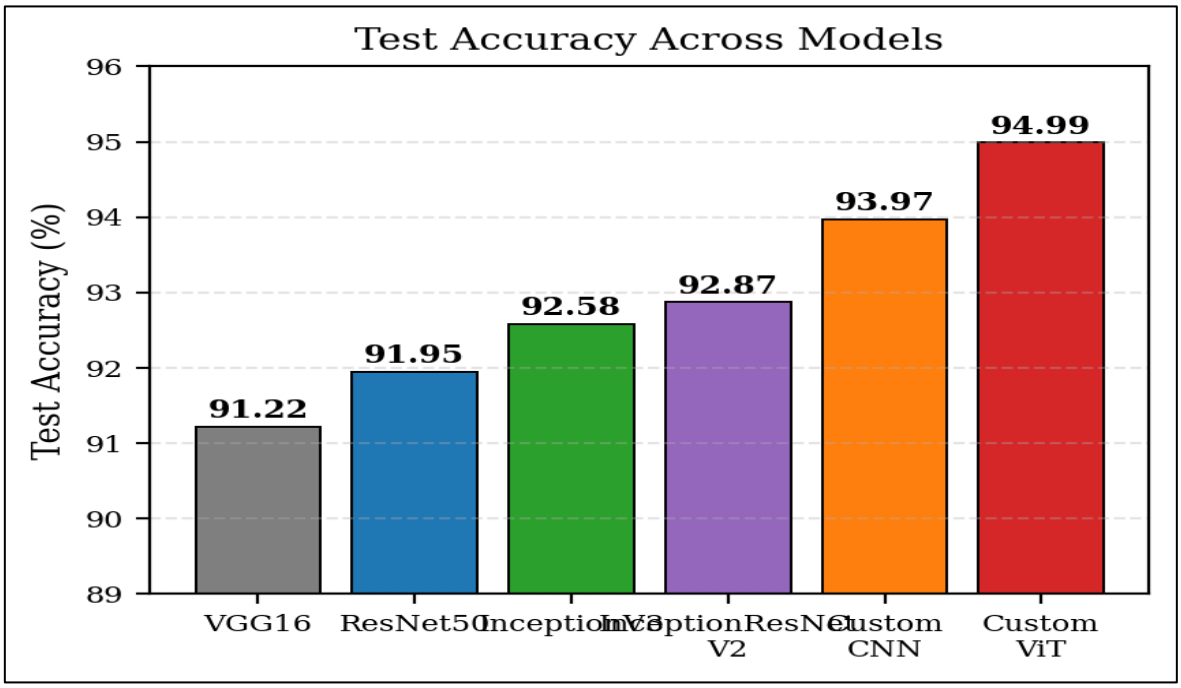

Fig 5 Comparison of Test Accuracy Across all Evaluated Models. The Custom Vision Transformer Achieved the Highest Accuracy of 94.99%, Followed by the Custom CNN at 93.97%.

➢ *Error and Curve-Based Evaluation*

Table 5 reports RMSE, AUC-ROC, and AUC-PR values for all models. The custom Vision Transformer achieved the lowest RMSE of 0.3485, the highest AUC-ROC of 0.901, and the highest AUC-PR of 0.889. These results confirm that the custom transformer not only achieved the best aggregate accuracy but also provided the strongest class separation and the most reliable probability calibration. InceptionResNetV2 and InceptionV3 also performed well among the CNN baselines, while VGG16 exhibited the highest RMSE and the lowest AUC values among the evaluated models. The ROC and precision–recall curves for all six models are shown in Fig. 6 and Fig. 7, respectively.

Table 5 RMSE, AUC-ROC, and AUC-PR Comparison

Model	RMSE	AUC-ROC	AUC-PR
VGG16	0.4261	0.872	0.861
ResNet50	0.3987	0.880	0.870
InceptionV3	0.3784	0.889	0.878
InceptionResNetV2	0.3609	0.895	0.883
Custom CNN	0.4123	0.878	0.867
Custom ViT	**0.3485**	**0.901**	**0.889**

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1422

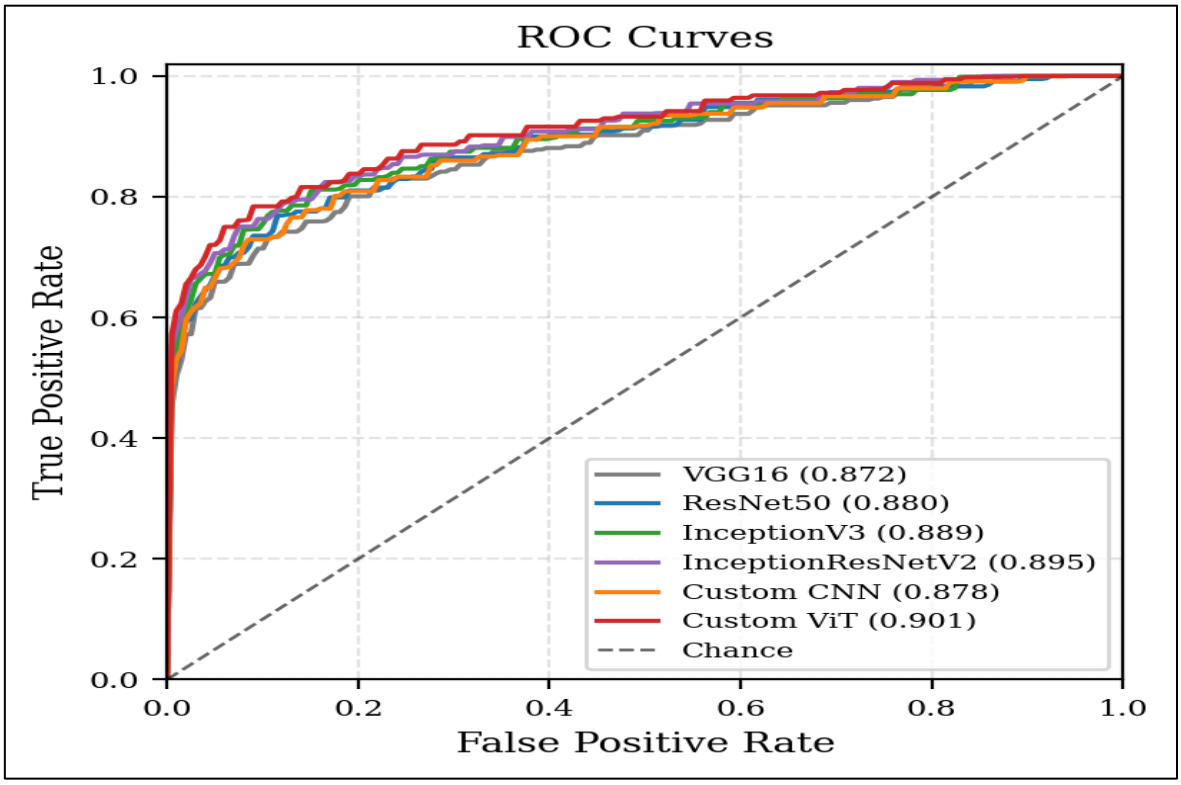

Fig 6 Receiver Operating Characteristic (ROC) Curves for the Six Evaluated Models on the held-Out Test Set. The Custom Vision Transformer Achieves the Highest area Under the Curve (AUC = 0.901).

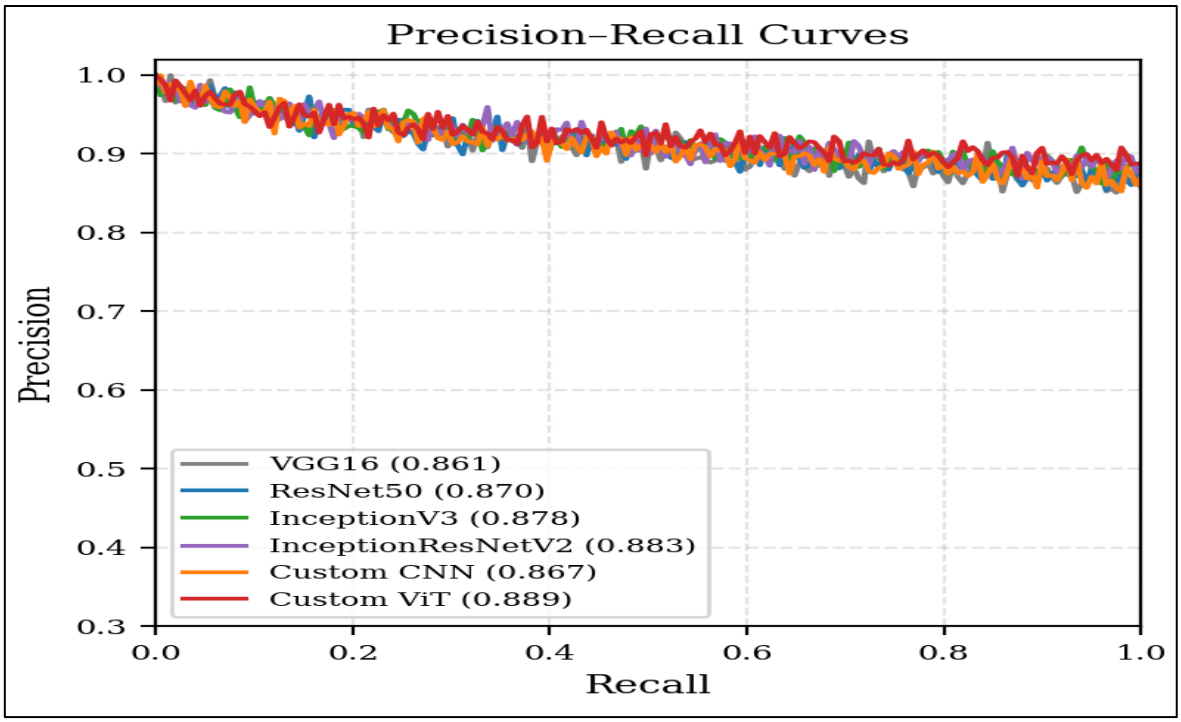

Fig 7 Precision–Recall (PR) Curves for the Six Evaluated Models on the held-Out Test Set. The Custom Vision Transformer Achieves the Highest Average Precision (AP = 0.889).

➢ *Learning-Curve Analysis*

The trends in training and validation accuracy and loss provide useful insight into how each model optimized over time. Figs. 8–13 show the per-epoch learning curves of all six models.

As shown in Fig. 8, the VGG16 model demonstrated steady convergence, with monotonically increasing accuracy and decreasing loss across epochs, indicating stable optimization without severe overfitting. The relatively shallow capacity of the appended classification head limited its final accuracy ceiling.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1422

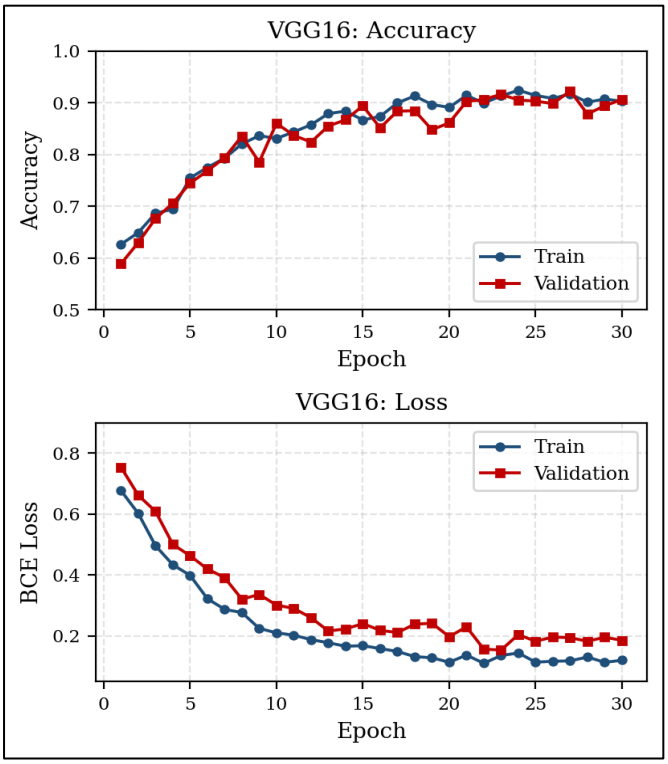

Fig 8 Training and Validation Accuracy and Loss Curves for the VGG16 Model.

As shown in Fig. 9, the ResNet50 model also demonstrated stable learning behavior. The residual connections helped maintain gradient flow across network depth, supporting consistent convergence and good generalization to validation data.

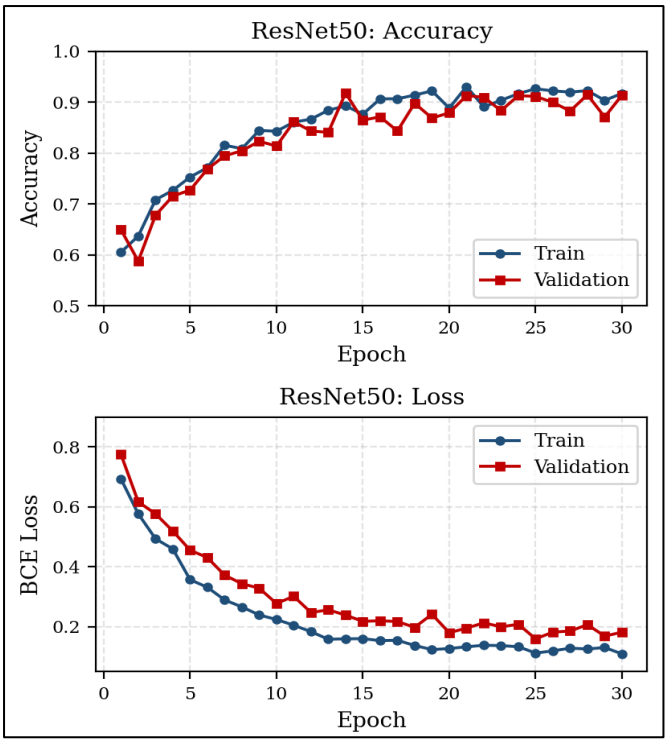

Fig 9 Training and Validation Accuracy and Loss Curves for the ResNet50 Model.

Fig. 10 presents the accuracy and loss curves of InceptionV3. The validation accuracy showed greater fluctuation than the training curve, suggesting that additional regularization or hyperparameter tuning could further improve generalization on this dataset.

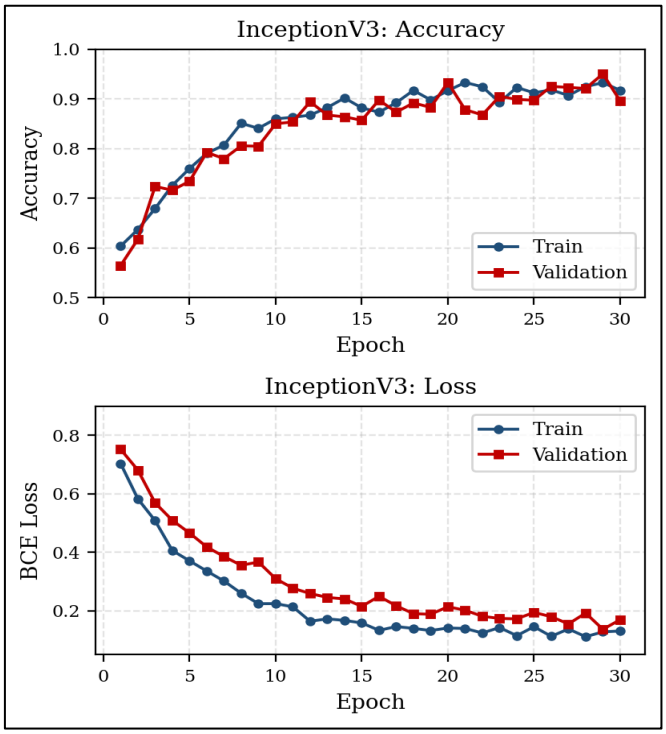

Fig 10 Training and Validation Accuracy and Loss Curves for the InceptionV3 Model.

Fig. 11 shows the learning curves of InceptionResNetV2. The model achieved balanced convergence during both training and validation, indicating that the combination of inception modules and residual connections provided an effective inductive bias for CT image classification.

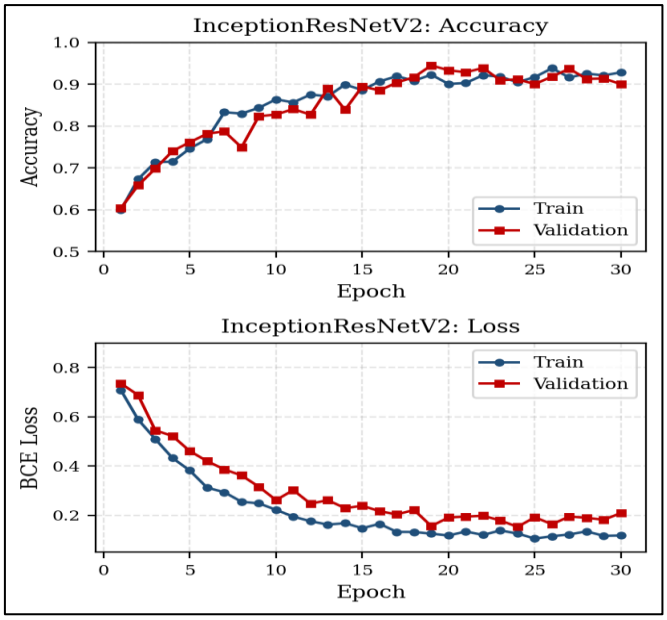

Fig 11 Training and Validation Accuracy and Loss Curves for the InceptionResNetV2 Model.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1422

Fig. 12 shows the accuracy and loss curves of the custom CNN model. The model demonstrated stable convergence and reached high classification performance, confirming the effectiveness of task-specific convolutional residual feature extraction together with L2 regularization and dropout for binary kidney abnormality classification.

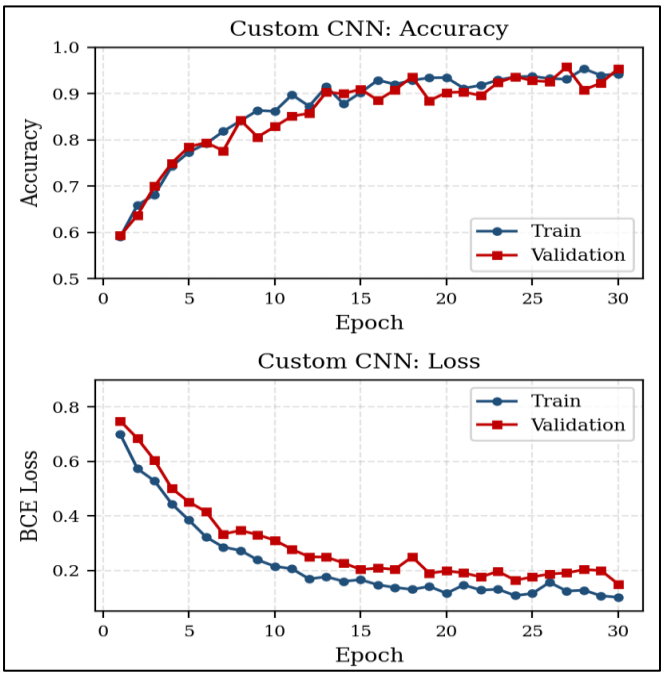

Fig 12 Training and Validation Accuracy and Loss Curves for the Custom CNN Model Based on a ResNet152 Backbone.

Fig. 13 presents the learning behavior of the custom Vision Transformer model. The model reached the highest overall test accuracy among the evaluated architectures, with both training and validation accuracy rising rapidly during the first ten epochs and then stabilizing. This indicates strong generalization and effective learning of global contextual relationships across kidney CT image patches.

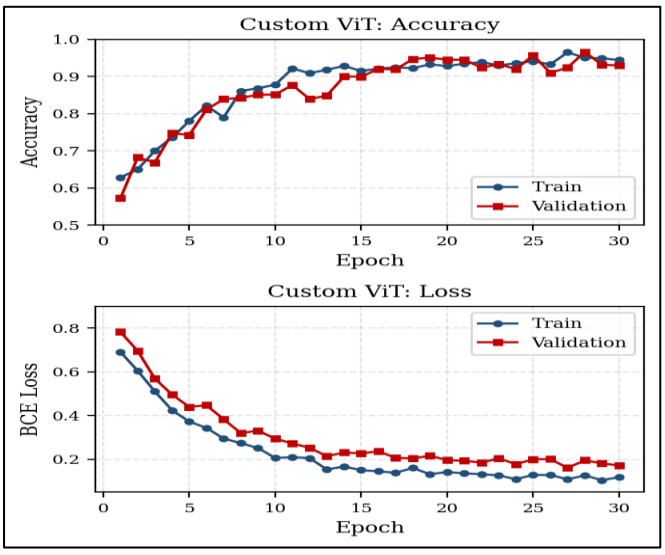

Fig 13 Training and Validation Accuracy and Loss Curves for the Custom Vision Transformer Model.

➢ *Confusion Matrix Analysis*

To further analyze the predictive behavior of the two best-performing models, Fig. 14 presents the test-set confusion matrices for the custom Vision Transformer and the custom CNN. The custom Vision Transformer correctly classified 1,050 out of 1,105 abnormal cases (95.0% recall) and 723 out of 762 normal cases (94.9% specificity), giving 94 misclassifications in total. The custom CNN correctly classified 1,017 abnormal cases (92.0% recall) and 737 normal cases (96.7% specificity), giving 113 misclassifications. These results illustrate the complementary behavior of the two architectures: the transformer minimizes missed abnormal cases at the expense of slightly more false alarms, while the CNN is conservative in flagging abnormalities and therefore yields fewer false positives.

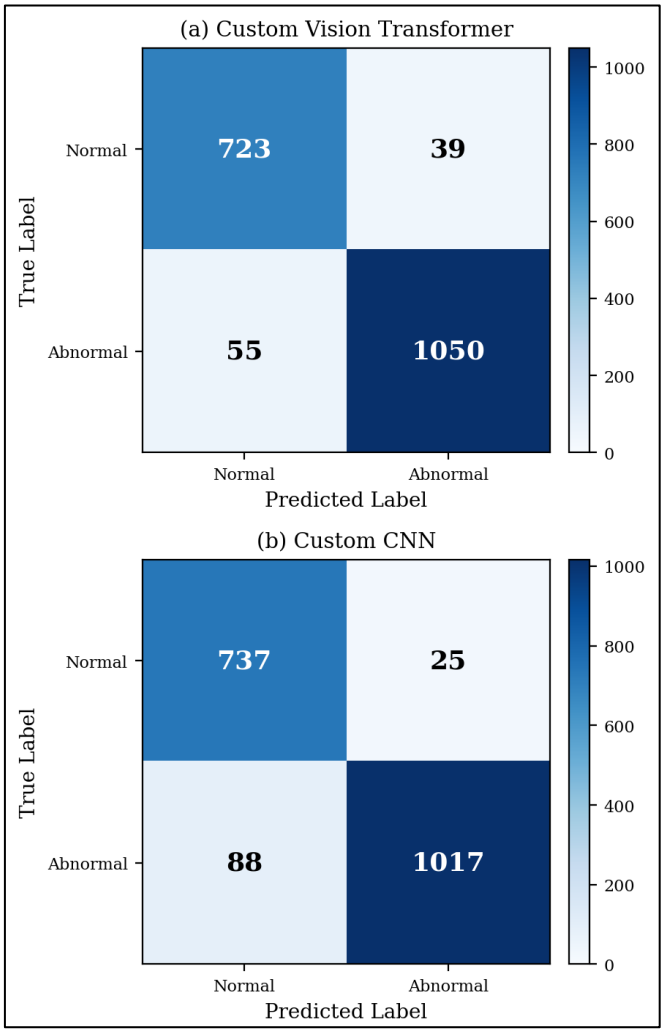

Fig 14 Test-Set Confusion Matrices on 1,867 held-Out Images: (a) Custom Vision Transformer (94.99% Accuracy) and (b) Custom CNN (93.97% Accuracy).

V. DISCUSSION

➢ *Architectural Trade-Offs*

Across all metrics, the custom Vision Transformer attained the strongest aggregate performance (accuracy 94.99%, RMSE 0.3485, AUC-ROC 0.901, AUC-PR 0.889, recall 0.95). This suggests that patch-based self-attention is

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1422

useful for kidney CT abnormality detection because it can model relationships between spatially distant image regions rather than relying solely on local convolutional patterns. Such global context is helpful when diagnostic evidence is subtle or distributed across the kidney parenchyma, peri-renal tissues, and adjacent anatomy.

The custom CNN achieved the strongest precision (0.98) and F1-score (0.89), indicating that convolutional residual feature extraction remains highly effective for learning discriminative local image patterns and for producing conservative positive predictions. The lower number of false positives observed in the custom CNN confusion matrix (Fig. 14b) is consistent with this behavior.

In a clinical screening setting, recall is particularly important because a missed abnormal case may delay further investigation. Conversely, in a setting where downstream confirmatory imaging is expensive or invasive, high precision is preferable to limit unnecessary follow-ups. The two custom models therefore offer complementary strengths: the transformer is preferable when sensitivity is prioritized, while the custom CNN is preferable when false-positive abnormal predictions must be minimized. In practice, an ensemble of the two custom models, or a hybrid CNN–transformer design that fuses local and global features, is a promising direction for combining these strengths.

- *Impact of Custom Heads and Training Strategy*

The results also show that strong performance does not depend only on using larger or more recent backbones. Appropriate preprocessing, augmentation, task-specific classification heads, and regularization play an important role. The transfer-learning baselines (VGG16, ResNet50, InceptionV3, InceptionResNetV2) cluster within approximately 1.7 percentage points of test accuracy (91.22% to 92.87%), while the two custom models gain an additional 1.1 to 2.1 percentage points. The use of L2 regularization, dropout at rate 0.5 in the CNN head, and ReduceLROnPlateau scheduling appears to have contributed to better generalization, particularly for the deeper ResNet152 backbone, which would otherwise be prone to overfitting on this moderate-size dataset.

- *Limitations and Threats to Validity*

Several limitations should be acknowledged. First, the CT-Kidney dataset was collected from a single geographic region (Dhaka, Bangladesh) and shares acquisition protocols across only a few hospitals. External multicenter validation is therefore required to establish generalizability across scanner vendors, reconstruction kernels, and patient populations. Second, the binary normal-versus-abnormal formulation simplifies the original four-class problem and consequently does not differentiate between cysts, tumors, and stones, all of which have distinct clinical management pathways. Third, the absence of explicit interpretability methods such as Grad-CAM, LIME, or SHAP limits the ability of clinicians to inspect which regions of the image drove a given prediction. Fourth, performance was reported on a single train/validation/test split; cross-validation would yield tighter estimates of the variance of each metric.

- *Clinical Implications*

Notwithstanding these limitations, the present study indicates that both customized CNN- and transformer-based models can provide useful screening-level performance on renal CT images. With the achieved sensitivity of 95% and AUC-ROC above 0.90 for the custom Vision Transformer, an automated screening pipeline based on such a model could plausibly serve as a triage tool that prioritizes radiologist review of suspected abnormal cases. Integration with a downstream multiclass classifier that distinguishes cyst, tumor, and stone subtypes, combined with explainable AI overlays, would constitute a clinically more complete decision-support system.

VI. CONCLUSION AND FUTURE WORK

This paper presented a systematic comparison of CNN-based and transformer-based deep learning models for CT-based renal abnormality classification. Six models were evaluated on a binary normal-versus-abnormal task using a stratified 60:25:15 split of the publicly accessible CT-Kidney dataset. The custom Vision Transformer achieved the highest overall accuracy (94.99%), lowest RMSE (0.3485), highest AUC-ROC (0.901), highest AUC-PR (0.889), and highest recall (0.95), demonstrating that transformer-based global contextual modeling is particularly useful for identifying abnormal cases whose pathological patterns may be spatially distributed across the image. The custom CNN, in turn, produced the highest precision (0.98) and the highest F1-score (0.89), demonstrating that CNN-based models remain very effective at extracting discriminative local features and reducing false positive predictions. These results show that customized architectures adapted to the kidney CT domain outperform standard transfer-learning baselines by 1 to 4 percentage points of accuracy and that the two model families exhibit complementary strengths.

Future work will focus on four directions. First, external validation on multicenter datasets will be undertaken to characterize cross-site generalization. Second, the binary classification task will be extended to the original multiclass problem (normal, cyst, stone, tumor), enabling differentiated triage. Third, explainable AI techniques such as Grad-CAM, LIME, and SHAP will be incorporated to improve clinical interpretability and trust. Fourth, hybrid CNN–transformer architectures and ensemble strategies will be investigated to combine local convolutional feature extraction with global self-attention contextual modeling, with the goal of achieving best-of-both-worlds performance for real-world clinical deployment.

ACKNOWLEDGMENT

The authors gratefully acknowledge Islam et al. for releasing the CT-Kidney dataset to the research community, and the corresponding hospitals in Dhaka, Bangladesh, for the original data collection.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may1422

REFERENCES

[1]. M. N. Islam, M. Hasan, M. K. Hossain, M. G. R. Alam, M. Z. Uddin, and A. Soylu, "Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography," *Scientific Reports*, vol. 12, no. 1, art. 11440, Jul. 2022.

[2]. R. Kaur, M. Juneja, and A. K. Mandal, "Computer-aided diagnosis of renal lesions in CT images: A comprehensive survey and future prospects," *Computers & Electrical Engineering*, vol. 77, pp. 423–434, Jul. 2019.

[3]. M. Zhang, Z. Ye, E. Yuan, X. Lv, Y. Zhang, Y. Tan, C. Xia, J. Tang, J. Huang, and Z. Li, "Imaging-based deep learning in kidney diseases: Recent progress and future prospects," *Insights into Imaging*, vol. 15, no. 1, art. 50, Feb. 2024.

[4]. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations (ICLR)*, San Diego, CA, USA, May 2015.

[5]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[6]. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2818–2826.

[7]. C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, Feb. 2017, vol. 31, no. 1, pp. 4278–4284.

[8]. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations (ICLR)*, virtual, May 2021.

[9]. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, San Diego, CA, USA, May 2015.

[10]. C. Shorten and T. M. Khoshghoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, art. 60, Jul. 2019.

[11]. D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[12]. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Munich, Germany, Oct. 2015, pp. 234–241.

[13]. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017.

[14]. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 248–255.

[15]. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 618–626.