

Deep Learning Based Framework for Early Detection and Classification of CKD Using Medical Imaging

A SYNOPSIS

submitted in partial fulfillment of the requirements of the
Degree of

DOCTOR OF PHILOSOPHY

By: Smit Thacker

Enrollment No.: 219999913010

Under the Supervision

of

Dr. Ravirajsinh S Vaghela



**GUJARAT TECHNOLOGICAL UNIVERSITY,
AHMEDABAD**

July – 2026

INDEX

Sr. No.	Topic	Page No.
	Abstract	3
1	Brief Description on the State of the Art of the Research Topic	4
1.1	Introduction	5
1.2	Background: Deep Learning in Kidney Disease Classification	7
1.3	Literature Survey	9
1.4	Research Gaps	12
2	Problem Definition	13
3	Objectives and Scope of Work	14
3.1	Research Objectives	14
3.2	Scope of Work	15
4	Original Contribution by the Thesis	15
5	Methodology, Results and Discussion	16
5.1	Dataset Description	16
5.2	Preprocessing and Data Augmentation	17
5.3	Baseline (CNN) Architectures	18
5.4	Proposed Methodology: RenalViT	19
5.5	Experimental Setup	20
5.6	Comparative Results	21
5.7	Confusion Matrix Analysis	24
5.8	ROC Curve and AUC Analysis	25
6	Achievements with respect to Objectives	26
7	Conclusion	27

8	List of Publications arising from Research Work	28
	References	29

Abstract

There is a rapid increase in Chronic Kidney Disease (CKD) around the world, and if abnormalities like cysts, stones and tumors in the kidney are to be successfully treated, they must be diagnosed accurately and in a timely fashion. Computed Tomography (CT) imaging is a key tool in the diagnosis of kidney disease. However, the process of manually analyzing a CT image is time-intensive and relies heavily on the expertise of a radiologist. Current deep learning model systems that are based on Convolutional Neural Networks (CNNs) show promise in automating this process, but have a stronger focus on local feature extraction and generally do not incorporate the global anatomical relationships necessary for discriminating visually similar pathologies of the kidney.

This study presents a novel hierarchical learning framework called RenalViT that employs Transformers to automate the multi-class classification of kidney CT images into four clinically relevant categories: Normal, Cyst, Stone, and Tumor. The proposed framework employs a patch-level tokenization approach, which learns positional embeddings and multiple transformer encoder blocks with multi-head self-attention to jointly learn fine-grained local pathological features and global contextual relationships. To account for class imbalance in the dataset, a clinically applicable training strategy was used that included a class-weighted loss function and domain-specific data augmentation methods [13].

Our proposed framework was further tested using the CT Kidney Dataset, which contains 12,438 images of kidneys. We compared our model against four of the most popular convolutional neural networks (CNNs): VGG16, ResNet50, DenseNet121, and EfficientNetB3, using the same experimental design. Our model achieved an accuracy of 98.96%, a weighted F1 score of 0.99, and an area under the curve AUC-ROC score of 0.997. Each of these measurements was significantly better than the baseline CNN networks. Furthermore, there was very little confusion among the kidney lesion classifications of our model. As such, we believe the CVK method will deliver a scalable and clinically-effective means of automatically classifying kidney lesions, making it a strong candidate for incorporation into computer-aided clinical decision-support systems.

1. Brief Description of the State of the Art of the Research Topic

Kidney diseases, including Chronic Kidney Disease (CKD), kidney cysts, kidney stones, and kidney tumors, have emerged as significant global health concerns due to their increasing prevalence and the risk of progressive renal failure if not diagnosed at an early stage [1], [2]. Noticeable kidney diseases/conditions are identified earlier, allowing for timely medical interventions, appropriate treatment regime and also having better outcomes for patients. Of the many different forms of medical imaging opportunities available, compute tomography (CT) is the most common due to its high definition and detailed manner of providing the Physicians information needed to assess the anatomy and pathologic processes occurring within a patient's kidney [3]. Manual readings of CT scans can be very time consuming, subjective and reliant on the radiologist experience, therefore may vary greatly with regard to diagnosis and create greater workloads on the clinical staff.

With the advances in artificial intelligence (AI) and more so deep learning technology the potential to have automated extraction of discriminating image feature sets directly from the original imaging data, therefore revolutionizing the ability to use AI and DL for medical image analysis [3]. In contrast with standard machine learning processes that are feature engineered from the data set, DL models are developed to inherently find multiple levels of features from the original imaging data making DL the preferred processing tool for many complex medical imaging applications. In addition to disease classification and lesions detection additional positive outcomes of DL models include organ segmentation and an improved computer-aided diagnostic capability across many clinical areas [3].

Convolutional Neural Networks (CNNs) are one of the most used architectures of deep learning algorithms in the classification of medical images as they are able to learn hierarchical spatial features through the use of the convolution and pooling operations [4]–[6]. As a result of this, many state-of-the-art models built upon certain CNN architectures like VGG16, ResNet50, DenseNet121 and EfficientNetB3 have been effectively implemented for the classification of kidney CT images and have produced promising results when classifying various forms of renal pathology [4]–[6]. While CNN approaches mainly use local feature extraction via a limited receptive field, long-range spatial dependency capture or global anatomical relation capture between regions of the body is also limited. Therefore, the classification of visually similar kidney

abnormalities such as cysts, stones and tumors may cause an increase in the level of confusion experienced by CNN-based models when classifying between each of these renal abnormalities.

In order to address these challenges, Vision Transformers (ViTs) have been garnering growing interest as they model all of the relationships between different parts of an image simultaneously using self-attention mechanisms [7], [8]. Instead of processing the entire image at once, as in traditional CNNs, Vision Transformers first divide the image into smaller fixed-sized patches and then treat those patches as sequential tokens. This allows for efficient learning of both local pathological features and global contextual information. Recent surveys have shown that Transformer-based architectures consistently perform better on contextual learning and provide comparable performance to CNNs on many types of medical imaging decoders [9], [10]. Therefore, ViTs have advantages in distinguishing between visually similar kidney abnormalities detected by computed tomography (CT).

The goal of this research is to develop a new automated kidney classification system using CT images called RenalViT: A hierarchical Transformer-based machine learning framework for kidney classification. To accomplish this, RenalViT will utilize Multi-Head Self-Attention mechanisms, which are capable of learning local pathological features and global contextual relationships in CT images, to classify kidney images. The performance of RenalViT will be compared to four widely used CNN models - VGG16, ResNet50, DenseNet121, and EfficientNetB3 on CT scans under identical conditions in order to provide a fair comparison. The final goal of this research is not only to increase the accuracy of classification but also to reduce confusion among normal, cysts, stones, and tumors categories as well as to create a strong platform for further development of automated computer assisted and clinical diagnosis systems [9]–[15].

1.1 Introduction

Chronic kidney disease (CKD) is a progressive disease in which there is a gradual decrease in the functioning ability of the kidneys over a period of time. This represents a worldwide epidemic and is one of the top causes of burden on the health care system across the globe [1],[2]. The primary function of the kidneys is to filter waste from the blood and regulate electrolyte levels, also to maintain the body's blood pressure and to produce hormones required to maintain normalised body functions. When kidney function is impaired, it may lead to significant renal complications in the future cardiovascular disease, kidney failure and also result in individuals developing a need for

dialysis or a kidney transplant, if not diagnosed and treated early [1]. Thus, it is important to detect kidney abnormalities early and reduce the rate of disease progression, increase the success of treatment interventions, and improve patients' quality of life.

Renal disorders have utilized the benefits provided by the detailed structural data from medical images in their diagnoses. Of the various imaging modalities available, CT scans are the most frequently used method of detection of abnormalities, cysts, stones and tumours, all of which are seen with greater spatial resolution than other imaging modalities and with excellent internal structure visualization [3]. Despite these advantages, CT scan images can only be interpreted manually and therefore require a high degree of expertise and time to interpret. Additionally, the rise in the quantity of imaging data means that additional volume of data to be processed has placed further demands on radiologists therefore, computer-assisted diagnosis systems are essential for a quicker and more accurate clinical diagnosis.

The technology known as Artificial Intelligence (AI) and its subfield Deep Learning (DL) has been revolutionary in the field of medical imaging by allowing for the ability to automatically extract image features and classify diseases based on imaging data [3]. The performance of many of these deep learning models have been proven to be fantastic in many different types of applications detecting diseases, segmenting lesions, classifying organs, and providing diagnostic decision support. One group of these deep learning models that have shown great success in kidney CT image classification are Convolutional Neural Networks (CNNs) due to the fact that they can learn how to automatically identify multi-level structure in space [4]–[6]. However, CNNs usually only focus on extracting local image features, which makes it difficult for them to identify long-range spatial relationships that will aid in classifying visually similar kidney abnormalities.

Recent studies in the field of medical imaging have shown increased interest in Vision Transformer (ViT) architectures due to their ability to address the limitations of traditional methods within this study area. These researchers theorized that by leveraging self-attention mechanisms, Vision Transformers can accurately represent a variety of indigenous pathological properties and produce accurate representations of the global context of a given image. As such, ViT architectures have proven effective at improving both the quality of feature representations and the accuracy of classification outcomes when working with CT and MRI images [7] - [10]. As a result of these advances, this project aims to develop a novel hierarchical learning framework using a

Transformer-based architecture for automated classification of renal CT images (RenalViT). The proposed framework will aim to increase the accuracy of renal classification by utilizing local to a global integration of features to provide a scalable and accurate computer-assisted method of diagnosing renal diseases.

1.2 Background: Deep Learning in Kidney Disease Classification

The advancement of artificial intelligence (AI) has had a huge impact on the advancement of intelligent healthcare systems, most notably with respect to the use of AI in medical images through image analysis. Artificial Intelligence is comprised of several different areas, Machine learning (ML), an area of A.I., is when computers use data to find patterns while Deep Learning (DL), a more specialized form of ML, uses multi-level Neural Networks to learn complex feature representations in very large data sets [3]. One of the key differences between D.L. and traditional machine learning algorithms is that traditional machine learning algorithms will extract features that have been designed by humans handcrafted. Instead, D.L. learns features that have been discriminated directly from images in real-time, leading to better classification accuracy than traditional methods as well as reducing the reliance on domain-specific feature engineering and/or design. As a consequence, D.L. has become the dominant technique for several areas of medical image analysis including: detection of disease, lesion segmentation, organ classification, and computer-aided diagnosis [3].

Convolutional Neural Networks (CNNs) are used most commonly by researchers using Deep Learning techniques for Automated Analysis of Medical Images. CNNs provide researchers with the ability to automatically learn to extract and identify spatial features using convolutional and pooling operations [4]-[6]. CNNs that have been indicated to have good performance during the identification of abnormalities in kidney CT imaging include those using models, such as VGG16, ResNet50, DenseNet121, and EfficientNetB3, with the ability to extract and identify pathological local features from renal images. Therefore, researchers have used CNNs for detecting kidney abnormalities since they have the capacity to effectively automate the extraction of features through transfer learning and through retrieval of the entire or global renal volume using feature extraction techniques. However, CNNs typically rely on small receptive fields for feature extraction, resulting in a lack of ability to retrieve or understand long-range spatial dependencies

and global anatomical relationships, which can be useful in distinguishing visually similar kidney abnormalities like cysts, stones, and tumours.

Vision Transformers (ViTs) are emerging as a solution to the limitations faced by traditional medical image classification methods. Vision Transformers (ViTs), which uses the Transformer architecture created for Natural Language Processing (NLP), represent input images as fixed-size patches and process each of these patches as "tokens," using the Multi-Head Self-Attention (MHSA) mechanism [7], [8]. While CNNs analyse an image's specific neighbourhoods, self-attention allows each patch in the image to interact directly with every other patch. Consequently, the model can learn both local pathological features and global contextual relationships in one model. These two factors contribute to a more comprehensive understanding of anatomical structures and better discrimination between disease categories, even if they look very similar.

Current research has shown that Transformer-based architectures are effective for multiple types of medical imaging tasks such as disease classification, organ segmentation, lesion detection, and decision support [9],[10]. Vision Transformers outperform traditional CNNs in terms of contextual feature representation, spatial dependency learning, and increased robustness when it comes to classifying complex images. Even though these models have demonstrated meaningful performance gains over traditional CNN models, they generally require much greater computational resources as well as larger datasets to train effectively. Due to these issues, there has been an increase in the number of researchers developing efficient Transformer-based frameworks specifically designed for medical imaging applications.

This research seeks to develop RenalViT, which aims to create an efficient and hierarchical learning framework that utilizes Transformer-based models to automate the classification process of kidney CT images. The proposed framework uses patch embedding, a sequence of Transformer encoder blocks, and Multiple-Head Self-Attention into order to learn both local pathological and global contextual data from the CT images. Additionally, the methodology employed in the study will systematically evaluate the performance of the RenalViT framework against four popular CNNs i.e., ResNet152, VGG19, DenseNet to validate the ability of the framework to accurately classify four clinically important kidney CT pathologies: normal, cyst, stone, and tumour.

1.3 Literature Survey

The use of deep learning for analysis of automated medical images has become the dominant technique for performing these tasks as it allows for hierarchical feature representation learning directly from imaging data. In the area of diagnosing kidney disease, numerous studies have utilized deep learning methods to classify abnormalities in kidney tissue present in Computed Tomography (CT) images. Much of this early work has been performed using Convolutional Neural Network (CNN) based methods, however, more recent investigations have started using Vision Transformer (ViT) architectures due to the limitations associated with CNNs for locally determining features. The purpose of this paper is to present a review of CNN and hybrid deep learning methods to classify kidney-related image features in CT scans.

CNN-Based Kidney CT Image Classification

Due to their capability of extracting spatial hierarchical features from computed tomography scanning images, Convolutional Neural Networks have found extensive implementation for the automated classification of kidney diseases. Pimpalkar et al. developed fine-tuned convolutional neural network architectures for the early diagnosis and classification of kidney disease via CT imaging. In their analysis, they demonstrated that convolutional neural networks based on transfer learning produced satisfactory classification results through efficiently learning local pathological feature information however, the ability to capture global contextual data was inadequate, limiting the discrimination of kidney abnormality examples that appeared visually similar [13]. Sharma et al. also developed a hybrid deep learning framework for classifying kidney CT images into the categories of normal, cyst, stone and tumour. Although the proposed framework improved classification accuracy, the cost associated with increased complexity and computational demands of the architecture prevented it from being practically implemented in settings with limited resources for clinical use [12].

Vision Transformer-Based Kidney CT Classification

In the past, significant improvements in medical imaging classification have been achieved using state-of-the-art methods involving vision transformers and self-attention mechanisms that allow for long range processing across multiple input sources or modalities. One example of this is the work of Islam et al., who first proposed an explainable model based on a vision transformer in combination with transfer learning for automatically detecting kidney cysts, stones, and tumours

from CT computerized tomography images. Their overall system achieved very high classification accuracy values and further supported the use of transformer based contextual learning approaches using the transformer architecture in renal image analysis. However, they did not provide any comparisons of their results against other leading-edge image classification methods, as described by state-of-the-art deep learning algorithms [11]. Similarly, the work of Subedi et al., who developed a modified vision transformer-based model for classifying kidney CT images, also improved the global feature representation and learned common spatial dependencies in kidney CT images but again they did not provide any meaningful validation tests beyond using a very limited data set and parameters from very controlled and likely artificial experimental settings [14].

Transformers in Medical Imaging

The rise in popularity of Transformer Architectures has sparked many extensive evaluations of their usefulness within the realm of medical imaging. One such evaluation, conducted by Shamshad and co-workers, was an exhaustive survey investigation into the expressed advantages of using a Transformer-based model versus traditional CNNs for tasks like Disease Classification, Segmentation, Detection, and Image Reconstruction. This survey determined that self-attention mechanisms allow for the creation of rich contextual representations through the use of long range or distance dependencies present in the images [9]. In a similar fashion, Azad and co-workers reviewed the state of the art advancements made with Vision Transformer Architectures and expressed how they utilise both local and global feature representations effectively. However, the authors also cited computationally intensive processes and excessive training time requirements as being two of the major barriers to implementing Transformer-based models in practical applications within the healthcare industry [10].

Comparative Studies of CNN and Transformer Architectures

In recent years, there has been an increase in research comparing CNN and Transformer models regarding classification of kidney disease. Hafiz et al. assessed a CNN and Vision Transformer model for classification of kidney stones. Their research found that the Transformer based models were consistently better than the CNN models in both the learning of spatial features as well as in the classification performance. However, their research investigated only the classification of kidney stones not investigating multiple classes of kidney disease [15]. Collectively, this research

shows the superiority of using Transformer models versus CNN for obtaining global context as well as better extraction of local features. However, when it comes to the use of hierarchical learning of Transformer models for classification of multiple classes of kidney disease using CT images, there is limited research.

Based on the review of this literature, there have been significant advances in the use of automated diagnosis of kidney diseases using deep learning. CNN-based models have been shown to have excellent performance in extracting local features however, Vision Transformer's ability to learn global context through a self-attention mechanism has dramatically improved. However, both CNN and Vision Transformer approaches continue to experience a set of challenges consisting of: inter-class confusion, computational complexity, limited ability to provide explainability and insufficiently comparative evaluation due to variable experimental conditions. The above observations represent a need for a robust framework using the Vision Transformer approach that will learn both the local pathological features and the global anatomical characteristics, thus, providing a motivation for the creation of the RenaViT framework for the classification of kidney CT images.

Table 1.1 Summary of Literature Review

Author(s)	Year	Method	Key Findings	Limitations
Islam et al.[11]	2022	Vision Transformer + Transfer Learning	High kidney CT classification accuracy	Limited comparative evaluation
Subedi et al.[14]	2023	Modified Vision Transformer	Improved spatial dependency learning	Limited dataset validation
Shamshad et al.[9]	2023	Transformer Survey	Superior contextual learning in medical imaging	Limited kidney-specific applications
Azad et al.[10]	2024	Vision Transformer Review	Strong global feature representation	High computational complexity

Sharma et al.[12]	2025	Hybrid Deep Learning	Improved multi-class classification	Computationally expensive
Pimpalkar et al.[13]	2025	Fine-tuned CNN Models	Effective local feature extraction	Limited global contextual learning
Hafiz et al.[15]	2025	CNN vs Vision Transformer	Transformer outperformed CNN	Limited to kidney stone classification

The table 1.1 presented compares various characteristics between CNN and transformer models outlining that CNNs utilize convolution operations where they focus on local feature extraction through a limited receptive field, making CNNs very good at extracting fine-detail spatial information. In summary, CNNs can capture local features very well but struggle with understanding of global relationships. Another way to say this is to use transformers that utilize self-attention mechanisms to understand relationships from all over an image and hence represent spatial relationships better while having a higher computational burden than CNNs.

1.4 Research Gaps

Existing literature suggests that automated identification of kidney diseases via DL-based systems has made considerable progress. Various CNN structures have proven to be effective at extracting local features and demonstrate promising results in classifying kidneys through CT images. In addition, VIT-based models have further improved global contextual learning by utilizing an attention mechanism to enhance classification performance through improved feature representation [9–15].

However, much work remains to be done in relation to unresolved research challenges. The primary shortcoming of most existing CNN-based systems is the lack of ability to learn long-range spatial dependence and global anatomical relationships present in the kidney CT image, as most systems rely solely on learning local image features via convolutions [4–6]. An increased level of confusion occurs among different classes of kidney abnormalities that may appear very similar visually (e.g., cysts, stones, tumours) as a result of this limitation. Some strategies have tried to use VIT attention mechanisms to overcome this challenge, however, relatively few studies have

actually described hierarchical learning via ViT technology for complete multi-class kidney CT image classification [11], [14].

In addition, many existing studies employ only one feature-learning method so do not effectively combine local pathological characteristics and global contextual information. Many of the reported models also have high computational complexity, inadequate comparative evaluation in identical experimental situations, and a lack of emphasis on explainability, limiting their use within clinical settings [9], [10], [12]. Therefore, these deficiencies indicate that there is a need for a more robust and efficient deep learning framework that can enhance classification accuracy, while also being computationally efficient and lessening the confusion between the classes.

Therefore, based on the limitations identified above, the current research proposes RenalViT, a Transformer-based hierarchical learning framework for classifying kidney CT images. The proposed framework integrates patch embedding, hierarchical Transformer encoder blocks and Multi-Head Self-Attention mechanisms to learn simultaneously both local pathological features and global contextual relationships. Furthermore, the proposed framework has been systematically evaluated against four well established CNN architectures using the same experimental conditions, such that a coherent comparison may be made and the performance of the proposed framework for the automatic classification of kidney disease may also be demonstrated.

2. Problem Definition

Clinical conditions like kidney disease have led to renal cyst disease, chronic renal insufficiency, diabetic nephropathy, and renal stones and tumours all require timely and accurate diagnosis so that appropriate therapeutic measures can be initiated to improve patient outcomes. Computed tomography is a common modality used to diagnose these types of renal diseases since it allows for better evaluation of anatomical structures of the kidneys as opposed to other imaging modalities. However, the major reliance upon the radiologist for the evaluation of radiology images creates a long and often subjectively interpreted process creating variability among different readers. Furthermore, the growing demand on healthcare professionals created by the increase in volume of medical images requires reliable and efficient computer-aided diagnosis tools to assist clinicians in their diagnosis [1], [3].

Recent advances in deep learning specifically convolutional neural networks allow for automated classification of CT-based renal images through the automated extraction of features present in the renal CT image from the medical image without the need for human involvement [4–6]. Although these systems have been shown to be accurate when classifying renal CT images of patients with renal abnormalities, CNNs have limitations in their focus on local feature extraction due to the use of convolutional operations, leading to problems when attempting to establish spatial relationships between renal structures that are far apart, these issues can create confusion between classes of abnormalities that appear similar kidney cysts, kidney stones and kidney tumours. In addition, publicly available datasets used to assess the performance of these models typically contain imbalanced classes that can impact the classification performance of the models and generalizability of the models to new data [9], [10], [13].

While the Ability of Vision Transformer (ViT) Architectures to Model Global Contextual Relationships Using a Self-Attention Mechanism is Superior Compared to Other Approaches, Their Use in Comprehensive Multi-Class Kidney CT Image Classification is Still Limited concerning Comprehensive Multi-Class Kidney CT Images Proves Very Limited [7]–[10]. Therefore, an efficient Transformer-based framework is required simultaneously learning local pathological features and global contextual information while effectively addressing class imbalance and reducing inter-class confusion. These limitations provide the motivation for developing the proposed RenalViT framework for automated kidney CT image classification.

3. Objective and Scope of Work

3.1 Research Objectives

This research's primary goal is to create a fully automated Deep Learning framework for Classifying Kidney CTs into categories (Normal, Cyst, Stone and Tumor) which will assist in making timely and accurate diagnoses of kidney conditions. This research will also compare multiple CNNs and Transformer-based architectures under identical experimental conditions to determine their efficacy for Classifying kidney diseases. Additionally, this research will improve overall Performance & Robustness as compared to the current state-of-the-art methods utilizing CNN-based architectures, particularly improving inter-class misclassifications of visually similar conditions using enhanced global context feature learning. Lastly, this framework will evaluate

performance using three measures, Accuracy, Precision, Recall, F1-Score, Confusion Matrix, and AUC-ROC, to provide a valid and comprehensive evaluation of the framework's performance.

3.2 Scope of Work

This research project includes an image-based (CT scan) classification of four different types of kidney disease based on publicly available medical imaging data. The project covers several topics including (i) preprocessing and data augmentation of kidney CT images, (ii) implementation and benchmarking four established CNN models (VGG16, ResNet50, DenseNet121, EfficientNetB3), (iii) design and implementation of a novel hierarchical transformer based classification framework known as “RenalViT” that uses patch-based tokenization and multi-head self-attention techniques and (iv) comprehensive comparison/ evaluation of all models against a common dataset and experimental protocols. This research project does not include real-time clinical deployment, multi-center clinical trials, or integration with Hospital Information Systems (HIS) as these are opportunities for future work.

4. Original Contribution by the Thesis

The major original contributions of this research are summarized as follows:

- **Propose a New Transformer Framework for Kidney CT Images:** This research develops RenalViT, a brand new hierarchical Vision Transformer-based algorithm for the purposes of automated classification of kidney CT images. The new architecture uses a 16×16 patch embedding, learnable positional encoding, Transformer encoder blocks and Multi-Head Self Attention (MHSA) to learn local pathological characteristics and global contextual relationships from kidney CT images.
- **Use Robust Training Strategy to Classify Kidney CT Images:** A comprehensive preprocessing pipeline has been constructed to provide image normalization, domain-specific data augmentation and class-weighted loss functions to provide better feature representation of the images and to more effectively handle the class imbalances in the CT Kidney dataset.
- **Conduct Comprehensive Benchmark Evaluation of Proposed Framework:** The proposed RenalViT framework has been systematically and comparably evaluated against four common CNN architectures - VGG16, ResNet50, DenseNet121, and EfficientNetB3 using

identical dataset, preprocessing pipeline, training configurations and evaluation methods to ensure an equitable and reproducible benchmark comparison.

- Better categorization effectiveness: The results of the experiment indicate that the suggested methodology has resulted in an overall classification accuracy of 98.96%. A weighted F1 score of 0.99 and an AUC-ROC of 0.997 are also produced. The establishment of the suggested model has been verified through extensive validation using various techniques and by successfully reducing inter-class mixing between the normal kidney category, the cyst category, the stone category and the tumors category.
- Validation of architecture: In order to verify the use of the GELU activation function within the Transformer encoder blocks, a systematic evaluation of the ReLU activation function (baseline) in comparison to the GELU activation function proposed was conducted. The results show that having the GELU activation function leads to a greater level of feature learning and classification accuracy during the analysis of CT scan images of the kidney.

5. Methodology, Results and Discussion

5.1 Dataset Description

The CT Kidney Dataset is a set of labelled images from a CT scan that show how kidneys are affected by four illnesses: Normal, Cyst, Stone and Tumour. This dataset was chosen for its diversity, and clinical relevance, making it an ideal benchmark for deep learning algorithms for classifying kidney CT images.

Specifically, all images were resized to 224×224 pixels and then split into three datasets Training:Validation:Testing - 80:10:10. The CT Kidney Dataset has 12,438 CT images, comprising 5,077 Normal, 3,701 Cyst, 2,283 Stone, and 1,377 Tumour images. Due to the difference in size between the four categories, there is a class imbalance problem in the CT kidney dataset, namely with regards to the Tumour category. To mitigate this issue, the use of a class weighted loss function was used during model training. Table 5.1 provides a summary of the characteristics of the CT kidney dataset.

Table 5.1: CT Kidney Dataset Description

Parameter	Details
Dataset Name	CT Kidney Dataset
Dataset Type	Medical CT Image Dataset
Imaging Modality	Computed Tomography (CT) Images
Total Images	12,438
Number of Classes	4 (Normal, Cyst, Stone, Tumor)
Image Format	PNG / JPEG
Image Size	224 × 224 pixels
Dataset Split	80% Training, 10% Validation, 10% Testing
Source	Kaggle CT Kidney Dataset [17]

The dataset is not equally represented between the 4 classes as shown Normal (5,077 images), Cyst (3,701 images), Stone (2,283 images), Tumour (1,377 images). To correct this class imbalance, a weighted loss function was applied to the training methodology.

5.2 Preprocessing and Data Augmentation

To achieve model development with consistent training, we preprocessed all CT images before using them for model building. Each image was resized to 224 × 224 pixels and pre-processed using normalization for pixel intensity values so that they would have a range from [0–1]. This was done to increase the stability of our training and speed up convergence of our model as well [3].

In order to enhance the quality of the CT images and increase the overall robustness of our proposed framework, we applied appropriate preprocessing techniques that were employed to reduce image noise and improve the visibility of pathological features. Additionally, we utilized domain specific data augmentation techniques to create new examples from our available examples to add to the diversity of data in the CT Kidney Dataset and minimize the effects of overfitting on our model. Examples of the augmentation techniques we used include random

rotations ($\pm 20^\circ$), horizontal flip, vertical flip, random zooming (up to 10%), random contrast change. Through these augmentation transformations we enabled our model to learn robust, invariant, and generalized feature representations while also providing solutions to the class imbalance of our CT Kidney dataset [13].

5.3 Baseline (CNN) Architectures

Four well-known Convolutional Neural Network (CNN) architectures will be used as baseline models for the comparison of the proposed RenalViT framework's effectiveness.

- VGG16 is an example of a pure deep sequential architecture and uses several 3×3 convolutional filters to perform hierarchical representations for input images. By being designed simply, this architecture can effectively and reliably extract low and intermediate-level visual features. [4].
- ResNet50 enables the training of deep networks by employing residual learning through skip (or shortcut) connections to mitigate the shrinking gradient problem that occurs when training very deep architectures. This architecture is well known to have produced state-of-the-art results in large-scale image classifications tasks. [4].
- DenseNet121 has been designed to create a dense connectivity to each previous layer's feature maps. This improves the propagation of gradients to produce better composite models by allowing reuse of features and eliminating unnecessary computations during training. [5].
- EfficientNetB3 uses a compound scaling approach to optimize depth, width, and input resolution. Thus, this architecture produces an appropriate balance between performance and economy when classifying images. [6].

Although these CNN architectures have demonstrated strong performance in medical image classification, they primarily focus on extracting local spatial features through convolution operations and have limited capability to model long-range contextual relationships across the entire CT image. Consequently, they may experience reduced performance when distinguishing visually similar kidney abnormalities, thereby motivating the development of the proposed RenalViT framework [9], [10].

5.4 Proposed Methodology: RenalViT

The proposed RenalViT framework is a hierarchical Vision Transformer-based architecture to overcome the limitations of local feature extraction in CNN-based systems, and is thus the focus of this research. The proposed framework consists of the following components:

- **Patch-level tokenization:** Each input CT image is divided into fixed-size 16×16 patches, which are linearly projected into token embeddings.
- **Positional encoding:** Learnable positional embeddings are added to the patch tokens to preserve spatial information that would otherwise be lost during tokenization.
- **Transformer encoding:** Four transformer encoder blocks, each employing four-head multi-head self-attention, process the token sequence to learn both fine-grained local pathological features and long-range global anatomical context.
- **Activation function:** The GELU (Gaussian Error Linear Unit) activation function was adopted in place of ReLU within the transformer encoder blocks, since GELU is a smooth, probabilistic activation that preserves useful negative information and provides better gradient flow properties that are particularly well suited to transformer-based architectures, at a modest cost in computational complexity relative to ReLU.
- **Class-imbalance handling:** An inverse-frequency class-weighted loss function was used during training, together with the domain-specific augmentation strategy described above, to robustly address the substantial class imbalance between the Normal, Cyst, Stone, and Tumor categories.
- **Classification head:** A dense classification head maps the learned representations to the four target classes - Normal, Cyst, Stone, Tumor.

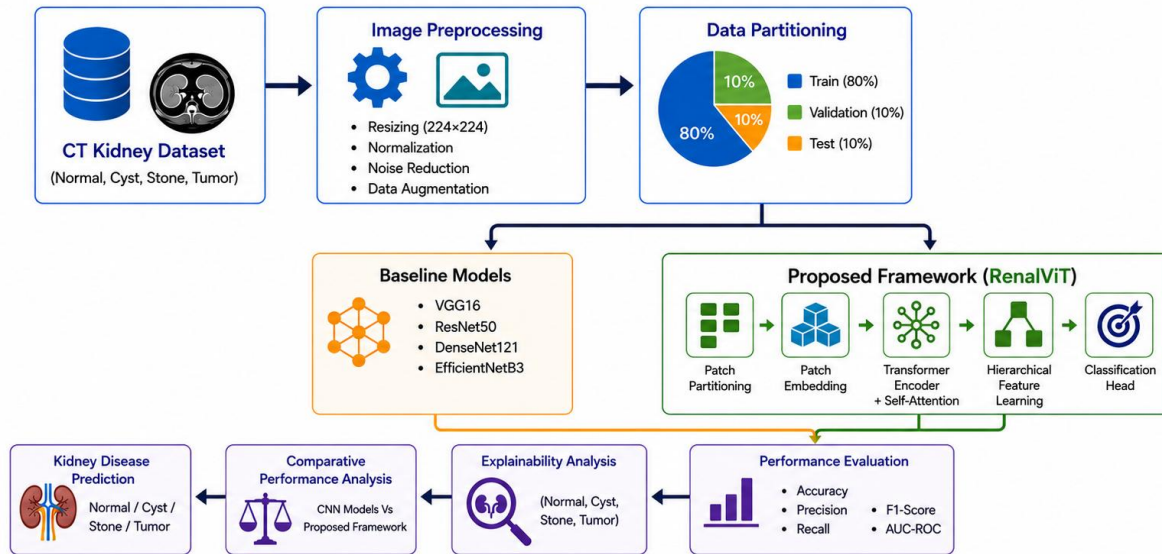


Figure 5.1: Work flow diagram for the proposed model

5.5 Experimental Setup

The experiments were carried out using a standardized deep learning configuration to allow for consistent results to be evaluated fairly and reproduced.

- **Platform:** Google Colab
- **GPU:** NVIDIA Tesla T4
- **Framework:** TensorFlow / Keras
- **Operating System:** Windows
- **RAM:** 16 GB
- **Batch Size:** 32
- **Epochs:** 50 to 65
- **Optimizer:** AdamW
- **Learning Rate:** 1×10^{-4}
- **Image Size:** 224×224 pixels

5.6 Comparative Results

The consolidated Table 5.2 shows a comparison of all of the model architectures and performance in accordance with the DPC review comment requesting a comparison of multiple algorithms with parameter settings.

Table 5.2: Consolidated Model Comparison with Architectural Parameters and Performance

Model	Layers	Parameters	Activation	Optimizer	Learning Rate	Accuracy (%)	F1-Score	Key Advantages
VGG16	16	138M	ReLU	Adam	1×10^{-4}	66.43	0.67	Strong low-level feature extraction
ResNet50	50	25M	ReLU	Adam	1×10^{-4}	72.76	0.73	Skip connections improve gradient flow
DenseNet-121	121	8M	ReLU	Adam	1×10^{-4}	67.07	0.66	Efficient feature reuse
EfficientNetB3	315	12M	Swish/ReLU	Adam	1×10^{-4}	75.16	0.76	Optimized accuracy-to-computation ratio
Proposed (RenalViT)	4 Encoder Blocks	22M	GELU	AdamW	1×10^{-4}	98.96	0.99	Global spatial dependencies

								es via self-attention
--	--	--	--	--	--	--	--	------------------------------

In evaluating the various deep learning models for classifying kidney CT images, extensive comparative evaluations were performed on the various models including the popular CNN architectures VGG16, ResNet50, DenseNet121 and EfficientNetB3, as well as the proposed Transformer-based model, RenalViT. Each of the models was evaluated under identical conditions to ensure a fair comparison of the results achieved by using standard classification metrics of accuracy, precision, recall, and F1 score.

The detailed results presented clearly support the proposed approach as compared to existing baseline models. The complete quantitative comparison of performance results for all of the evaluated models is also shown in Table 5.3.

Table 5.3: Comparative Performance Analysis of All Evaluated Models

Model	Accuracy (%)	Precision	Recall	F1-Score
VGG16	66.43	66.61	66.43	66.24
ResNet50	72.76	73.39	72.76	72.89
DenseNet121	67.07	67.52	67.07	66.94
EfficientNetB3	75.16	76.22	75.16	75.68
Proposed (RenalViT)	98.96	99.02	98.96	98.99

RenalViT received a highest classification accuracy of 98.96% & 0.99 weighted F1 Score of all architecture evaluated showing large margins of performance with EfficientNetB3 being the strongest performing CNN baseline model at 75.16% accuracy, with RenalViT outperforming it by over 23 percentage points. Additionally, RenalViT was able to achieve an area under the ROC curve (AUC-ROC) score of 0.997 indicating excellent sensitivity, specificity, and class separation along with an inference throughput on approximately 243 images per second, thus demonstrating its suitability for practical medical use in a clinical environment.



Figure 5.2: Accuracy Comparison of All Evaluated Deep Learning Models

Comparative accuracy performances of VGG16, ResNet50, DenseNet121, EfficientNetB3, and proposed RenalViT are illustrated in Figure 5.1 with each architecture being demonstrated under identical experimental conditions with their classification accuracies varying greatly as a function of architecture. It can be concluded from the graphical comparison of performance across these different architectures that the RenalViT model significantly outperformed all baseline models and obtained the highest level of accuracy overall compared to other benchmark deep learning approaches.

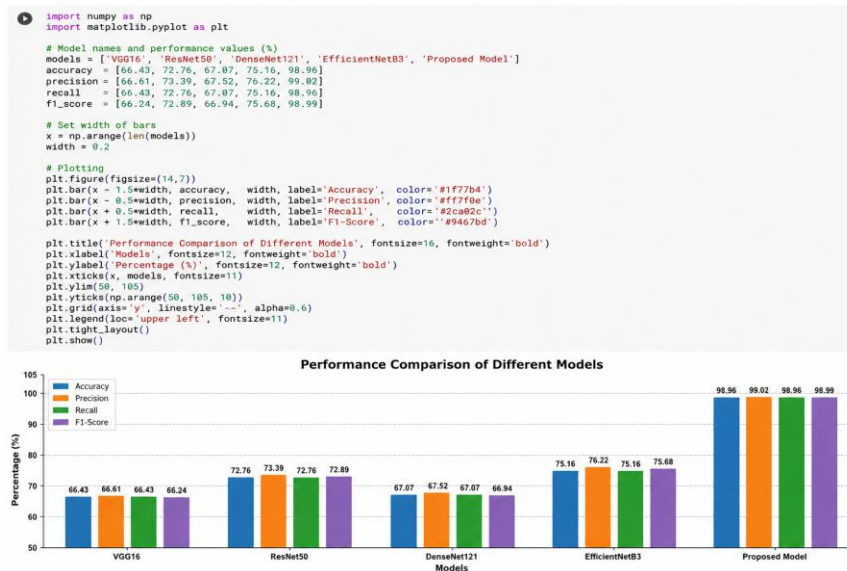


Figure 5.3: Performance Comparison Chart Across All Metrics

A performance comparison analysis of all studied deep learning models is seen across various evaluation metrics including accuracy, precision, recall, F1 scores as represented in Figure 5.2. The efficiency of the proposed RenalViT model vs. the efficiency of the baseline traditional CNN models (VGG16, ResNet50, DenseNet121 and EfficientNetB3) under identical experimental setups is displayed - clearly demonstrating that the performance of the proposed model is greater than each baseline approach across all metrics examined (classification ability, quality of feature representation and overall robustness). As represented in Figure 2, the results of this comparison demonstrate that the proposed framework has been successful in producing good kidney disease classification results that are balanced in performance.

5.7 Confusion Matrix Analysis

The method used to evaluate classification performance was via the confusion matrix which compares the actual and predicted class labels, including the true positives, true negatives, false positives, and false negatives, for each of the four categories associated with kidney diseases. The proposed model consistently had higher values along the diagonal of the confusion matrix indicating that most of the CT images are classified into their appropriate categories (Normal, Cyst, Stone and Tumor) with few misclassified images found near the off-diagonal. This directly addresses the DPC review comment that requested more clarity on how to interpret the confusion matrix and supports the conclusion that the proposed transformer-based self-attention approach has significantly less confusion between classes compared to the baseline CNN classifiers.

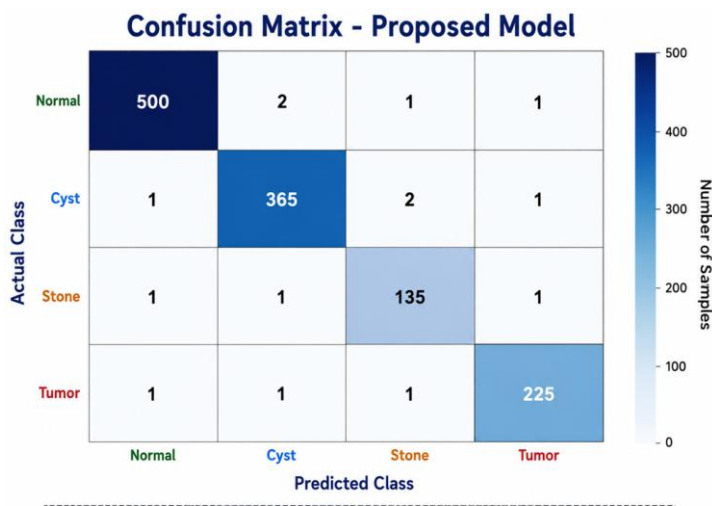


Figure 5.4: Confusion Matrix of the Proposed RenalViT Model

As illustrated in Figure 5.3, using the RenalViT model provides a clear picture of the classification performance of the proposed model across all four classes of kidney disease Normal, Cyst, Stone and Tumor. The values shown on the diagonal indicate the number of correctly classified images for each class with a large number of images correctly classified for each class, while the values shown off-diagonal are significantly lower indicating that few images were misclassified. From the results of this evaluation, the proposed model achieved a high level of performance through all four class models with excellent balance across the four class models and minimal classification errors.

5.8 ROC Curve and AUC Analysis

The Receiver Operating Characteristic curve and Area Under the Curve score were used to quantify how well the proposed deep learning framework classifies kidney disease into the four categories. The proposed approach produces an AUC-ROC score of 0.997, showing a high level of sensitivity and specificity for distinguishing between all four categories of kidney disease.

The near-perfect AUC demonstrates strong separability between classes and lends support to the transformer based self-attention mechanism's ability to distinguish visually similar kidney conditions with minimal overlap between their class distributions.

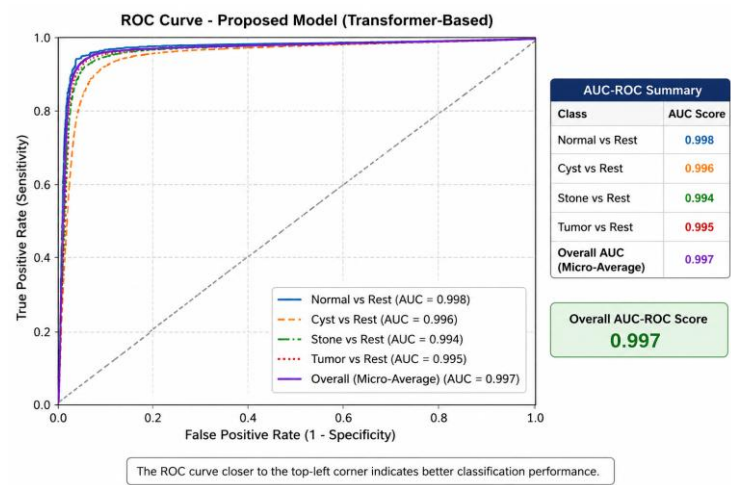


Figure 5.5: ROC Curve and AUC Analysis of the Proposed Model

Figure 5.4 shows the AUC and ROC curves of the proposed RenalViT model by each of the four categories of kidney disease: normal, cyst, stone, and tumor. The curves are all located near the top-left corner of the plot, meaning that all four categories have relatively high true positive rates

and low false positive rates, suggesting strong classifier performance across all four categories. The AUC score for each of the four categories demonstrates that the RenalViT model is able to distinguish between all four categories of kidney disease at a very high level therefore, the model is capable of robustly and reliably distinguishing between kidney disease categories as illustrated in Figure 4.

6. Achievements with respect to Objectives

This study has been mapped to the stated study objectives and presents the following results as an overview of the study's results:

- The first objective was to develop an automated, deep-learning framework to classify kidney CT scans this was accomplished through the design, development, and validation of a new deep learning framework called RenalViT to classify 4 classes of CT scan images: Normal, Cyst, Stone, and Tumor.
- The second objective was to evaluate CNN models against compare with Transformer models. This was done by benchmarking 4 standard CNN models (VGG16, ResNet50, DenseNet121, and EfficientNetB3), using the same pre-processing, training, and evaluation protocols and providing a basis for assessing the performance of the proposed Transformer-based model, using a single pooled dataset of 12,438 CT images.
- The third objective provided a basis for improving both classification accuracy and classification robustness. The proposed RenalViT achieved a classification accuracy of 98.96% with a weighted F1 score of 0.99, which is a significant improvement over the highest-performing CNN baseline, EfficientNetB3, which achieved a classification accuracy of 75.16%.
- The fourth objective required reducing the amount of misclassification between classes in the Confusion Matrix analysis of inter-class misclassifications, which were negligible in each case. This indicates that the self-attention mechanism is effective in improving global context learning and effectively reducing confusion between visually similar classes for kidney disease, such as cyst, stone, and tumor.
- The last goal to pursue your final evaluation of a model required the use of several different measurements of model performance with the aim of comprehensive testing. All were successfully met via the use of many different methods to assess accuracy, precision and

recall as well as data analysis on model prediction accuracy using F1 score, confusion matrix, and AUC-ROC curve (AUC 0.997). Thus validating your model through testing, via multiple dimensions.

Additionally, during this research there were 3 published articles consisting of 3 peer-reviewed journal/conference articles, and 1 more article is currently under review, in reputable peer-reviewed journal, thus contributing to the broader research community and disseminating of our research.

7. Conclusion

This research shows that using deep learning in a transformer-based hierarchical structure for the automated classification of renal (kidney) computed tomography (CT) images significantly outperformed standard convolutional neural networks (CNNs) as the sole means of classifying renal CT images. The RenalViT framework used a patch-based tokenization and multi-head self-attention mechanism to capture the local pathological details and global anatomical relationships necessary to differentiate between potentially visually similar renal conditions (e.g., cysts, stones, and tumors).

The proposed model performed systematic benchmark evaluations against VGG16, ResNet50, DenseNet121, and EfficientNetB3 using a common dataset of 12,438 CT images to achieve a classification accuracy of 98.96%, a weighted F1 score of 0.99, and an area under the curve for receiver operating characteristic output (AUC-ROC) score of 0.997 minimal inter-class confusion and a high inference throughput. These results indicate that the proposed framework is a scalable, accurate, and clinically applicable solution, and therefore has great potential to be incorporated into computer-aided diagnosis and clinical decision-support systems to allow for earlier and more accurate identification of renal abnormalities.

The framework being proposed allows for early and accurate identification of any kidney abnormalities using CT images, will lessen the workload of diagnosing and analyzing images from a professional radiologist and will speed up a radiologist or healthcare professional's ability to make a diagnosis and create a plan of treatment. These capabilities will lay a strong groundwork for the development of future AI-based computer-aided diagnosis systems.

The limitations of this work include the need for a significant amount of processing resources to run transformer-based models in addition to a long time to train the model and the fact that we have only validated this model on one publicly available dataset in clinical practice. Also, interpretability will continue to be a broader challenge for medical diagnosis using deep learning. Future work will focus on incorporating Explainable AI (XAI) approaches for interpretability and trustworthiness in the clinical setting, validating the framework on large-scale, multi-center datasets, developing lightweight versions for real-time use in clinical settings, and expanding the framework to encompass other forms of medical imaging.

8. Copies of Papers Published and List of All Publications Arising from the Thesis

The research work carried out during this thesis has resulted in three publications, summarized in the table 8.1.

Table 8.1: List of Publications Arising from the Thesis

Sr. No.	Title of the Paper	Journal / Conference	Status
1	Assessing Chronic Kidney Disease Prediction Models: A Comparative Analysis of Smart AI and Intelligent Machine Learning Approaches	2nd International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), 2024 (IEEE)	Published
2	Kidney Disease Prediction Using Machine Learning	IAPQR Transactions, Vol. 49, Issue 02, 2024	Published
3	KTransNet: A Hybrid Transformer-CNN Framework for Early-Stage Chronic Kidney Disease Detection Using Multi-Modal Renal Imaging	International Journal of Innovative Science and Research Technology (IJISRT), Vol. 11, Issue 5, 2026	Published

References

- [1] Chittora P, Chaurasia S, Chakrabarti P, Kumar V, Ghosh A, Khandelwal N. Prediction of chronic kidney disease: a machine learning perspective. *IEEE Access*. 2021;9:17312-17334.
- [2] Arulanthu P, Perumal E. Predicting chronic kidney disease using various classifiers. In: *Proceedings of the 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*; 2019. p. 70-75.
- [3] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60-88.
- [4] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016. p. 770-778.
- [5] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2017. p. 4700-4708.
- [6] Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*; 2019. p. 6105-6114.
- [7] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems (NeurIPS)*; 2017. p. 5998-6008.
- [8] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16×16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)*; 2021.
- [9] Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, et al. Transformers in medical imaging: a survey. *Med Image Anal*. 2023;88:102802.
- [10] Azad R, Kazerouni A, Heidari M, Rajebi H, Aghdam EK, Molaei S, et al. Advances in medical image analysis with vision transformers: a comprehensive review. *Med Image Anal*. 2024;91:103000.
- [11] Islam MN, Hasan M, Hossain MK, et al. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. *Sci Rep*. 2022;12(1):11440.

- [12] Sharma K, Uddin Z, Wadal A, Gupta D. Hybrid deep learning framework for classification of kidney CT images: diagnosis of stones, cysts, and tumors. arXiv. 2025;2502.04367.
- [13] Pimpalkar A, Saini DKJB, Shelke N, Balodi A, Rapate G, Tolani M. Fine-tuned deep learning models for early detection and classification of kidney conditions in CT imaging. Sci Rep. 2025;15(1):10741.
- [14] Subedi R, Timilsina S, Adhikari S. Kidney CT scan image classification using modified vision transformer. J Eng Sci. 2023;2(1):24-29.
- [15] Hafiz AA, Vericho D, Carter VJ, Thio DC, Isnan M, Pardamean B. Vision transformer and CNNs in kidney stone classification: a comparative study. Procedia Comput Sci. 2025;269:1466-1473.
- [16] Wu D, Xu J, Ding B, et al. A generative vision transformer model for kidney tumor classification in computed tomography images. Alex Eng J. 2025;128:259-267.
- [17] Nazmul0087. CT Kidney Dataset: Normal, Cyst, Tumor and Stone [Internet]. Kaggle; 2022. Available from: <https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone>. Accessed 2026 Jul 2.