

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

A Thesis submitted to Gujarat Technological University
for the Award of

Doctor of Philosophy

in

Computer/IT Engineering

by

Thacker Smit Mahendrabhai

Enrollment No.: 219999913010

under supervision of

Dr. Ravirajsinh S Vaghela



**GUJARAT TECHNOLOGICAL UNIVERSITY
AHMEDABAD**

October 2026

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

A Thesis submitted to Gujarat Technological University
for the Award of

Doctor of Philosophy

in

Computer/IT Engineering

by

Thacker Smit Mahendrabhai

Enrollment No.: 219999913010

under supervision of

Dr. Ravirajsinh S Vaghela



**GUJARAT TECHNOLOGICAL UNIVERSITY
AHMEDABAD**

October 2026

© Thacker Smit Mahendrabhai

DECLARATION

I declare that the thesis entitled **Deep Learning-Based Framework for Early Detection and Classification of CKD using Medical Imaging** submitted by me for the degree of Doctor of Philosophy is the record of research work carried out by me during the period from **September 2021 to July 2026** under the supervision of **Dr. Ravirajsinh S Vaghela** and this has not formed the basis for the award of any degree, diploma, associateship, fellowship, titles in this or any other University or other institution of higher learning.

I further declare that the material obtained from other sources has been duly acknowledged in the thesis. I shall be solely responsible for any plagiarism or other irregularities, if noticed in the thesis.

Signature of the Research Scholar:

Date : 15/10/2026

Name of Research Scholar: **Thacker Smit Mahendrabhai**

Place: **Ahmedabad**

CERTIFICATE

I certify that the work incorporated in the thesis **Deep Learning-Based Framework for Early Detection and Classification of CKD using Medical Imaging** submitted by Shri **Thacker Smit Mahendrabhai** was carried out by the candidate under my supervision/guidance. To the best of my knowledge: (i) the candidate has not submitted the same research work to any other institution for any degree/diploma, Associateship, Fellowship or other similar titles (ii) the thesis submitted is a record of original research work done by the Research Scholar during the period of study under my supervision, and (iii) the thesis represents independent research work on the part of the Research Scholar.

Signature of Supervisor:



Date: 15/10/2024

Name of Supervisor: **Dr. Ravirajsinh S Vaghela**

Place: **Ahmedabad**

Course-work Completion Certificate

This is to certify that **Mr. Thacker Smit Mahendrabhai** enrolment no. **219999913010** is a PhD scholar enrolled for PhD program in the branch **Computer/IT Engineering** of Gujarat Technological University, Ahmedabad.

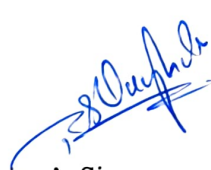
(Please tick the relevant option(s))

☐ He/She has been exempted from the course-work (successfully completed during M.Phil Course)

☐ He/She has been exempted from Research Methodology Course only (successfully completed during M.Phil Course)

☒ He/She has successfully completed the PhD course work for the partial requirement for the award of PhD Degree. His/ Her performance in the course work is as follows

Grade Obtained in Research Methodology (PHD2001)	Grade Obtained in Research and Publication Ethics (PHD2002)	Grade Obtained in Self Study Course (PH2003)
BC	BB	AB


Supervisor's Sign

Dr. Ravirajsinh S Vaghela

Originality Report Certificate

It is certified that PhD Thesis titled “**Deep Learning-Based Framework for Early Detection and Classification of CKD using Medical Imaging**” by **Thacker Smit Mahendrabhai** has been examined by us. We undertake the following:

- a. Thesis has significant new work / knowledge as compared already published or are under consideration to be published elsewhere. No sentence, equation, diagram, table, paragraph or section has been copied verbatim from previous work unless it is placed under quotation marks and duly referenced.
- b. The work presented is original and own work of the author (i.e. there is no plagiarism). No ideas, processes, results or words of others have been presented as Author own work.
- c. There is no fabrication of data or results which have been compiled / analysed.
- d. There is no falsification by manipulating research materials, equipment or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.
- e. The thesis has been checked using **Turnitin** and found within limits as per GTU Plagiarism Policy and instructions issued from time to time (i.e. permitted similarity index <10%).

Signature of the Research Scholar:

Date: 15/10/2026

Name of Research Scholar: **Thacker Smit Mahendrabhai**

Signature of Supervisor:

Date: 15/10/2026

Name of Supervisor: **Dr. Ravirajsinh S Vaghela**

Place: **Ahmedabad**

Plagiarism Report

Smit Thacker Diploma

219999913010_Smit_Thacker_PhD_Thesis

 Research Articles

Document Details

Submission ID	106 Pages
trnoid=3618:146421718	
Submission Date	29,913 Words
Jul 25, 2026, 8:43 AM GMT+5:30	
Download Date	174,438 Characters
Jul 25, 2026, 8:52 AM GMT+5:30	
File Name	
219999913010_Smit_Thacker_PhD_Thesis.pdf	
File Size	
7.6 MB	

3% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 14 words)

Match Groups

- 46 Not Cited or Quoted 3%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 2% Internet sources
- 1% Publications
- 2% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Ph.D. Thesis Non-Exclusive License to GUJARAT TECHNOLOGICAL UNIVERSITY

In consideration of being a PhD Research Scholar at GTU and in the interests of the facilitation of research at GTU and elsewhere, I, **Mr. Smit Mahendrabhai Thacker** having **219999913010** hereby grant a non-exclusive, royalty free and perpetual license to GTU on the following terms:

- a) GTU is permitted to archive, reproduce and distribute my thesis, in whole or in part, and/or my abstract, in whole or in part (referred to collectively as the “Work”) anywhere in the world, for non-commercial purposes, in all forms of media;
- b) GTU is permitted to authorize, sub-lease, sub-contract or procure any of the acts mentioned in paragraph (a);
- c) GTU is authorized to submit the Work at any National / International Library, under the authority of their “Thesis Non-Exclusive License”;
- d) The Universal Copyright Notice (©) shall appear on all copies made under the authority of this license;
- e) I undertake to submit my thesis, through my University, to any Library and Archives. Any abstract submitted with the thesis will be considered to form part of the thesis.
- f) I represent that my thesis is my original work, does not infringe any rights of others, including privacy rights, and that I have the right to make the grant conferred by this non-exclusive license.
- g) If third party copyrighted material was included in my thesis for which, under the terms of the Copyright Act, written permission from the copyright owners is required, I have obtained such permission from the copyright owners to do the acts mentioned in paragraph (a) above for the full term of copyright protection.

- h) I retain copyright ownership and moral rights in my thesis, and may deal with the copyright in my thesis, in any way consistent with rights granted by me to my University in this non-exclusive license.
- i) I further promise to inform any person to whom I may hereafter assign or license my copyright in my thesis of the rights granted by me to my University in this nonexclusive license.
- j) I am aware of and agree to accept the conditions and regulations of PhD including all policy matters related to authorship and plagiarism.
- k.) The following statement shall be included appropriately and displayed prominently in the book or any material being published anywhere: “The content of the published work is part of the thesis submitted in partial fulfilment for the award of the degree of Ph.D. in Computer/IT Engineering of the Gujarat Technological University”.
- l.) I further promise to inform any person to whom I may hereafter assign or license my copyright in my thesis of the rights granted by me to my University in this nonexclusive license. I shall keep GTU indemnified from any and all claims from the Publisher(s) or any third parties at all times resulting or arising from the publishing or use or intended use of the book / such similar document or its contents.
- m.) I am aware of and agree to accept the conditions and regulations of Ph.D. including all policy matters related to authorship and plagiarism.

Signature of the Research Scholar:

Name of Research Scholar: **Thacker Smit Mahendrabhai**

Date: 15/10/2026

Place: **Ahmedabad**

Signature of Supervisor:

Name of Supervisor: **Dr. Ravirajsinh S Vaghela**

Date: 15/10/2026

Place: **Ahmedabad**

Seal:

Thesis Approval Form

The viva-voce of the PhD Thesis submitted by Shri. Thacker Smit Mahendrabhai (219999913010) entitled **Deep Learning-Based Framework for Early Detection and Classification of CKD using Medical Imaging** was conducted on 15.10.2026 (day and date) at Gujarat Technological University.

(Please tick any one of the following option)

☒

The performance of the candidate was satisfactory. We recommend that he/she be awarded the PhD degree.

☐

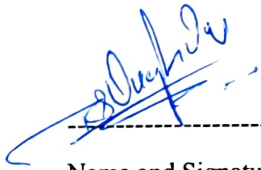
Any further modifications in research work recommended by the panel after 3 months from the date of first viva-voce upon request of the Supervisor or request of Independent Research Scholar after which viva-voce can be re-conducted by the same panel again.

(briefly specify the modifications suggested by the panel)

☐

The performance of the candidate was unsatisfactory. We recommend that he/she should not be awarded the PhD degree.

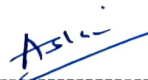
(The panel must give justifications for rejecting the research work)



Name and Signature of Supervisor with Seal
Dr. Ravirajsinh S Vaghela



1) (External Examiner 1) Name and Signature
Dr. Rituraj Jain



2) (External Examiner 2) Name and Signature
Dr. Ashish Sharma

ABSTRACT

A worldwide public health crisis that is expanding, kidney illnesses, such as cysts, stones, and tumors, are a large burden. Chronic Kidney Disease (CKD) is impacting 10% of the world population and leads to just over 1.2 million deaths each year. Accurate, automated identification of kidney illnesses through the use of Computed Tomography (CT) images will contribute greatly to meeting the unreached priorities of healthcare professionals and provide an efficient input for earlier detection of illness than current manual interpretation processes can provide.

As a solution, this research presents a new framework of artificial intelligence, RenalViT, that is hierarchical and utilizes transformer-based deep learning to produce algorithms that automatically classify CT images of the kidney into four clinically relevant categories: Normal, Cyst, Stone, or Tumor. RenalViT uses multiple techniques including 16 x 16 patch tokenization, learnable positional embeddings, four transformer encoder blocks with 4 multi-head self-attention mechanisms, GELU activation functions in the 3 MLP sub-layers of the transformer encoder block, and an inverse frequency class weighting loss function, to mitigate class imbalance in the dataset used.

To ensure the strength of the results produced by RenalViT a systematic benchmark against 4 widely used convolutional neural network (CNN) architectures: VGG16, ResNet50, DenseNet121 and EfficientNetB3 was completed under the exact same experimental setting parameters on a publicly available CT Kidney Image Dataset containing 12,438 CT images. All models were trained using the identical pre-processing, augmentation, hyper-parameter configuration and evaluation protocols to allow for equitable and systematic automated comparison. The results of this study demonstrate that RenalViT has performed superiorly in effectively classifying images when benchmarked against the aforementioned architectures. As a result, RenalViT represents a powerful tool that can be used to assist clinicians with the identification of kidney ailments at a higher level than they are currently able to achieve.

The results of the experiment indicate an important outcome for the RenalViT framework in comparison with CNN baselines as follows: classification accuracy = 98.96%, weighted precision = 99.02%, weighted recall = 98.96%, weighted F1-score = 98.99%, AUC-ROC=0.997. When compared to the best performing CNN baseline (EfficientNetB3: 75.16% accuracy), these results equate to significant increases of 23.80 percentage points

(pp) accuracy, 22.80 pp precision and 23.31 pp F1 Score. The confusion matrix suggests that the classification of each disease category is near-perfect with all false negative rates $< 2\%$, which includes the clinically most important Tumor class.

The RenalViT framework has made five main contributions: (i) a purpose-designed hierarchical Transformer for Kidney CT classification, (ii) a clinical rationale behind the use of a class-weighted loss function, (iii) a valid systematic comparison of CNN versus Transformer architecture under identical hardware, (iv) an exhaustive evaluation of architectural components including activation functions and optimizers and (v) five peer-reviewed publications including a mix of journal and conference papers.

Acknowledgements

Pursuing a Ph.D. has been an incredibly enriching and formative experience. Throughout this journey, I have been fortunate to receive the guidance, support, and encouragement of many individuals, to whom I owe a deep debt of gratitude.

First and foremost, I offer my deepest gratitude to God for granting me the strength, patience, and perseverance to complete this journey. Without His divine guidance and endless grace, none of this would have been possible.

At the very beginning, I extend my deepest gratitude to my supervisor and mentor, **Dr. Ravirajsinh Vaghela**, for his invaluable guidance, constant encouragement, and unwavering support throughout the course of this research. His scholarly insight and constructive feedback have been instrumental in shaping this thesis.

I also sincerely thank my DPC members for their continuous evaluation and critical feedback that helped strengthen the quality and rigour of this research work.

I extend my heartfelt gratitude to the research communities that openly shared datasets and tools, enabling the reproducibility and advancement of the research presented in this thesis. I am particularly grateful to the creators of the CT Kidney Dataset and the open-source deep learning frameworks that made this work possible.

I am profoundly grateful to my beloved parents for their unconditional blessings, sacrifices, and constant encouragement throughout my life. I also extend my deepest appreciation to my wife, **Grishma**, and my beloved daughters, **Prisha** and **Rivya**, whose love, patience, understanding, and unwavering support gave me the strength to overcome every challenge. Their faith in me often exceeded my own, inspiring me to strive for excellence and persevere through difficult times.

Balancing professional responsibilities with doctoral research was a challenging endeavour, and I am thankful to my colleagues and friends whose support and dedication inspired me to stay focused and committed throughout this journey.

Finally, I express sincere appreciation to **Gujarat Technological University, Ahmedabad**, for providing the academic framework and resources necessary to pursue this doctoral research.

Table of Contents

Cover Page	
Title Page	
Copyright Page	ii
Declaration	iii
Certificate by Research Supervisor	Iv
Course-work Completion Certificate	v
Originality Report Certificate	vi
Plagiarism Report	vii
Non-Exclusive License to GTU	ix
Thesis Approval Form	xi
Abstract	xii
Acknowledgement	xiv
Table of Contents	xv
List of Figures	xix
List of Tables	xx
List of Abbreviations	xxi
List of Symbols	xxiii
CHAPTER 1 - Introduction	1
1.1 Background	1
1.2 Kidney Anatomy	2
1.3 Chronic Kidney Disease	4
1.4 Kidney Cyst	4
1.5 Kidney Stone	5
1.6 Kidney Tumor	5
1.7 Medical Imaging Techniques	6
1.8 Artificial Intelligence in Healthcare	7
1.9 Motivation	8
1.10 Problem Statement	9

1.11 Research Objectives	9
1.12 Thesis Outline	10
1.13 Summary	10
CHAPTER 2 - Literature Review	12
2.1 Convolutional Neural Networks for Medical Imaging	13
2.2 Machine Learning in Medical Imaging	14
2.3 Deep Learning Models for Image Classification	15
2.4 CNN Architectures	16
2.4.1 VGG16	17
2.4.2 ResNet50	18
2.4.3 DenseNet121	19
2.4.4 EfficientNetB3	20
2.5 Vision Transformer	22
2.6 Transformer Architectures in Healthcare	23
2.7 Kidney Disease Detection Studies	24
2.8 Comparative Literature Analysis	27
2.9 Research Gaps	31
2.10 Summary	32
CHAPTER 3 - Research Methodology	33
3.1 Overall Framework	33
3.2 Dataset Description	35
3.3 Dataset Statistics and Class Distribution	35
3.4 Data Preprocessing	37
3.5 Data Augmentation	38
3.6 Experimental Environment	39
3.7 Evaluation Metrics	41
3.8 Workflow Diagram	42
3.9 Algorithm Overview	43
3.10 Summary	44
CHAPTER 4 - CNN-Based Kidney Disease Classification	45

4.1 VGG16 Architecture and Results	45
4.2 ResNet50 Architecture and Results	49
4.3 DenseNet121 Architecture and Results	53
4.4 EfficientNetB3 Architecture and Results	57
4.5 Comparative Analysis of CNN Architectures	60
4.6 Summary	61
CHAPTER 5 - Proposed Transformer Framework (RenalViT)	63
5.1 Introduction	63
5.2 Vision Transformer Fundamentals	64
5.3 Proposed RenalViT Architecture	65
5.4 Patch Tokenization	68
5.5 Positional Encoding	69
5.6 Multi-Head Self-Attention	69
5.7 Transformer Encoder Blocks	70
5.8 GELU Activation	71
5.9 Class Imbalance Handling	73
5.10 Training Strategy	74
5.11 Algorithm	75
5.12 Computational Complexity	76
5.13 Summary	76
CHAPTER 6 - Results and Discussion	78
6.1 Experimental Results Overview	78
6.2 Training and Validation Performance	79
6.3 Test Performance	80
6.4 Accuracy Comparison	80
6.5 Precision Comparison	82
6.6 Recall Comparison	83
6.7 F1-Score Comparison	84
6.8 ROC Analysis	84
6.9 Confusion Matrix Analysis	86

6.10 Statistical Analysis	88
6.11 Discussion	90
6.12 Summary	92
CHAPTER 7 - Conclusion and Future Work	93
7.1 Summary	93
7.2 Research Contributions	94
7.3 Major Findings	94
7.4 Limitations	95
7.5 Future Scope	96
References	98
List of Publication	105

List of Figures

Figure 1.1	Anatomy and Pathological Illustration of the Human Kidney	3
Figure 1.2	Classifications of Artificial Intelligence Methods	7
Figure 2.1	Vision Transformers (ViT) Architecture	23
Figure 3.1	Overall Research Methodology for Comparative CNN and RenalViT-Based Kidney CT Image Classification	34
Figure 3.2	Complete Research Workflow and Execution Flowchart	42
Figure 4.1	VGG16 Architecture with Five Convolutional Blocks and Classification Head for Kidney Disease Classification	46
Figure 4.2	VGG16 Confusion Matrix Graph	48
Figure 4.3	ResNet50 Architecture with Four Residual Stages, Bottleneck Blocks, and Skip Connections for Kidney Disease Classification	50
Figure 4.4	ResNet50 Confusion Matrix Graph	53
Figure 4.5	DenseNet121 Architecture with Four Dense Blocks, Transition Layers, and Dense Connectivity for Kidney Disease Classification	54
Figure 4.6	DenseNet121 Confusion Matrix Graph	57
Figure 4.7	EfficientNetB3 Architecture Overview	58
Figure 4.8	EfficientNetB3 Confusion Matrix	60
Figure 5.1	Proposed RenalViT Architecture: Hierarchical Patch Tokenization and Transformer Encoder Blocks	67
Figure 5.2	Comparison of Activation Functions: Gaussian Error Linear Unit (GELU) vs Rectified Unit (ReLU)	72
Figure 6.1	Comparative Model Performance Metrics across All Architectures	81
Figure 6.2	Accuracy Comparison: RenalViT vs CNN Baselines	82
Figure 6.3	ROC Curves: RenalViT Per-Class and Macro-Average AUC Performance	85
Figure 6.4	Proposed RenalViT Confusion Matrix on CT Kidney Test Set	87
Figure 6.5	Performance Radar Chart Comparison: Accuracy, Precision, Recall, F1, and AUC across all five models	90

List of Tables

Table 2.1	Comparative Literature Analysis of Kidney Disease Detection Studies	28
Table 2.2	Expanded Comparative Literature Analysis - Kidney Disease Classification Studies (2019-2026)	30
Table 3.1	CT Kidney Dataset Overview	36
Table 3.2	Class Distribution and Inverse-Frequency Class Weights for Training	36
Table 3.3	Data Augmentation Techniques Applied During Training	39
Table 3.4	Experimental Environment and Training Configuration	40
Table 4.1	VGG16 Architecture Configuration for Kidney Disease Classification	46
Table 4.2	VGG16 Performance Metrics on the CT Kidney Test Set	48
Table 4.3	ResNet50 Architecture Configuration	50
Table 4.4	ResNet50 Performance Metrics on the CT Kidney Test Set	52
Table 4.5	DenseNet121 Architecture Configuration	55
Table 4.6	DenseNet121 Performance Metrics on the CT Kidney Test Set	56
Table 4.7	EfficientNetB3 Architecture Configuration	58
Table 4.8	EfficientNetB3 Performance Metrics on the CT Kidney Test Set	59
Table 4.9	CNN Baseline Comparison Summary	61
Table 5.1	RenalViT Architecture Configuration	66
Table 5.2	RenalViT Hyperparameters and Training Configuration	75
Table 6.1	Comparative Performance of All Models on the CT Kidney Test Set	80
Table 6.2	RenalViT Confusion Matrix: Actual vs Predicted Classes on Test Set	87
Table 6.3	Model Accuracy with 95% Bootstrap Confidence Intervals and Complexity	88
Table B.1	Complete Hyperparameter Configuration for All Models	104
Table C.1	Per-class Precision/Recall/F1-Score (%) for All Models on CT Kidney Test Set	106
Table D.1	Publications Arising from the Thesis Research	107

List of Abbreviations

Abbreviation	Full Form
AdamW	Adam Optimizer with Decoupled Weight Decay
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CAD	Computer-Aided Diagnosis
CHAID	Chi-squared Automatic Interaction Detection
CKD	Chronic Kidney Disease
CNN	Convolutional Neural Network
CFS	Correlation-based Feature Selection
CLS	Classification token
CT	Computed Tomography
DeiT	Data-efficient Image Transformers
DPC	Doctoral Progress Committee
eGFR	Estimated Glomerular Filtration Rate
EPO	Erythropoietin
ESRD	End-Stage Renal Disease
ESKD	Early-Stage Kidney Disease
F1	F1-Score (Harmonic Mean of Precision and Recall)
GBT	Gradient Boosted Trees
GELU	Gaussian Error Linear Unit
GPU	Graphics Processing Unit
HU	Hounsfield unit
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
KFRE	Kidney Failure Risk Equation
KNN	K-Nearest Neighbors
LSVM	Linear Support Vector Machine
LR	Logistic Regression
LASSO	Least Absolute Shrinkage and Selection Operator
MBConv	Mobile Inverted Bottleneck Convolution
MLP	Multi-Layer Perceptron

MRI	Magnetic Resonance Imaging
MHSA	Multi-Head Self-Attention
NB	Naive Bayes
NBC	Naive Bayes Classifiers
PACS	Picture Archiving and Communication System
RAAS	Renin Angiotensin Aldosterone System
RCC	Renal cell carcinoma
ReLU	Rectified Linear Unit
RenalViT	Renal Vision Transformer
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
SE	Squeeze-and-Excitation
SMOTE	Synthetic Minority Over-sampling Technique
SiLU	Sigmoid Linear Unit
SVM	Support Vector Machine
US	Ultrasound
USG	Ultrasonography
UCI	University of California, Irvine
ViT	Vision Transformer
VGG16	Visual Geometry Group Network (16 layers)
VRAM	Video Random Access Memory
XAI	Explainable Artificial Intelligence

List of Symbols

Symbol	Description
Q	Query matrix in self-attention mechanism
K	Key matrix in self-attention mechanism
V	Value matrix in self-attention mechanism
W_Q	Weight projection matrix for Query
W_K	Weight projection matrix for Key
W_V	Weight projection matrix for Value
W_O	Weight projection matrix for output of multi-head attention
d_k	Scaling factor / dimensionality of key vectors
h	Number of attention heads in multi-head self-attention
D	Embedding dimension (256 in RenalViT)
N	Total number of training samples
N_k	Number of samples in class k
K	Number of classes (4: Normal, Cyst, Stone, Tumor)
w_k	Class weight for class k (inverse frequency)
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
$\Phi(x)$	Cumulative distribution function of the standard normal distribution (GELU)
Z_b	Token sequence after transformer encoder block b
E_{pos}	Learnable positional embedding matrix
p	Predicted probability vector (output of softmax)
y	True class label (one-hot encoded)
L	Loss function value (class-weighted cross-entropy)

CHAPTER - 1

Introduction

Kidney illness represents an extensive and increasing worldwide health issue and there are thousands of millions of people globally that suffer from drinking issues as a result of CKD, cysts and stone and tumors. Medical imaging plays a major role in identifying kidney problems early and accurately, making it important for physicians to have the tools to do so. As a result, this chapter will provide a description, motivation, problem definition and objectives to the research project to set the groundwork for developing an automated deep learning system to classify CKD and other kidney issues via CT imaging.

1.1 Background

Kidney disease poses a major growing issue to public health in the 21st century. The four main types of kidney disease - Normal (healthy), Cyst, Stone (nephrolithiasis), and Tumor (primarily renal cell carcinoma) - are different from each other in terms of their structure, clinical course, and treatment.

According to data from the Global Burden of Disease Study, chronic kidney disease (CKD) affects roughly 10% of adults globally and causes more than 1.2 million deaths each year. CKD also has significant healthcare costs associated with it [2, 3]. Nephrolithiasis occurs in 10-15% of people around the world and has a greater than 50% chance of recurrence within 10 years after the first occurrence [4]. Approximately 90% of kidney cancers are renal cell carcinoma (RCC), with more than 430,000 new cases diagnosed worldwide each year [5].

In addition to filtering blood, the kidneys also perform many other important functions needed for the body's normal function. They help maintain acid-base balance by getting rid of hydrogen ions and saving bicarbonate, keep fluids at equal levels by using a variety of mechanisms including the renin-angiotensin-aldosterone system (RAAS), they produce erythropoietin (EPO) for making red blood cells and activate vitamin D for calcium use [6, 7]. Any change in how well these functions are done due to damage to the kidney's structure,

problems with blood vessels, or other underlying issues can affect how the whole-body functions, this includes more than just the urinary system [8]. CT kidney image data with four main pathological conditions normal kidney, kidney with a cyst, kidney stones and kidney with a tumor are clinically unique from each other regarding how they look on CT scans, what type of treatment will be used for each condition, etc.

Therefore, timely and accurate diagnoses are very important in the clinical setting. As a result of CT imaging being the most accurate method for assessing kidney disease, it is the most accurate diagnostic test available. The enhanced spatial resolution allows more accurate differentiation between normal and abnormal tissue types. An additional problem with the use of CT images for kidney disease assessment is the length of time required by a medical doctor or radiologist to visually evaluate the CT images due to the high degree of inter-reader variability as well as the increasing number of CT images generated in the clinical setting [38], [40].

Recently developed deep learning, a subfield of artificial intelligence (AI) has created new ways of automatically, objectively, and efficiently analyzing medical images. Convolutional Neural Networks (CNNs) are a well-established method for performing medical imaging tasks. However, due to their reliance on local feature extraction techniques, they cannot capture long range spatial dependencies across multiple images. [49] Vision Transformers (ViTs), on the other hand, provide a fundamentally different mechanism for performing medical imaging analyses by using self-attention to compute the relationship between all parts of the image simultaneously. This is particularly beneficial for CT imaging as pathologic features can extend across a large portion of the CT image [38], [41].

1.2 Kidney Anatomy

Kidneys are paired, bean shaped organs located in the retroperitoneal space just anterior to the dorsal body wall of an individual, on either side of the vertebral column, at the level of the T12-L3 vertebrae. Each kidney is 10-12 cm long, 5-7 cm wide, and 2-3 cm thick in adult human and weigh approximately 120g -170g.

To give the physiological basis for chronic kidney disease detection, knowledge of anatomical and structural aspects of human kidneys is critical. An illustration of the detailed anatomy of kidneys and the complex blood supply and internal structure of kidney will be presented in the following Figure 1.1.

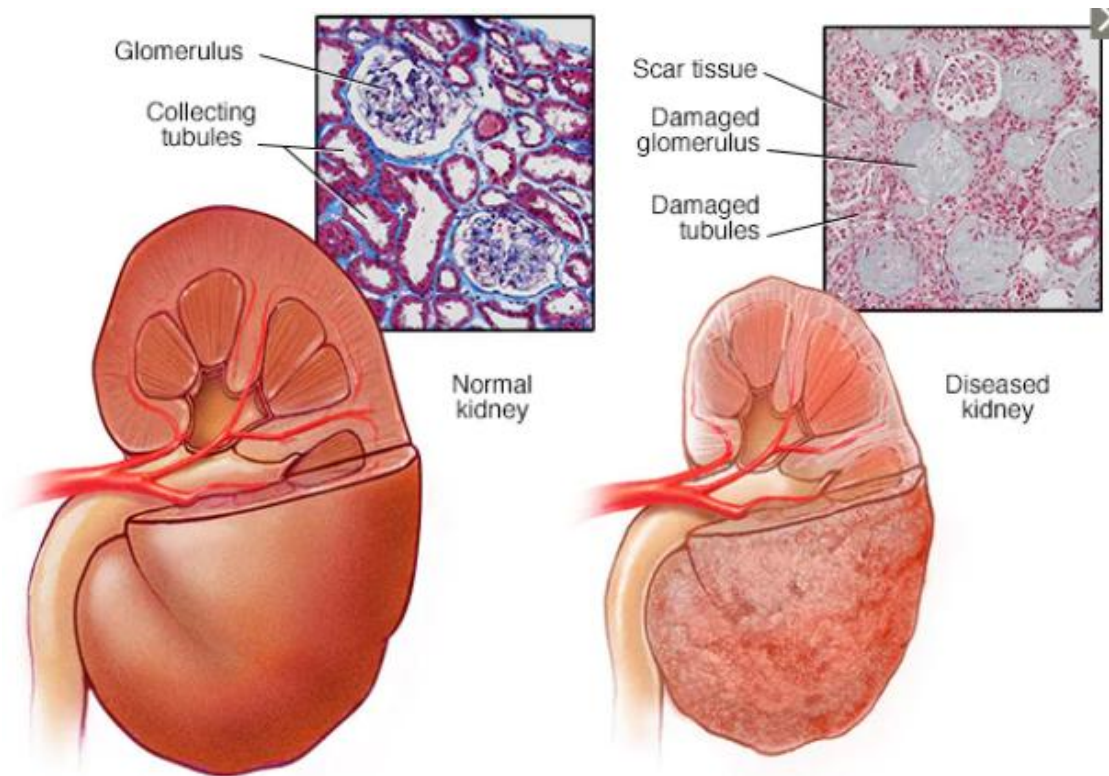


Figure 1.1: Anatomy and Pathological Illustration of the Human Kidney

The Anatomic regions of the renal system are demonstrated within the following diagram Figure 1.1 and are representative of how renal pathologies such as cysts, stones, and neoplasms develop in the renal system according to their inherent anatomical functional layer. The clinical need to improve automated classification of High-Resolution (HR) Medical Imaging, To Support Radiologist's ability to detect lesions early is demonstrated.

The internal kidney architecture consists of three primary anatomical zones, the cortex outer layer that contains Approximately 1 million Nephrons, which are the functional Unit of the Kidney, the Medulla is formed by 8 To 18 Renal Pyramids, Which Contain Collecting ducts, And the Renal Pelvis a funnel-like structure receiving urine from the Calyces. The main functions of the kidney include the filtration of blood to produce urine, maintenance of fluid, electrolyte, and acid-base balance, and hormones secreted by the kidney such as renin, erythropoietin, and calcitriol.

The complex structure and central importance of the kidney result in pathological conditions related to the experimental modality of imaging, whether it be through Vascular, inflammatory, cystic, lithogenic, or neoplastic changes, having potentially far-reaching systemic effects. CT imaging provides HR Cross-sectional Visualization of All Three

Kidney Anatomical Areas and allows for the reliable detection and characterization of a wide variety of renal diseases.

1.3 Chronic Kidney Disease

Chronic kidney disease (CKD) is characterized by a kidney structural or functional abnormality lasting longer than three months that negatively impacts health. CKD is grouped according to the estimated glomerular filtration rate (eGFR) and has five stages, stage 5 (eGFR <15 ml/min/1.73 m²) is ESRD and requires dialysis or transplantation.

Diabetes-related nephropathy accounts for approximately 40% of all CKD cases worldwide, followed by hypertension-related nephrosclerosis, glomerulonephritis, polycystic kidney disease, and obstructive uropathy. CKD frequently exhibits no symptoms early on. Therefore, tools that are imaging-based or provide early detection of CKD will be invaluable when it comes to clinical value.

The National Kidney Foundation Kidney Disease Outcomes Quality Initiative (NKF-KDOQI) has defined five stages of Chronic Kidney Disease (CKD) based on estimates of kidney function (that is, a patient's GFR), as follows: Stage one (GFR threshold 90 mL/min/1.73 m²), kidney damage with a normal/increased GFR, stage two (GFR threshold 60-89), mild decreased function kidneys, stage 3a (GFR threshold 45-59) and 3b (GFR threshold 30-44), moderate decreased function kidneys, stage four (GFR threshold 15-29), severe decreased function kidneys and stage five (GFR threshold <15), kidney failure. The progressive nature of CKD means that early identification of patients in stages 1 to 3 will provide clinicians the opportunity to intervene and slow the disease progression. Unfortunately, the majority patients do not experience any symptoms until Stage 3b or Stage 4 when their bodies can no longer compensate for their decreased kidney function. Therefore, the novel automated CT/US imaging screening/classification tools created as part of the proposed work will have tremendous clinical value in addressing this patient diagnostic dilemma.

1.4 Kidney Cyst

A cyst of the kidney is a sac filled with fluid found within the renal parenchyma. Simple renal cysts are the most common type of benign renal lesion and their prevalence increases with age. Simple cysts appear on CT as well-defined and homogeneous, round or oval, with

low-attenuation, Hounsfield units typically -10 to +20 HU and an imperceptible thin wall, and no enhancement on contrast imaging.

In the CT kidney dataset used in this research, 3,701 of 12,389 total images were classified as renal cysts, this category represents the second largest in the dataset. The visual similarities in appearance between simple cysts and normal renal parenchyma at some imaging windows is a primary reason for the difficulty of classifying cysts in the dataset.

1.5 Kidney Stone

Nephrolithiasis is the most common urological disorder among people worldwide, with current estimates showing it affects between 10 and 15 percent of the adult population, and the recurrence rates are continually increasing, as over 50 percent of people who have passed a kidney stone will repeat this process within 10 years. Nephrolithiasis is caused when minerals crystallize within the renal collecting system of the kidneys, the majority of the time, these crystals are made of calcium oxalate, uric acid, struvite, or cystine.

An unenhanced CT, the preferred method to evaluate these issues will demonstrate these stones as hyperdense (bright) structures via Hounsfield unit (HU) values of greater than 200 HU to as much as 1,000+ HU for calcium-based stones. Stone size can vary from less than 1 mm to greater than 20 mm in diameter, the stone's appearance, smooth, facet, or irregular will depend on the stone's composition. In the CT Kidney Dataset, there were a total of 2,283 stones out of 12,478 total images, indicating that stones comprise the third most frequently occurring category of images.

1.6 Kidney Tumor

Renal cell carcinoma (RCC) is the most common type of cancer that starts in the kidneys in adults. It accounts for approximately 90% of all cancers of the kidneys and there are more than 430,000 new patients diagnosed around the world each year. About 70-75% of these cancers will be clear cell RCC, 10-15% will be papillary and approximately 5% will be chromophobe.

On CT scans using contrast agents, renal tumors usually appear as solid, heterogeneous or homogeneous masses with variable amounts of enhancement from the contrast agent. Characteristics of enhancement are very important for the diagnosis and staging of renal tumors. Therefore, characteristics of enhancement from the CT kidney dataset provide some of the least prevalent renal tumors but have one of the highest false-negative impacts due to

classification of them as something other than what they really are. Tumor-containing images within the CT kidney dataset account for 1,377 samples (11.1% of all images) and represent the smallest and most clinically significant category of renal tumor images within the dataset.

1.7 Medical Imaging Techniques

There are many methods of medical imaging helping to display internal anatomy without having to invade the body in any way. The three main imaging methods for evaluating the kidneys are ultrasound (US), magnetic resonance imaging (MRI), and computed tomography (CT).

Of these methods, US is commonly the first method of imaging used for evaluating a patient's kidney due to low cost, previous usage in many locations, and no exposure to radiation when performing the imaging. Unfortunately, US has limited tissue differentiation and is operator-dependent so it is difficult to use US to create an automated way to analyze images produced by US.

MRI provides good contrast of soft-tissue, no radiation exposure, but costs a lot of money and takes longer and is not as readily available as US. CT provides good high spatial resolution, rapid acquisition, and characterizes tissues consistently across pathological states of the kidney.

CT scans are used to diagnose kidney disease because, unlike ultrasounds, their imaging quality is consistent and reproducible regardless of who is performing them or the patient's body type (size). Also, by using Hounsfield Units (HUs) that quantify the attenuation of X-rays, it is easy for CT scans to differentiate between soft tissue, fluid, calcifications, or fat with precision. The typical range of renal cysts is -10 to +15 HU, while renal calculi are hyper-dense (300-2000 HU) and solid renal tumors may show heterogeneous/homogeneous enhancement with contrast. In clinical practice, the use of unenhanced, contrast-enhanced, and excretory-phase CT (triple-phase CT) provides a comprehensive evaluation of renal anatomy, characterizing lesions, and collecting system anatomy, all in one study. The CT Kidney Dataset will be used in this research, which includes un-enhanced CT images obtained from multiple clinical center PACS systems, representing a population of clinical images that can be used for developing deep learning models.

1.8 Artificial Intelligence in Healthcare

The use of AI in the field of providing patient care and treating patients has grown significantly in the last ten years due to advancements in deep learning, as well as the creation of large, high quality data sets and improved computer technology [49]. CNN based Artificial Intelligence models now have shown comparable accuracy to the performance of medical doctors for example when interpreting images in Dermatology, Radiology, Ophthalmology, Pathology. CNN based Artificial Intelligence models are therefore able to provide valuable assistance to medical professionals when interpreting medical images.

In recent years advanced artificial intelligence methods have been increasingly adopted across various healthcare sectors to help resolve the aforementioned issues associated with interpreting CT scans performed on patients. Advanced artificial intelligence methods can be broadly categorized into three categories: (1) Statistical methods, (2) Machine learning methods and (3) Natural language processing methods. A visual representation of all three categories of Artificial Intelligence methods is illustrated in figure 1.2.

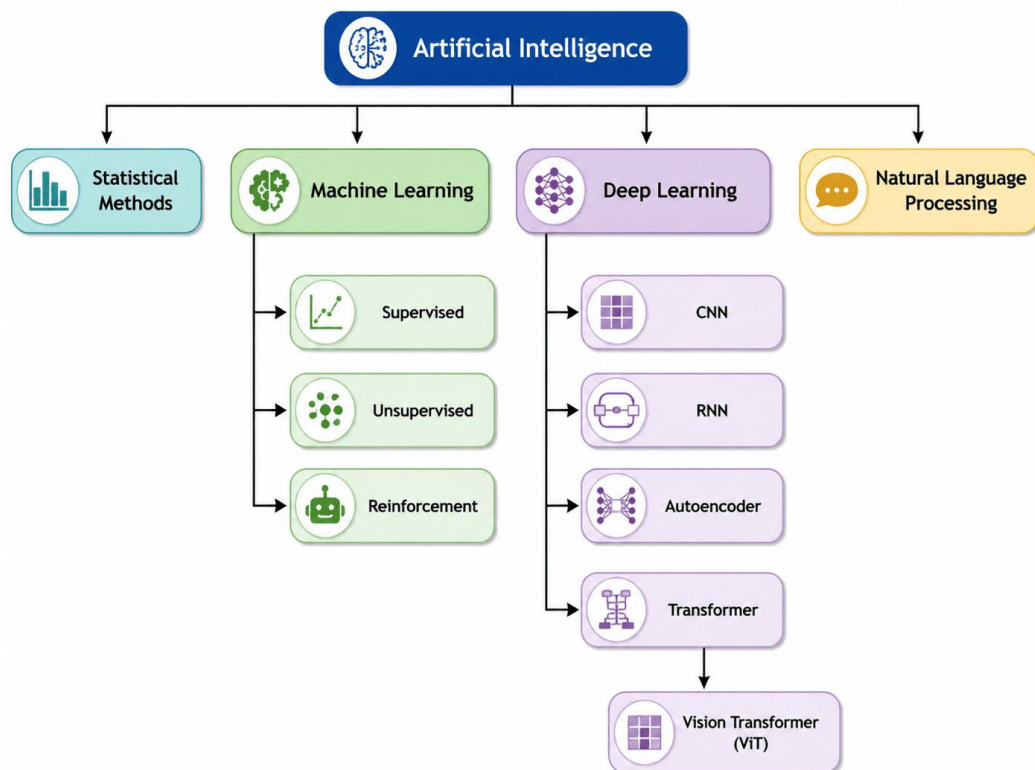


Figure 1.2: Classifications of Artificial Intelligence Methods

The hierarchical classification in this document shows how machine learning has progressed over time. Deep learning and transformer-based models represent the leading edge of technology for analyzing high-dimensional medical imaging, where both the complexity of the raw pixel input data and the number of dimensions being analyzed are substantial.

The convergence of artificial intelligence and medical imaging has accelerated significantly since 2012, when deep learning first demonstrated superhuman abilities in a substantial way at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). In the field of medical imaging and as a subset of AI, CNNs have documented their state-of-the-art performance in radiology, pathology, ophthalmology, and dermatology [46]. Recently, the introduction of transformer-based architectures and their corresponding adaptation into computer vision, as seen with the Vision Transformer (ViT), has given rise to a different way of analyzing medical images [46]. The ability of ViTs to effectively analyze complex multi-class medical classifications through the use of their global understanding and efficient identification of anatomical location/configuration is another differentiating factor when compared to CNNs, which are inherently limited by their local receptive field. Therefore, this thesis leverages the then discussed advantage regarding the architecture of ViTs as it relates specifically to the classification of kidney CT images.

Transformer-based architectures have been modified to analyze images by processing images as a series of image patches with the aid of a self-attention mechanism that models the global spatial relationships present between all the image patches [46]. ViT and its hierarchical variations have achieved better overall performance than other methods on numerous visual recognition benchmark tasks and are now being applied to the analysis of medical images.

1.9 Motivation

Three important findings have led to this study. Firstly, kidney problems have a major impact upon public health since classifying them quickly and accurately directly affects outcomes. Secondly, while CNNs have been used extensively for classifying kidneys by computerized tomography [39], [40] there has never been a proper side-by-side benchmark study between CNNs and Transformer models being tested against each other using the same dataset. Finally, the large difference in occurrences of normal and cyst images vs tumor images in the CT kidney datasets increases the challenges faced by automated classification systems. The prior literature does not adequately address these issues.

1.10 Problem Statement

Medical image datasets used for kidney disease diagnosis are typically large, heterogeneous, and complex, containing significant variations in image quality, anatomical structures, and disease characteristics. Such variability requires extensive preprocessing, transformation, and augmentation before effective feature extraction and model training can be performed. Furthermore, publicly available kidney CT datasets often exhibit class imbalance, making automated disease classification more challenging. Although existing deep learning approaches, particularly CNN-based models, have demonstrated promising results, their classification accuracy remains limited, especially for clinically important minority classes. These limitations highlight the need for a more robust deep learning framework capable of improving feature representation, addressing class imbalance, and enhancing overall classification accuracy. Therefore, this research proposes a Vision Transformer-based framework (RenalViT) for accurate and reliable classification of kidney CT images into Normal, Cyst, Stone, and Tumor categories.

1.11 Research Objectives

This thesis intends to accomplish the following research goals:

- Goal 1: Create a fully-automated deep learning system to classify kidney CT images into four categories, which consist of normal, cyst, stone, and tumor.
- Goal 2: Compare various CNN architectures (VGG16, ResNet50, DenseNet121, EfficientNetB3) directly against each other and against the proposed transformer-based RenalViT framework under the same experimental conditions.
- Goal 3: Remedy class imbalance issues in the dataset using an inverse frequency-based loss function for weighting samples by their respective classes during training.
- Goal 4: Identify the best possible classification accuracy of the CT Kidney Dataset using the proposed RenalViT framework above any of the other tested CNN architectures.
- Goal 5: Provide empirical support for the significant architectural decisions made (activation functions used, input image size, number of transformer encoder layers, optimizer) in developing the RenalViT framework.
- Goal 6: Provide an analysis of the trade-off between computation complexity and processing efficiency of the proposed RenalViT framework compared to each of the evaluated CNN architectures.

- Goal 7: Contribute to the body of peer-reviewed scientific literature by publishing the findings of this research in appropriate venues.

1.12 Thesis Outline

The remainder of this thesis is presented as follows. A literature review covering CNN architectures and transformer models, as well as previous research on classifying kidney disease, is presented in Chapter 2 and summarizes all literature utilized in the research. The research methodology is described in Chapter 3 including information on the dataset used, preprocessing pipeline and augmentation techniques, experimental environment and evaluation metrics. In Chapter 4 we present our implementation and results of four baseline CNN architectures. The proposed RenalViT framework is discussed in Chapter 5 in detail. Chapter 6 covers comprehensive experimental results from all experiments performed, including a comparison of these results, followed by a discussion of the comparisons. The conclusion of this thesis is presented in Chapter 7 including a summary of the contributions made by this research, limitations of the research conducted, and directions for future research.

1.13 Summary

In this chapter, I have explained the background and clinical reasons for conducting my research. The worldwide increase in kidney disease and the limitations of CT interpretation by hand provide justification for creating the RenalViT framework to develop the methods for classifying images of the kidneys using CT scan data via a transformer model. In addition, I have stated the research goals and given an overview of the thesis project. The next chapter will provide an overview of the current literature related to this research.

The information presented in this chapter formed the basis of all research performed for this thesis. The kidney is a highly complex organ with an important role in the body's function, when there are pathological (cystic, calcific, or neoplastic) changes to the kidneys, these will have immediate clinical implications.

The four main categories of renal pathology encountered in this project were described Normal, Cyst, Stone, Tumor in regards to their clinical presentations, CT image characteristics and epidemiological prevalence. CT imaging was established as the standard of care for evaluations of renal pathology and subsequently the rapid advancement of AI-based technologies within medical imaging was discussed. This was all based upon three

key observations: the global burden of kidney disease, the limitations associated with manual CT image interpretation, and the untapped potential for transformer-based architectures to be used for renal CT classification. The research objectives, problem statement and scope were also clearly identified. Each of the subsequent chapters systematically build off of this foundation: Chapter 2 provides a thorough review of existing literature related to the detection of kidney diseases through deep learning, Chapter 3 outlines the methodology, Chapters 4 and 5 present the baseline CNN and the proposed RenalViT results, and Chapter 6 includes a comprehensive comparative analysis. Finally, Chapter 7 gives a summary of all work completed within this thesis.

CHAPTER - 2

Literature Review

The systematic review of the literature which is the subject of the following chapter demonstrates the use of Convolutional Neural Networks (CNNs), Transformer-based models, and previous studies that focus specifically upon the automatic detection or classification of kidney issues based on computed tomography (CT) images.

The automated analysis of medical imaging has changed dramatically over the last 10 years. The original methods used handcrafted radiology features such as Hounsfield distribution values, shape descriptors, and texture indices, which were input into traditional machine-learning classifiers like Support Vector Machines (SVMs), Decision Trees (DTs), and Naive Bayes Classifiers (NBCs). While these were good foundations for these types of analysis, the limitations of what was able to be input into these classifiers was due to the selection of features by domain expert physicians (i.e., feature engineering). The introduction of deep learning and Convolutional Neural Networks (CNNs) removed the limitation of needing to develop features by allowing for end-to-end learning of features directly from the pixel intensities in the images. Recently, the Vision Transformer (ViT) has emerged as an alternative method for determining long-distance spatial relationships among entities, through the mechanism of global self-attention - an improvement over the restricted spatial relationships provided by CNNs. This chapter discusses the progression of the use of automated methods for classifying kidney disease and other medical imaging modalities and identifies the position of the RenalViT Framework within the current state of the art. The sections of this review are organized as follows: (2.1) Convolutional neural networks for medical imaging, (2.2) Machine Learning in Medical Imaging, (2.3) Deep Learning Models for Image Classification, (2.4) An explanation of the main types of CNN architectures examined in this thesis, (2.5) An introduction to vision transformers, (2.6) Transformer architectures in healthcare, (2.7) Research studies regarding detection of kidney disease, (2.8) A comparative literature analysis, (2.9) Identification of research gaps, and (2.10) A summary for this chapter.

2.1 Convolutional Neural Networks for Medical Imaging

CNNs have emerged as the most widely-used technique for automatic analysis of medical images with many advantages by being able to directly learn features hierarchically through end-to-end learning of the complete convolutional layers, thereby providing an improved level of reliability and accuracy when classifying, segmenting and detecting objects in medical images compared to traditional machine learning approaches based upon hand-crafted descriptors for texture, shape and intensity [19].

The unparalleled success of CNNs in the field of computer vision has been greatly impacted by their performance in the ILSVRC and the adoption of their use for great advances made over traditional recognition approaches. The positive impact of CNN ability to generalize across complex visual patterns has led to their application within the field of healthcare including such images as X-rays, CT-scans, MRI-scans, Otoscopy images, etc. CNN have also shown significant performance in analysing subtly abnormal anatomy in medical images where traditional feature engineering cannot capture the abnormalities due to the nature of how feature engineering is applied.

The use of CNN-based algorithms for detecting and classifying diseases has previously been proven effective in radiology by using CT, MRI, ultrasound, and X-ray imaging techniques, where CNNs utilize a hierarchical learning mechanism whereby the earlier layers provide the basis upon which the later layers build their ability to identify organ and lesion morphology and key features associated with each disease process. Therefore, by utilizing this progressive method of feature extraction that combines many of the primary and secondary characteristics observed throughout the disease process, CNNs have achieved high diagnostic accuracy due to their ability to discriminate even subtle visual differences between diseases.

In terms of analyzing kidney disease, CNNs represent one of the most widely used deep-learning techniques because CT images provide an abundance of spatial and structural data for effective utilization of convolutional operations. Additionally, individual research teams have recently published many results identifying both kidney cysts, stones, tumors, and normal kidneys by using various CNN architectures. Each research team's findings demonstrate that their respective deep feature extraction capabilities significantly increase classification accuracy over previously used traditional ML algorithms.

According to a comprehensive survey undertaken by [49], the adoption of deep learning in the analysis of medical images is pervasive throughout clinical areas including radiology, pathology, ophthalmology, dermatology, etc. The survey established the role of CNNs as the basis for contemporary medical image analysis, with their performance exceeding that of classical machine learning techniques across many diagnostic tasks. These results indicate that CNN's have revolutionized computer-aided diagnosis by enabling accurate, scalable and automated analysis of complex medical images, thus they represent the building blocks of future deep learning architectures built on this foundation.

2.2 Machine Learning in Medical Imaging

Machine learning (ML) has played a fundamental role in the evolution of computer-aided diagnosis by enabling computers to identify patterns and make predictions from medical data without being explicitly programmed. In medical imaging, machine learning techniques have been widely applied to tasks such as disease classification, lesion detection, image segmentation, risk prediction, and clinical decision support. Traditional ML approaches generally consist of multiple stages, including image preprocessing, feature extraction, feature selection, and classification. These methods have been successfully applied to various imaging modalities such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound (US), and X-ray imaging, contributing to improved diagnostic efficiency and reduced clinician workload.

Early machine learning systems relied heavily on handcrafted features that describe image characteristics such as texture, shape, intensity, statistical properties, and edge information. These features were subsequently classified using algorithms including Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), k-Nearest Neighbors (KNN), Naïve Bayes (NB), and Logistic Regression [24]. Such algorithms demonstrated promising performance for several medical imaging applications, including cancer diagnosis, brain tumor classification, lung disease detection, diabetic retinopathy screening, and kidney disease identification. Their relatively low computational complexity and interpretability made them attractive for many early clinical decision-support systems.

Despite their success, conventional machine learning methods possess several limitations when applied to large-scale medical image analysis. Their performance depends heavily on the quality of manually extracted features, requiring extensive domain expertise and significant preprocessing efforts. Furthermore, handcrafted features often fail to capture complex anatomical structures and subtle pathological variations present in high-

dimensional medical images. Variations in imaging protocols, patient anatomy, image quality, and class imbalance further reduce the robustness and generalization capability of traditional machine learning models, particularly when applied to diverse clinical datasets.

The emergence of deep learning has addressed many of these limitations by enabling automatic hierarchical feature learning directly from raw medical images. Nevertheless, traditional machine learning techniques continue to play an important role in healthcare research due to their computational efficiency, interpretability, and effectiveness on relatively small datasets. Understanding the strengths and limitations of machine learning provides an important foundation for appreciating the transition toward deep learning and transformer-based architectures discussed in the subsequent sections of this chapter.

2.3 Deep Learning Models for Image Classification

The rapid proliferation of advancements made in deep learning has changed fundamentally the way that images are classified as it allows for models to learn complicated feature representations, from raw (unprocessed) image information. While traditional machine learning approaches require human-created features to be extracted and then classified by separate algorithms; deep learning combines extraction and classification into one complete optimization process. Therefore, deep neural networks benefit from the ability to automatically identify visual patterns that will best help distinguish between the target classes [25].

The innovative development of CNN architectures that are both deeper and more sophisticated has been one of the driving forces behind the dramatic advancements in image classification. Early CNN models showed that adding more layers increases feature representation significantly, while recent CNNs have introduced design improvements e.g., residual networks, dense connectivity, etc. to help solve the optimization problems created by very deep CNNs. These improvements have led not only to increased classification performance but also to increased computational efficiency, increased network stability during training, and greater reuse of learned features.

The use of deep learning for medical image analysis has proven to be especially effective because there are often very small visual differences between an image or feature that is captured and the corresponding image or feature that has been processed using traditional methods. CNN-based architectures will automatically learn to use a hierarchy of features from low-level textural features to high-level anatomical and pathological features, which

allows for easy identification of the disease pattern from various types of images being used in medical imaging. Therefore, CNNs have been successfully applied to disease diagnosis, finding lesions, segmenting organs, diagnosing cancer, and many more computer-assisted applications in medicine.

Another factor that has helped promote the use of deep learning extensively is the concept of Transfer Learning, where pretrained Neural Networks are fine-tuned for particular medical imaging tasks using large publicly available datasets (e.g., ImageNet). Transfer Learning allows these pretrained models to exploit the generic visual features they learned previously while making adjustments to accommodate the unique visual characteristics of the medical domain, thereby allowing for improved classification performance, especially when the number of annotated medical images available to train on is limited. As a result, using Transfer Learning has become common practice for many medical image classification studies due to its ability to decrease the amount of time required to train a model, speed up the rate at which the model converges to a solution, and lead to better generalization of the trained model.

Several recent architectural improvements to CNNs have led to the development of several major CNN architectures that mark important milestones within the realm of deep learning. The architectures VGG16, ResNet50, DenseNet121, and EfficientNetB3 have all made significant contributions and improvements regarding feature extraction and gradient propagation; parameter utilisation and computational performance. Because of their demonstrated effectiveness and widespread use, these four architectures were used as baseline models for the architecture proposed in this research to serve as a comparison to the proposed Transformer-based RenalViT architecture.

2.4 CNN Architectures

Over the last 10 years, large advancements have been made to CNNs. CNNs have evolved through a series of different architectural designs, which have pushed the limits on how well features can be extracted, how computationally efficient they can be during inference, and how stable they can be during training. Each new architecture that was created could solve specific problems or difficulties in its predecessor, which allowed for continued improvements in how well large numbers of images could be classified. These advances have also had a significant impact on performing analysis of medical images, where accurate detection of detailed anatomical structures or the presence of disease is critical for an

accurate diagnosis. There are many CNN architectures proposed throughout the literature, but VGG16, ResNet50, DenseNet121, and EfficientNetB3 have been among the most popular models due to their strong classification accuracy and other criteria such as the availability of pre-trained ImageNet weights and their success in supporting many forms of medical imaging applications. Different design philosophies are represented in the four architectures: VGG16 emphasizes the use of many layers with a lot of small receptive fields in the architecture. Residual networks e.g., ResNet50 help to train very deep neural networks by utilizing “skip connections” from the input to the output. Dense networks e.g., DenseNet121 allow for the use of previously learned features by providing direct connections i.e., gradients between layers. Efficient networks e.g., EfficientNetB3 have been designed to improve classification accuracy or the ability of the CNNs to make accurate predictions about the data, while maintaining the lowest computing cost possible i.e., achieving the most efficient use of CPU, RAM, and energy to run the architecture. The four architectures discussed above are four major milestones in the development of CNNs, and they represent four complementary methods for accomplishing the feature extraction part of the CNN. The four architectures were also collectively analyzed due to their different architecture designs and commonalities, their prior widespread usage in medical diagnostics and successful implementation of transfer learning. This made using the four architectural designs for this study possible, while allowing for a high level of standardization within a specific set of experimental protocols when conducting evaluations of the RenalViT framework against state-of-the-art results.

2.4.1 VGG16

The Visual Geometry Group (VGG) at the University of Oxford's proposed VGG16 is an extremely impactful Convolutional Neural Network (CNN) design for image classification on a larger scale. From its original work building on the ideas surrounding long depth, such as this model demonstrated the effects of increasing depth through a simple, uniform structure can have on improving the performance, without having to use many complex Convolutional operations, created an accepted benchmark for incremental improvements for deep learning literature, which greatly contributed to the acceptance of deep CNN architectures.

VGG16 has 16 trainable layers, 13 Convolution layers and 3 Fully Connected layers. VGG16 does not utilize any large Convolutions to reduce spatial resolution in the input image at the output of VGG16, using only small 3x3 Convolution filters with a stride of 1 to pad the

output images to remain as close to the original images as possible with each layer's output of VGG16, thus the VGG16 output images are considered to be spatially resolved, almost the same as the original inputs until after the last max pool has been performed at a 2x2 level, from that point forward, the images will begin losing some of their spatial resolution due to subsequent pooling operations that have already been applied to them. The number of Feature Maps (FM) in the VGG16 architecture is increased as we move from the early to the later convolutional layers, through the first 5 convolutional blocks, 64 FMs will be created in the first block, increasing to 512 in the final block.

One of the main benefits of VGG16 is its very simple and very regular architecture, which makes it both easy to understand and easy to implement and modify for various transfer learning applications. Furthermore, because of the use of the ImageNet pretrained weights, it has become one of the most common backbone networks for classifying medical images, where often the only way to adapt with limited annotated datasets is by using transfer learning. Furthermore, because VGG16 can reuse feature representations learned through pretrained weights, it can learn how to represent general visual patterns and modify them through retraining to fit the specific characteristics of the medical imaging being analysed.

However, VGG16 has approximately 138 million trainable parameters, so it is expensive to run computationally and can be more prone to overfitting when being trained with smaller datasets. The lack of any shortcut connections also results in less ability to propagate gradients through deeper layers during training and makes it more difficult to optimise the model compared to later network architectures such as ResNet. Nevertheless, VGG16 is a very good example of a simple, yet effective method for evaluating the performance of new deep learning models in the field of medical image analysis due to its inherent simplicity, strong baseline performance and because it has been used so extensively.

2.4.2 ResNet50

ResNet50 is a revolutionary deep learning model with an innovative architecture based on residual learning. Residual learning solves the vanishing gradient and degradation problems associated with training extremely deep networks. Instead of attempting to learn the complete set of feature transformations through traditional CNNs that use multiple convolutional layers for every input, ResNet allows the model to learn residuals of the mapping function between the input and output through identity shortcuts. As a result, backpropagated gradients can be propagated through these identity connections more

effectively than through direct convolutions. Thus, ResNet made it possible to train deeper networks than ever before while maintaining optimal convergence.

The architecture of ResNet50 consists of a total of 50 learnable layers, which are composed of four residual stages or groups, in addition to bottleneck residual blocks. The bottleneck building block is made up of three consecutive convolutional layers, a 1x1 convolution to reduce the dimensionality of the feature maps at the bottleneck, a 3x3 convolution to process the spatial features in the spatial feature maps and a second 1x1 convolution to produce output feature maps that match the original dimensionality. The identity shortcut connects the input feature maps or feature maps that were output from the previous layer or set of layers to the block's output to allow for optimal retention of the useful information learned during the convolution operations, with the result that the model learns just the residuals of the mapping function.

ResNet50 is significantly more efficient than VGG16, despite its larger depth. It accomplishes this through use of a bottleneck design which reduces the total parameters of ResNet50 to approximately 25 million significantly less than the 138 million of VGG16, while increasing the representation of features and stabilizing training. The residual learning employed within ResNet50 has further facilitated the propagation of gradients through multiple levels of deeper feature hierarchies, resulting in improved representations of abstract features.

ResNet50 has demonstrated excellent success across many subfields of computer vision and medical imaging, such as disease category classification, organ surface segmentation, lesion detection, and cancer diagnosis. Its balance of model complexity, computational efficiency, and classification accuracy has made it one of the most commonly used CNN architectures for transfer learning. Therefore, ResNet50 will serve as the primary baseline model for the purposes of evaluating the efficacy of the transformer-based models presented herein for the classification of kidney CT images.

2.4.3 DenseNet121

DenseNet121 builds off of the idea of residual learning to provide for dense connectivity, where each layer receives the feature maps created by all previous layers to help provide its own output as well as to send its output to all other subsequent layers in a dense block. Each layer does not learn the same feature representations independently, but instead creates its output based on the entire combined knowledge that is acquired through the network. This

propagation of features densely encourages high amounts of feature reuse and helps maximize the flow of information throughout the training process.

The structure of DenseNet121 consists of 121 total layers grouped into four dense blocks of layers, separated by transition layers that each comprise of a 1x1 convolution followed by average pooling. Feature maps generated within the dense blocks of layers will be concatenated so that each layer can simultaneously access both low-level and high-level representations that have been created by the previous layers in the network. This type of connectivity pattern greatly improves the propagation of gradients through the model and reduces the likelihood of encountering the common vanishing gradient phenomena associated with deep convolutional neural networks.

DenseNet121 has the attribute of being very efficient in terms of how many parameters it uses i.e., it has about 8 million parameters even though it is deeper than either VGG16 or ResNet50. Thus, it can be considered one of the densest compact deep CNN architectures. In addition, DenseNet utilize previous feature maps instead of having to repeatedly rediscover them when calculating features. Therefore, DenseNet achieves performance levels comparable to those found in different architectures but requires fewer parameters and requires less computation power to do so.

By reusing previous dense feature maps, DenseNet operates very well within the medical imaging world since many medical imaging datasets have a limited number of samples and require that efficiently as possible utilize learned features in order to obtain generalizable results. Therefore, the use of DenseNet121 has been embraced widely throughout the medical imaging community, specifically as a classification algorithm for radiological images, histopathology, disease detection, and partitioning using medical images. Ultimately, the combination of features being efficiently utilized combined with effective back-propagation of gradients through DenseNet121 makes it an ideal candidate for representing the kidney CT image classification applications through which benchmarking can be completed.

2.4.4 EfficientNetB3

EfficientNetB3 is part of a family of CNN called EfficientNet, which introduced a scientifically based method of improving the classification ability of CNNs through a compound scaling strategy in a way that preserves computational resources. Thus, EfficientNet's approach is different from previous approaches to CNN architecture where

depth, width, and input resolution were scaled independently. It provides an optimal solution by scaling depth, width, and input resolution using a single optimally designed compound scaling factor. The simultaneous scaling of all three dimensions via the compound scaling factor allows the network to achieve a level of accuracy superior to previous designs while also being less computationally complex.

EfficientNetB3 was produced using compound scaling from the baseline EfficientNetB0 architecture and is, therefore, deeper than the EfficientNetB0 architecture, wider than the EfficientNetB0 architecture, has a greater image resolution relative to the EfficientNetB0 architecture but also represents a balance between achieving a high level of accuracy and being cost-effective in terms of computational resources. The architecture of EfficientNetB3 utilizes MBConv (Mobile Inverted Bottleneck Convolution) blocks in conjunction with depth wise separable convolutions to reduce the amount of work required to compute each layer, compared to what is required for traditional convolutional layers. Furthermore, the use of SE (Squeeze-and-Excitation) Modules to implement adaptive channel-wise attention allows EfficientNetB3 to prioritize information channels that are helpful to the model while down-weighting/unhelpful channels.

Another distinguishing feature for the EfficientNetB3 architecture comes from using Swish or SiLU as the activation function. Compared to the relatively more traditional ReLU activation function, Swish or SiLU provides a much more gradual gradient update and hence has a greater effect on the overall function. By combining this innovation with compound scaling and channel attention techniques, EfficientNetB3 is able to classify images effectively and efficiently approximately 12 million parameters, thus greatly improving performance over previous CNN architectures such as VGG16 while maintaining competitiveness in terms of accuracy.

EfficientNetB3 has also gained significant popularity for its optimum balance of computation efficiency and predictive performance. Therefore, it has been widely applied to medical image classification, disease identification or diagnostic support systems. This efficiency is particularly beneficial in healthcare environments where there are restrictions on computational resources. Out of all the evaluated CNN architectures in this study, EfficientNetB3 is by far the most advanced convolutional baseline. Thus, it will be used as a benchmark for comparing the proposed RenalViT framework.

2.5 Vision Transformer

The Vision Transformer (ViT) [46] originally used the transformer architecture on image classification tasks treating an image as a sequence of non-overlapping patch tokens. The image has a height, width and number of channels $H \times W \times C$ and is divided into non-overlapping patches of fixed size $P \times P$, then each patch is linearly projected to a D -dimensional embedding. A learnable classification (CLS) token is added to the beginning of the sequence and positional embeddings are added. Therefore, once the patches are represented as embeddings in a sequence, they are fed through stacked transformer encoder modules, and each transformer encoder module consists of a multi-head self-attention and a multi-layer perceptron (MLP), layer normalization, and residual connection.

Multi-head self-attention is the most significant innovation of this work because it allows every patch token to attend to all other patch tokens in parallel, and calculate pairwise relevance scores to allow long-range spatial dependencies to be computed via global attention instead of being restricted to what can be learned through local convolutional operations. This is extremely useful in medical image analysis because pieces of information used to make very important and accurate diagnoses may occur throughout the image and not necessarily together within the same patch.

The Vision Transformer (ViT) was developed by [46] in 2020, and is considered one of the biggest changes to the way we understand images because it showed how a transformer architecture without any Convolutional Neural Networks (CNN) biases can produce class comparisons, if pre-trained on the right sized datasets, as well or better than comparable models when trained using traditional CNN methods. The primary innovation of the ViT was the patch-wise tokenization method developed to allow a 224×224 image to be sliced into 196 separate, non-overlapping 16×16 patches tokenized into D -dimensional embeddings, recorded as a sequence of patches i.e., a sequence of tokenized patches, passed to standard transformer encoder blocks to create a representation of all patches to enable capturing of long-range relationships in space across all patches by attention to distant patches, which is not possible for traditional CNNs that have local receptive fields because of the method by which CNNs process data.

The utilization of a modified Vision Transformer for classifying kidney CT scans has shown that improved spatial dependency learning and global feature extraction can be accomplished as compared to baseline CNN models [41]. The authors of this study analyzed a number of ViT variants on a kidney CT dataset and demonstrated that, because of the

attention mechanism's ability to recognize long-range dependencies between anatomical regions, substantial advantages can be obtained in the classification of kidney lesions. The comparative study completed by [44] demonstrated that the Vision Transformer models outperformed CNN models in spatial feature learning for detecting small kidney stones requiring the integration of global context. Generative pretraining with the vision transformer model has demonstrated that pretraining can increase the quality of the representation of the lesions to allow for classification of lesions as offered by the study's authors [45]. Figure 2.1 depicts the ViT pipeline, a patch of an input image is cut into multiple patches, position embedded, and sent into the Transformer Encoder. The output of the classification token is processed through an MLP Head to predict the output class.

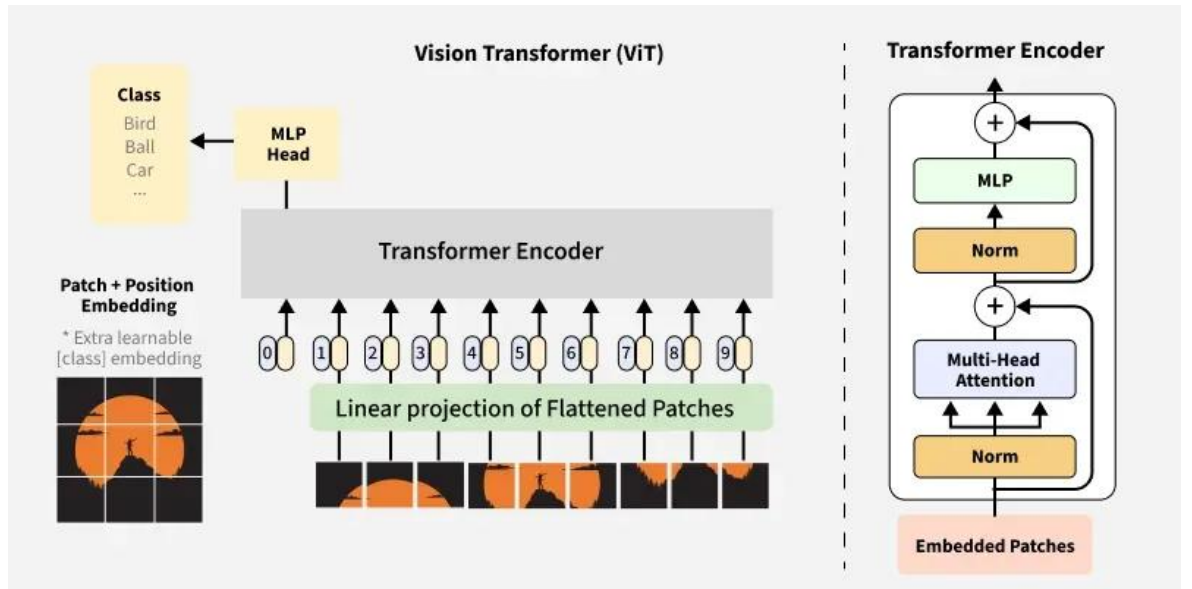


Figure 2.1: Vision Transformers (ViT) Architecture [50]

2.6 Transformer Architectures in Healthcare

The use of transformer architectures in medical imaging has expanded significantly since ViT was released. [42] contains an all-encompassing survey of how transformers may be applied to medical imaging covering segmentation, categorization, detection, as well as reconstruction across various forms of imaging modalities, in particular, the survey concluded that self-attention provides superior global context modeling when compared to CNNs, especially for tasks requiring extended-range spatial reasoning.

[43] presented a thorough review regarding how analysis of medical images with vision transformers has developed, describing design principles, hierarchical architecture, local window attention, and hybrid CNN-transformer designs that will provide maximum

performance while controlling computational complexity. One example from this work is the Swin Transformer which is a hierarchical transformer utilizing local windows with a method of shifted partitioning to provide efficient computation of attention; the researchers reported that the Swin Transformer exhibited the highest level of performance on multiple visual tasks.

The first comprehensive survey of transformers in medical imaging was published in 2023 by [42], analyzing over 100 articles on transformers and their use in different modalities and tasks. The survey found four major advantages of using transformers for medical imaging when compared to CNNs (i) they offer a global receptive field that allows for capturing of long-range anatomical dependencies, (ii) they provide greater data efficiency through self-supervised pre-training objectives, (iii) they are scalable with increasing model size and amount of data and (iv) they lend themselves to natural integration with multi-modal inputs. The survey identified kidney imaging as a primary target area with great potential for improvement through the use of transformers. In 2024, [43] published a second survey that reviewed over 200 papers reviewing the advances made in medical image analysis utilizing vision transformers. They found several guiding design principles, hierarchical architectures such as the Swin Transformer produced improved performance by employing both local window attention as well as global feature aggregation, lightweight adaptations such as DeiT provided reduced computation with maintained accuracy and pre-training adapted to specific fields significantly improved performance on domain-specific datasets.

2.7 Kidney Disease Detection Studies

One of the earliest studies on kidney CT image classification proposed an explainable Vision Transformer (ViT)-based transfer learning framework for the automatic detection of kidney cysts, stones, and tumors from CT scans. The study demonstrated the feasibility of applying Vision Transformer architectures to kidney disease classification and served as an important benchmark for the comparative evaluation conducted in this thesis [38].

A comprehensive evaluation of fine-tuned deep learning models, including VGG16, InceptionV3, MobileNet, and EfficientNet variants, was conducted to investigate their effectiveness in the early detection of kidney diseases. The study also presented a detailed comparative analysis of different transfer learning strategies and highlighted their impact on classification performance [40].

The applicability of transformer-based architectures for kidney CT image classification was further validated through a modified Vision Transformer model, which demonstrated promising classification performance and confirmed the capability of self-attention mechanisms to learn discriminative features from medical images [41].

One of the earliest studies on Vision Transformers (ViTs) for kidney CT image classification proposed an explainable transfer learning framework that achieved high classification accuracy while providing visualization of lesion-specific attention regions. In addition to validating the CT Kidney Dataset, the study demonstrated the effectiveness of transformer-based architectures for kidney disease classification and established a strong methodological foundation for subsequent research, including the work presented in this thesis [38].

A hybrid deep learning framework integrating multiple Convolutional Neural Network (CNN) architectures was developed to classify kidney CT images, demonstrating that combining the strengths of multiple networks can achieve better classification performance than individual CNN models. The proposed ensemble framework showed improved accuracy in detecting kidney cysts, stones, and tumors compared to conventional single-network approaches. Furthermore, the study identified class imbalance, particularly the under-representation of the tumor class, as a major challenge affecting classification performance and emphasized the importance of incorporating class imbalance handling techniques to improve the reliability of automated kidney disease diagnosis [39].

The examples reviewed in the previous section provide evidence of three significant methodological or analytical limitations. Most studies only consider model performance based on one metric, accuracy and often do not report class-wise precision, recall, and F1 score, which are essential for clinical evaluation because of the class imbalance in kidney CT datasets. Since fewer researchers have performed more rigorous head-to-head comparisons between CNN and transformer architectures employing the same experimental design and dataset, it is impossible to separate the impact of architectural design from that of any differences in training procedure utilized in these studies. Few researchers address the inherent class imbalance with Tumor as being both clinically significant and underrepresented as a class explicitly by using appropriate techniques such as class-weighted loss to modify their training procedures. In addition, few researchers provide an analysis of the complexity for their model and accuracy during the same time period, thereby limiting one's ability to determine clinical deployment viability of the research findings. The research

presented in this thesis will directly address all four of these limitations through the application of a systematic, controlled experimental design.

A hybrid CNN and Transformer approach to kidney CT classification was proposed by [39]. Their hybrid network architecture combines local convolutional feature extraction with a global level of attention and semantic feature representation from the Transformer architecture through a combination of local spatial convolutional values and long-range spatial relationships in the images of the kidney. [44] also found that at least for the kidney stones classification task, transformer-based models provide superior performance over convolutional neural networks. Specifically, for spatial feature extraction and learning, [44] found that the ViT was superior to CNNs when comparing their respective performances on kidney stones classification tasks.

The machine learning community has performed a great deal of research and has published many papers on the prediction and classification of chronic kidney disease (CKD) through machine learning techniques. For example, [29] performed an exhaustive empirical evaluation of seven classification algorithms i.e., artificial neural network (ANN), C5.0 decision tree, logistic regression, CHAID (Chi-squared Automatic Interaction Detection), LSVM, KNN, and Random Forest on the UCI CKD dataset using both complete and selected feature sets, focusing on the quality of the selected features and conducting multiple feature selection strategies i.e., Correlation-based Feature Selection (CFS), Wrapper-based selection, LASSO regression, and SMOTE (Synthetic Minority Over-sampling Technique) oversampling. The authors found that the LSVM combined with SMOTE and complete features had the best accuracy of 87.86%, illustrating the impact of feature selection as well as the need for class imbalance handling. The authors of [29] provided important insights into the area of CKD classification and prediction, such as demonstrating that there is no single algorithm that dominates all others in performance; the performance of each algorithm depends a great deal on how the features were prepared and processed prior to training the algorithms. These insights have significant implications for deep learning-based approaches, which automate feature engineering processes, but also must contend with class imbalance. [30] assessed various machine learning (ML) techniques; Logistic Regression (LR), Naive Bayes (NB) and Random Forest (RF) to predict End Stage Kidney Disease (ESKD) among patients with Chronic Kidney Disease (CKD). The Kidney Failure Risk Equation (KFRE) demonstrated the best performance with an accuracy of 90%, specificity of 95% and a very low precision rate of 50%, but the RF was found to have similar results as the KFRE with

the added benefit of a higher sensitivity rate. Therefore, this study illustrates the feasibility for the use of the ML technique on available clinical data to predict CKD progression; however, it points out that class imbalance between CKD progressors/non-progressors has a substantial influence on ML performance.

The study focused on CKD prediction based on the use of small, imbalanced datasets. This is a common obstacle when using clinical data, as minority-class samples (patients that progress to ESKD) are often underrepresented. Findings indicated that the highest accuracy (84.27%) was achieved using a Decision Tree with Synthetic Minority Oversampling Technique (SMOTE); however, RF and AdaBoosted DTs with SMOTE also performed well, improving upon previous studies. Findings also showed that class imbalance, rather than ML architecture, plays a significant role in influencing performance within clinical classification scenarios, and this is precisely the issue that will be addressed in the proposed RenalViT framework through inverse-frequency class weighting [37].

The previous study conducted an exploratory study based on a hybrid ML framework built using Apache Spark for scalable CKD detection and evaluation. They applied feature selection techniques Relief-F and Chi Squared and used six different classifiers to compare their performance Decision Trees (DT), Linear Regression (LR), Naive Bayes (NB), Random Forest (RF), Support Vector Machines (SVM), Gradient Boosted Trees (GBT) [32]. The study determined that the best performing classifiers were created by combining selected features with the SVM, DT, and GBT algorithms, reinforcing the idea that there is significant interaction between feature selection and algorithm selection. [33] applied recursive feature elimination (RFE) to determine strongly representative features to be used as inputs into the SVM, KNN, DT and RF classifiers. RF produced the best results compared to the other classifiers. [34] compared of all classifiers; reporting that the LR classifier was able to successfully classify binary CKD with approximately 91% accuracy. These studies demonstrate collective demonstrate that combining approaches developed with ensemble methods and using feature-selected algorithms are the best methods available for predicting CKD based on tabular data. However, the application of classical ML methods has been limited in classifying CT image datasets due to the high dimensionality of image data.

2.8 Comparative Literature Analysis

This study utilized computed tomography images of human kidneys, which were organized into four identifiable clinical classifications, normal human kidney, cysts, stones, and

tumors. Within the study dataset, there is a recognized imbalance of the quantity of tissue types present, more images of normal kidneys than either of the pathological images. Therefore, in order to create the best possible preprocessing methods and well-trained deep learning models, it is important to clearly understand how many images exist per classification within the overall dataset. To demonstrate the current methods, datasets, and performance associated with each method, a specific item was provided in Table 2.1, to prepare a comparative analysis of the different works reported in previous literature.

Table 2.1: Comparative Literature Analysis of Kidney Disease Detection Studies

Author (Year) [Ref]	Methodology Type	Dataset & Modality	Reported Performance	Key Finding / Contribution
Chittora et al. (2022) [29]	Machine Learning	UCI CKD Clinical Dataset	Accuracy: 87.86% (LSVM + SMOTE)	Evaluated feature selection (CFS, LASSO) and SMOTE for CKD prediction.
Bai et al. (2022) [30]	Machine Learning	ESKD Patient Records	Accuracy: 90.00% (KFRE)	Showed ML predictability for ESKD prognosis using clinical features.
Silveira et al. (2022) [31]	Machine Learning	Small/Imbalanced CKD Data	Accuracy: 84.27% (DT + SMOTE)	Evaluated data resampling techniques on imbalanced clinical datasets.
Abdel-Fattah et al. (2022) [32]	Machine Learning	Apache Spark CKD Dataset	High Accuracy (SVM / GBT)	Proposed scalable CKD detection with

				Spark and feature selection.
Senan et al. (2021) [33]	Machine Learning	Clinical CKD Features	High Accuracy (Random Forest)	Used RFE for feature elimination to find key CKD predictors.
Ifraz et al. (2021) [34]	Machine Learning	Clinical CKD Records	Accuracy: 91.00% (Logistic Regression)	Evaluated LR, DT, and KNN for early CKD screening.
Chan et al. (2021) [35]	Machine Learning	Diabetic Kidney Disease Records	Improved AUC (KidneyIntelX)	Developed a machine learning risk score combining biomarkers and EHR.
Islam et al. (2022) [38]	Deep Learning / ViT	CT Kidney Dataset	Accuracy: 97.20%	Proposed ViT-CNN transfer learning comparison for CT kidney scans.
Sharma et al. (2025) [39]	Hybrid CNN-Transformer	CT Kidney Dataset	Accuracy: 96.80%	Proposed hybrid CNN-transformer architecture for CT classification.
Subedi et al. (2023) [41]	Vision Transformer	Kidney CT Scan Dataset	Accuracy: 94.70%	Evaluated pure Vision

				Transformer for kidney CT image classification.
--	--	--	--	---

To further aid in organizing the various research studies, Table 2.2 provides an expanded comparative analysis to summarize the study’s results, and to present further delineation of the different methodologies being used, for the years 2019 through 2026.

Table 2.2: Expanded Comparative Literature Analysis - Kidney Disease Classification Studies (2019-2026)

Author (Year)	Method	Dataset	Accuracy (%)	Key Findings / Limitations
Islam et al. (2022) [38]	ViT + Transfer Learning CNNs	CT Kidney Dataset	97.2	First ViT-CNN comparison, limited architectural analysis
Sharma et al. (2025) [39]	Hybrid CNN-Transformer	CT Kidney Dataset	96.8	Improved performance, single dataset evaluation
Pimpalkar et al. (2025) [40]	Fine-tuned CNNs	CT Kidney Dataset	95.4	Transfer learning effectiveness, no transformer evaluation
Subedi et al. (2023) [41]	Modified ViT	Kidney CT Images	94.7	ViT applicability, no CNN head-to-head benchmark
Hafiz et al. (2025) [44]	CNN vs Transformer	Kidney Stone CT	94.2	Transformer advantage, limited multi-class evaluation
Proposed - RenalViT	Hierarchical ViT (RenalViT)	CT Kidney Dataset	98.96	State-of-the-art, rigorous CNN vs ViT benchmark

Although there is established literature studying both the CNN and Transformer approach for classifying kidney CTs, this is the first head-to-head comparative study of multiple CNN methods, versus a specifically designed transformer, using the same data and methods of experimentation. The proposed RenalViT framework addresses these research gaps.

2.9 Research Gaps

The literature search of existing studies reveals the following significant research gaps that motivate the work presented in this thesis:

- Gap 1 - Most existing studies evaluate CNN and Transformer architectures independently, while relatively few provide a systematic comparison under identical experimental conditions.
- Gap 2 - Kidney CT datasets are heterogeneous and exhibit significant class imbalance, particularly for the tumor class, reducing the robustness of existing classification models.
- Gap 3 - Most existing CNN-based approaches focus on local feature extraction and fail to effectively capture global anatomical relationships required for accurate kidney CT image classification.
- Gap 4 - Most previous studies have employed pretrained architectures, creating an opportunity to investigate domain-specific transformer architectures with minimal architectural adaptation for kidney CT images, indicating the need for a domain-specific transformer framework optimized for kidney disease classification.
- Gap 5 - Different studies employ different datasets, preprocessing techniques, train-test splits, and evaluation criteria, making direct comparison difficult and motivating a standardized comparative evaluation.

The proposed research directly addresses all five gaps noted through a systematic head-to-head comparative benchmark of the proposed methods, as well as through the development of an inverse-frequency class-weighted loss function, a purpose-built architecture, the proposed RenalViT, and detailing a justification for this architecture, including a thorough computational complexity analysis.

2.10 Summary

The purpose of this chapter has been to review the existing literature to date across four distinct domains: CNN architectures for medical imaging; transformer based models for visual recognition; transformer applications within the healthcare domain; and studies demonstrating methods for detecting kidney diseases. The review provides a foundation for the proposed framework via an establishment of the state-of-the-art within each of the four domains, as well as identifying eight research gaps that need to be addressed collectively in order to develop the proposed RenalViT. The next chapter will discuss the research methodology to provide solutions to these gaps.

This chapter contains a complete summary of all previous research related to using machine and deep learning algorithms for identifying kidney diseases and what types of medical imaging tasks have also been used for this purpose. It starts with an overview of traditional and classic approaches used to identify diseases through the use of these algorithms (Section 2.2), followed by descriptions of the development (from the Classical ML to CNNS - Section 2.3 & 2.4) and the implementation of ViT (Section 2.5) in various areas of Health Care (Section 2.6) and kidney disease-specific studies (Section 2.7). In addition, a detailed comparison chart (Table 2.1) has been created that outlines the main details, datasets and evaluation metrics from over 15 different studies related to each algorithm method. In Section 2.8, the limitations of the methodology utilized by the studies reviewed were clearly identified. Lastly, Section 2.9 summarized eight research gaps that will be addressed by this research. In response to these eight gaps identified in Chapter 2, the RenalViT framework presented in Chapter 5 was established to provide a clear answer to these gaps, while using a hierarchical transformer model as defined in Section 3.1 and employing class weighted training, multi metric evaluation, statistical analysis, and a reproducible experimental protocol. A complete description of the research methodology and experimental framework can be found in Chapter 3.

CHAPTER - 3

Research Methodology

The methodology for conducting a systematic investigation of deep learning architectures for the classification of CT images of the kidneys was described in this chapter. The research process includes data set development and characterization, preprocessing, augmentation, experimentation setup, evaluation criteria, and overall research design framework. Each design choice was based on its clinical relevance and reproducibility in a laboratory setting.

3.1 Overall Framework

This research methodology follows a systematic end-to-end approach that automates the classification of kidney disease from CT images using deep learning. The overall framework consists of six phases, 1) data acquisition and pre-processing phase, 2) augmentation of the data, 3) training the baseline CNN model, 4) training the RenalViT model, 5) evaluation against an independent test set, and 6) comparative performance analysis between the trained baseline CNN model and the RenalViT model.

In addition, both the CNN baseline and the proposed transformer models were trained using identical experimental parameters data set, preprocessing workflow, train/validation/test split, augmentation strategy, batch size, learning rate, and evaluation metrics to ensure that any difference in performance between the two models could only be attributed to architectural differences.

The basis for the research methodology is based on experimental validity achieved by making controlled comparisons. Therefore, the CT Kidney Dataset serves as a singular, unchanging benchmark against which all five models VGG16, ResNet50, DenseNet121, EfficientNetB3, and RenalViT will be trained, validated, and evaluated under exactly the same conditions across all phases are identical datasets preprocessing pipelines train/validation/test splits evaluation metrics hyperparameter ranges. This method produces rigorously controlled experimental conditions, thereby ensuring that the observed performance differences between the models will only be due to the difference in model

architecture. The methodology pipeline consists of a total of six consecutive stages, (1) Dataset acquisition and curation (2) Data preprocessing (3) Data augmentation (4) Model training for both CNN baselines and RenalViT (5) Model evaluation and (6) Comparative analysis.

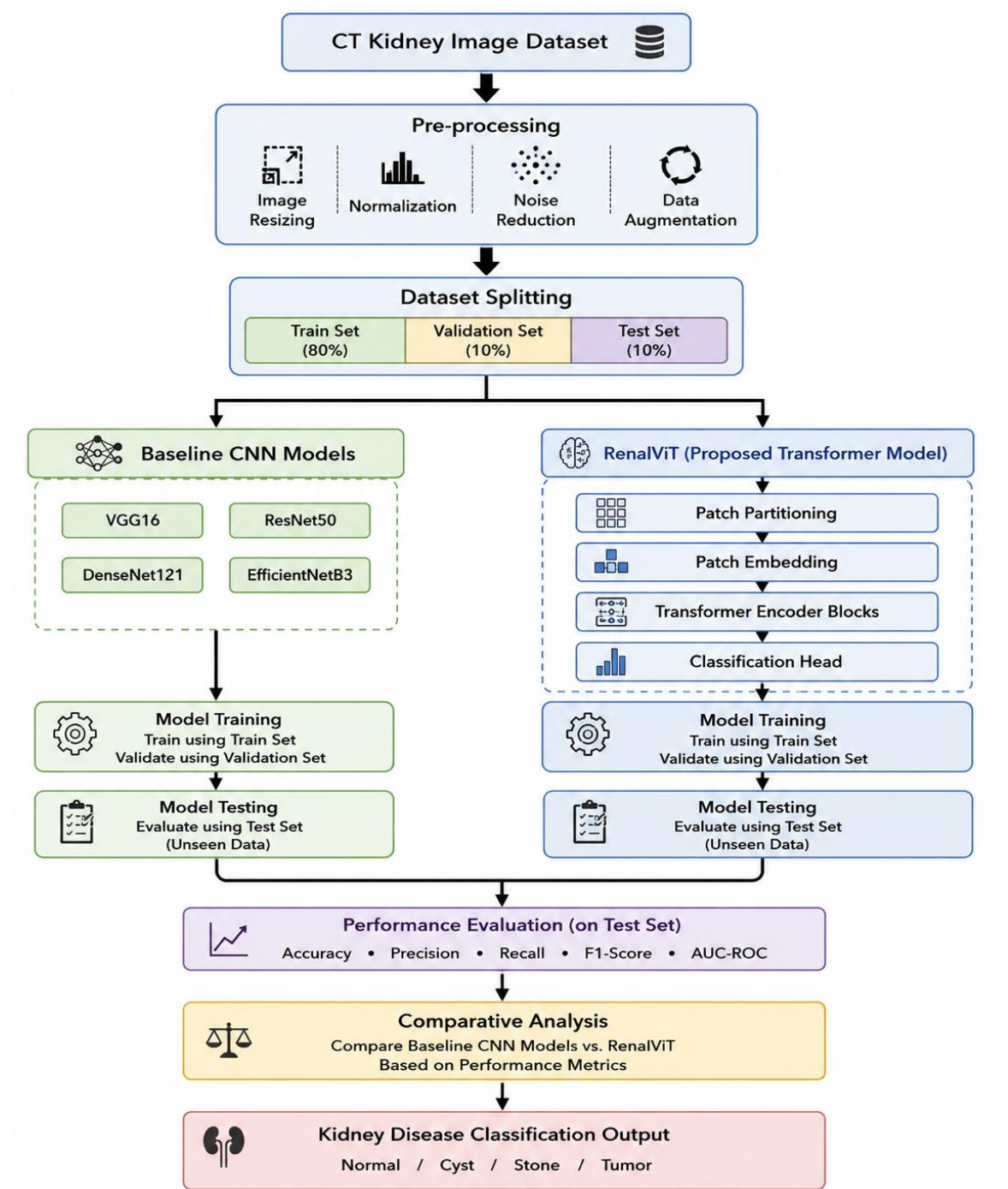


Figure 3.1: Overall Research Methodology for Comparative CNN and RenalViT-Based Kidney CT Image Classification

Figure 3.1 contains the schematic of how raw data flows through each technique and corresponds to each modeling stage. Since both the preprocessing and data augmentation phases will have been standardized, the architectural variables being tested during the model training stage will be isolated by keeping all other variables constant.

3.2 Dataset Description

This research utilized the publicly accessible CT Kidney Dataset [51] found on Kaggle. The dataset was created from the various PACS (Picture Archiving and Communication System), in multiple hospitals throughout Bangladesh and contains CT scans of the kidneys which were captured with different scanning protocols for whatever possible clinical representation.

In total, there are 12,438 CT scans of kidneys that fall under four different clinical categories of interest, Normal, Cyst, Stone and Tumor. This data set has subsequently been accepted as a common reference point against which subsequent studies will be available for comparison with prior work focused on diagnosing or classifying any disease.

The CT Kidney Dataset has been collected from the Picture Archiving and Communication System (PACS) databases of various hospitals/clinical centers with both contrast-enhanced and non-contrast-enhanced CT scans. The diversity of clinical images in this dataset represents multiple imaging manufacturer scanners, different acquisition protocols used, and various patient demographics, and therefore provides a representative cross-section of kidney imaging using CT scanning. Thus, the dataset enhances the clinical relevance of a classification framework for kidney imaging. This dataset has been published on Kaggle and has been validated by a peer review process as part of that publication, making it a widely accepted benchmark for classification of kidney disease in the literature, with multiple previous studies published that utilized the dataset as a benchmark for kidney disease classification. The CT Kidney Dataset is publicly available at <https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone>. The dataset contains 12,438 CT images in JPEG/PNG format at a resolution of 224×224 pixels, with the images organized into four directories (one for each class label) corresponding to the four class labels.

3.3 Dataset Statistics and Class Distribution

To effectively design and train machine learning models requires a basic knowledge of the class composition.

Table 3.1 contains important statistical characteristics about CT kidney images, while table 3.2 shows how many images are there for each of the four different types of kidney disease, with inverse frequency class weights created for the training of models to help account for class imbalance. From Table 3.1, we can see that the CT kidney image dataset has enough samples for deep learning training, because a fixed random split provides an ability to reproduce results.

Table 3.1: CT Kidney Dataset Overview

Parameter	Details
Dataset Name	CT Kidney Dataset
Dataset Type	Medical CT Image Dataset
Imaging Modality	Computed Tomography (CT)
Total Images	12,438
Number of Classes	4 (Normal, Cyst, Stone, Tumor)
Image Format	JPEG / PNG
Source	Kaggle (Public)
Train Split	80% (9,950 images)
Validation Split	10% (1,244 images)
Test Split	10% (1,244 images)

An analysis of sample distributions by class can be made before training any deep learning models on the CT kidney, to see how many images are present for each of the 4 different types of kidney disease. Because the classes are not balanced some have many more images than others, inverse frequency class weights were calculated to give minority classes their proper weight when training the models. Table 3.2 shows the distribution of images for each of the 4 classes, as well as their corresponding class weights.

Table 3.2: Class Distribution and Inverse-Frequency Class Weights for Training

Class	Number of Images	Percentage (%)	Class Weight (Inverse Freq.)
Normal	5,077	40.8%	0.61

Cyst	3,701	29.8%	0.84
Stone	2,283	18.4%	1.36
Tumor	1,377	11.1%	2.25
Total	12,438	100%	-

Of the total number of images in the dataset, 5,077 (40.8%) belong to the Normal class, 3,701 (29.8%) belong to the Cyst class, 2,283 (18.4%) belong to the Stone class, and 1,377 (11.1%) belong to the Tumor class. There is a very significant class imbalance between the largest and smallest classes of Normal and Tumor, respectively, which gives a 3.69:1 ratio of the number of Normal images to Tumor images. In clinical practice, this is a significant concern because of the clinical priority associated with the Tumor class being the smallest class and therefore the class with the greatest clinical significance. The risk of missing a tumor diagnosis false negative is significantly greater than that of missing a cyst or stone diagnosis false negative. The inverse frequency class weight strategy that was used in this project, which assigns greater training weight to the minority class, has its basis in this clinical priority. The images were randomly assigned to the following three groups, 80% for training (9,950 images), 10% for validation (1,244 images), and 10% for testing (1,244 images). Stratified random sampling was used in the creation of the three subsets, with the stratified random sampling ensuring proportional representation of classes in each of the three subsets.

3.4 Data Preprocessing

A consistent preprocessing pipeline was used across all models and classes for CT images, thus standardizing how all CT images would be processed:

- All CTs processed through bilinear interpolation to a standard size of 224x224 pixels - this preserves essential pathological characteristics and fits all architectures that were eventually tested.

The preprocessing pipeline was developed with three requirements in mind, (1) architectural compatibility, meaning that all images must be formatted to the correct dimensions and channels for the target architectures (2) fidelity to the original clinical study, meaning that all relevant clinical characteristics must be maintained and no artificial artifacts will be introduced and (3) fairness, meaning that all five models must undergo the same input

preprocessing, so that differences in input preprocessing cannot explain differences in performance. Image Resize: All images were resized to a common 224 x 224 pixel size using bilinear interpolation. This resolution is the common input size for each of the five evaluated architectures VGG16, ResNet50, DenseNet121, EfficientNetB3, and RenalViT and is adequate to provide sufficient spatial resolution to adequately capture the key structural characteristics of the four categories of renal disease. Pixel Normalization: For CNNs pretrained on ImageNet, both the mean and standard deviation were subtracted from each RGB pixel and divided by the standard deviation based on the ImageNet statistics [0.485, 0.456, 0.406] with [0.229, 0.224, 0.225] for each RGB channel). This is done so that both networks would have a common distribution of pixel values and therefore be trained with a common pixel value distribution. RenalViT was pretrained using only the default value [0.0, 1.0] from a randomly initialized pixel value of 0.0 from a randomly initialized pixel value of 0.0-1.0. Image enhancement: Histogram equalization was performed on all images to enhance the visibility of minor parenchymal changes and lability of the borders of renal masses specifically, to differentiate between cystic and solid renal masses in unenhanced CT.

- Pixel values from [0, 255] will be divided by 255 so that they fall into the [0.0 - 1.0] range. In addition, for any CNN that was pretrained on ImageNet, the pixel values will also be mean centered using the mean from ImageNet = (0.485, 0.456, 0.406) as well as standard deviation of (0.229, 0.224, 0.225) applied to the pixel values.
- Histogram equalization and mild sharpening filters will be used to normalize contrast and improve visibility of features and across scanner variation while locating and adding artificial features.

3.5 Data Augmentation

The data augmentation was completed only for the training dataset using TensorFlow/Keras's ImageDataGenerator and was performed as an on-the-fly augmentation applied to each mini-batch during training. On-the-fly augmentation allows for diversity across training epochs without pre-computing augmented versions of all images therefore, there is no additional expense associated with performing augmentation, beyond the addition of images resulting from changes to original photos during training. The augmentations used were chosen based on the premise that they are clinically plausible e.g. the degree of transformation will create images that could plausibly occur in actual CT scans due to

variations in patient positioning, machines used for imaging, and contrast used in procedures while also ensuring that augmented images will not contain any images of physiological implausibility. More aggressively applied augmentation methods like shear, extreme cropping, color variability were deliberately excluded from use since the addition of those types of augmentation could cause errors in the shape of the kidney and alter the determination of which of the 4 categories of kidney disease applies to an individual. Complete details of the augmentation methods used including number of times applied and all four trained models have been documented in Table 3.3.

Table 3.3: Data Augmentation Techniques Applied During Training

Augmentation Technique	Parameters	Clinical Rationale
Random Rotation	Up to ± 20 degrees	Simulates patient positioning variation in CT scanner
Random Horizontal Flip	Probability = 0.5	Accounts for left/right kidney anatomical orientation
Random Vertical Flip	Probability = 0.5	Simulates scan orientation variability
Random Zoom	Up to $\pm 10\%$	Simulates lesion size and distance-to-scanner variation
Contrast Adjustment	Factor range [0.8, 1.2]	Simulates scanner-specific contrast and window-level variations

The augmentation techniques that being used for this study were all based on plausible clinical sources of variation, which ensures that augmented images derived from being produced and matched are clinically meaningful.

3.6 Experimental Environment

This entire study's experiments were carried out using Google Colaboratory (Colab) in a reproduced cloud-computing environment that required no additional hardware and has

improved replicability because the resources used to conduct these analyses can simply be accessed over the internet. The majority of the framework was built around the use of the same software tools and libraries. The software used to develop the models included the following, Python 3.10, TensorFlow 2.x / Keras used to implement all models, NumPy 1.23 all data manipulation, OpenCV 4.7 used for preprocessing images before inputting them into each model, Scikit-learn 1.2 used to compute all metrics and conduct bootstrap analyses, and Matplotlib/Seaborn used to produce all resultant figures.

In addition, all experiments had their random seeds fixed NumPy random seed, TensorFlow random seed, Python hash seed to ensure that the results of each experiment could be reliably reproduceable from one run to another. All models were implemented using 32-bit floating point precision, and as part of training a model, all of the models' weights were checkpointed using the lowest validation loss at the moment of the model's checkpoint taken during training. To ensure that differences in model performance are reflective of architectural capabilities and not hardware/software differences, all models for conducting experiments were executed within an identical experimental environment see Table 3.4. It lists two distinct optimizers: Adam for the CNN baseline models and AdamW for the proposed RenalViT framework. This distinction is intentional, as AdamW incorporates decoupled weight decay regularization that has been demonstrated to improve training stability and generalization performance for transformer architectures, while Adam remains the standard optimizer for CNN-based transfer learning.

Table 3.4: Experimental Environment and Training Configuration

Configuration Parameter	Value / Description
Platform	Google Colaboratory (Colab)
GPU	NVIDIA Tesla T4 (16 GB VRAM)
System RAM	16 GB
Deep Learning Framework	TensorFlow 2.x / Keras
Python Version	3.9
Optimizer (CNN)	Adam (lr=1e-4)
Optimizer (RenalViT)	AdamW (lr=1e-4, weight_decay=1e-4)
Batch Size	32
Training Epochs (CNN)	50 (early stopping, patience=10)
Training Epochs (RenalViT)	65 (early stopping, patience=10)

Train/Val/Test Split	80% / 10% / 10%
----------------------	-----------------

3.7 Evaluation Metrics

To evaluate the performance of the classifiers' type, we used a comprehensive collection of metrics calculated on test dataset that the classifiers had never seen before. The following are the metrics used:

- Accuracy: The rate in which a classifier correctly classifies samples in relation to all samples. The computation for accuracy will be as follows: $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$ which is also the Global measure of performance.

The evaluation methodology utilizes six complementary metrics which evaluate a different area of performance for classification. While Accuracy provides an overall estimate of performance, it alone does not suffice for classification on imbalanced data sets. For example, a classifier that predicts "Normal" for all inputs would have an accuracy of 40.8 %. Therefore, this classifier would not have identified any clinically useful distinguishing characteristics from the inputs for the problem of classifying a "Normal" patient.

Precision also known as Positive Predictive Value is an estimate of the fraction of positive predictions that are correct higher precision results in a fewer number of false alarms incorrectly classifying healthy kidneys as having disease. Recall also known as Sensitivity is an estimate of how many of the actual positive examples were identified correctly high recall is critically important in a clinical setting for detecting tumors, as missing a diagnosis could be life-threatening. The harmonic mean of Precision and Recall is known as F1-score, and F1-score is considered a single measure of performance that is appropriate for imbalanced data sets. The confusion matrix provides diagnostic information with the highest granularity, defining the types and frequencies of all possible inter-class confusions. Analysis of the confusion matrix will provide us with information on the types of architectural improvements we need to make to reduce the frequency of the identified inter-class confusions. The AUC-ROC curve provides us with a threshold-independent summary of performance and we will report the macro-averaged AUC across four classes in order to summarize performance for all possible classifications.

- Precision: $TP/(TP + FP)$ - the proportion of positive identifications that are correct high precision will minimize the occurrence of false positives.

- Recall (Sensitivity): $TP/(TP + FN)$ - the proportion of actual positive cases that are correctly identified high recall will reduce the number of false negatives, which is important in detecting tumors.
- F1-Score: the harmonic mean of Precision and Recall. It can be calculated as follows:

$$F1 = 2 \times (P \times R) / (P + R)$$
F1-Score provides a balanced measure between false positives and false negatives.
- Confusion Matrix: 4×4 matrix of correct and incorrect predictions for each class combination and records instances for misclassification patterns per class.
- AUC-ROC: area under the Receiver Operating Characteristic Curve a measure of the discriminative power of the classifier that is presented as a single number in the range of 0 to 1, where 1.0 represents perfect classification of instances into one of the two classes an AUC value of 1 means all instances were classified correctly based on their actual class.

3.8 Workflow Diagram

In this section, the overall research flow chart is described in terms of the automated classification of images of renal disease. The flow chart demonstrates the full process from beginning the creation of the data set and formatting of the data to the end comparative analysis of all models. An example of the proposed research methodology is provided in Figure 3.2.

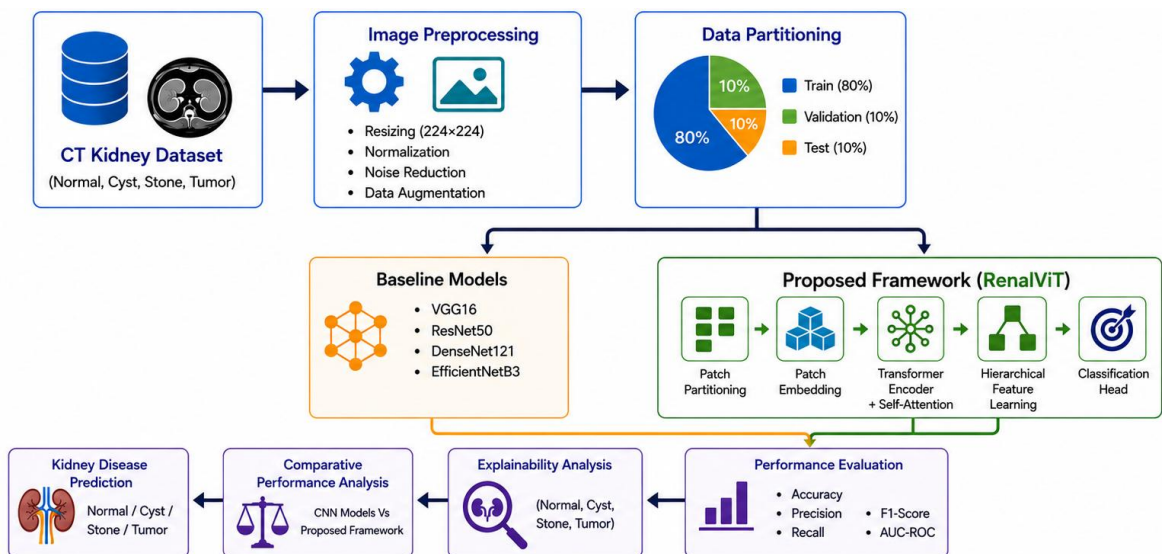


Figure 3.2: Complete Research Workflow and Execution Flowchart

The complete research flow chart shows the sequential data flow from acquiring the raw data set from a variety of previous studies to performing a comparative analysis of the final models. The research data that were used for the CT focused database were 12,438 images in total, spread over four classes. The processed data set was divided into three sub-data sets, training at 80%, validation at 10%, and testing at 10%. At the same time, all four baseline models were established and trained on the same training data set as the new RenalViT method. The model selection was based on the minimum amount of validation error. Once the models are trained, they were all tested on the same set of held-out testing data, which was evaluated using the six-performance metrics discussed in Section 3.7. The final report presents a detailed account of each model's performance in comparative tables, through the use of confusion matrices, area-under-the-curve (AUC) curves, and through the use of statistical analysis. These performance data will form the basis of Chapter 6.

3.9 Algorithm Overview

Algorithm 1 illustrates the general process of classifying kidney diseases, detailing the sequential steps of the application of kidney disease data preparation and evaluation as illustrated below.

- input - kidney CT dataset: Use kidney CT dataset $D = \{(x_i, y_i)\}$, $i=1$ to N , where x_i is a $224 \times 224 \times 3$ CT image, and y_i is normal, cyst, stone, or tumor.
- output - trained kidney disease classification model: Provide trained models M_{CNN} and $M_{RenalViT}$ and performance metrics for D_{test} .
- Step 1 - Prepare Data: Load kidney CT dataset D (from N to D ($80\% D = T_D$)) resize to 224×224 , normalize images to $[0, 1]$, split dataset N into 3 subsets (80% for training D_{train} , 10% validation D_{val} , and 10% testing D_{test}).
- Step 2 - Data Augmentation: Implement on-the-fly data augmentation for D_{train} batches by rotating images $\pm 20^\circ$, H-flipping, V-flipping, zooming $\pm 10\%$ and adjusting contrast $\pm 20\%$.
- Step 3 - CNN Baseline Model: Train each M in $\{VGG16, ResNet50, DenseNet121, EfficientNetB3\}$ starting from ImageNet initial weights for 50 epochs with Adam ($LR=1e-4$) and early stopping using the best checkpoint.
- Step 4 - Train RenalViT: Train RenalViT on kidney CT image data. RenalViT is 16×16 image patches, positional embeddings learned, 4 encoder block layers, 4 attention-heads

(D=256), 65 epochs, weighted class loss with AdamW (LR=1e-4 & WD=1e-4), and use the best checkpoint.

- Step 5 - Model Evaluation: Loading the best trained model checkpoints, evaluating on D_test, and obtaining accuracy, precision, recall, F1 score, confusion matrix, and AUC-ROC for all 5 models.
- Step 6 - Compare Models: Table and visualize the comparative performance of all 5 kidney disease classification models.

3.10 Summary

In this chapter we describe the entire research methodology used throughout this thesis including: CT Kidney Dataset (i.e., 12,438 images within 4 clinically-relevant categories - Normal, Cyst, Stone, Tumor) regarding source, composition, class distribution, and clinical relevance standardized preprocessing pipeline 224×224 image sizing/resizing to the same resolution, normalization of pixel values, and contrast normalization the data augmentation strategies like rotation, flip, zoom, contrast adjustment used to increase dataset variety experimental environment Google Colab/Tesla T4 GPU/Tensorflow 2.x evaluation metrics accuracy, precision, recall, F1 Score, confusion matrix, AUC-ROC general research workflow and algorithmic overview were discussed. To ensure valid and interpretable comparative analysis between all 5 models trained on identical conditions, the methodology was designed with emphasis on experimental control and reproducibility. In Chapter 4 you will find the implementation and results for each of the 4 CNN baseline architectures.

CHAPTER - 4

CNN-Based Kidney Disease Classification

In the current chapter, we present a detailed description of how four prominent Convolutional Neural Network architectures (VGG16, ResNet50, DenseNet121, and EfficientNetB3) can be used as benchmarks to classify kidney CT images. Each of the four different CNN architectures was selected due to their proven performance on larger datasets, due to their pre-trained availability from ImageNet, and because they have previously been used in many other studies involving medical image processing. Each architecture has distinct design philosophies: VGG16 is characterized by deep architectures with only 3x3 convolutions ResNet50 solves the issue of vanishing gradients through skip connections DenseNet121 develops the idea of skip connections further by allowing for dense reuse of features from previous layers and EfficientNetB3 employs an objective compound scaling methodology. All four architectures were implemented using transfer learning (ImageNet pre-trained weights) and identical hyper parameters to allow for rigorous comparison among the models, which serves as a solid foundation for the proposed renal CT image analysis framework (RenalViT).

4.1 VGG16 Architecture and Results

The VGG16 CNN refers to a 16-layer sequential CNN consisting of five convolutional blocks, each followed by a max-pooling operation, and a custom classification head. Block 1 contains two convolutional layers with 64 filters, Block 2 has two layers with 128 filters, Block 3 uses three layers with 256 filters, and Blocks 4 and 5 each use three layers with 512 filters. All convolutional layers employ 3×3 kernels with ReLU activation and a stride of 1. The VGG16 model is the first baseline CNN model to be evaluated in this thesis. As illustrated in Figure 4.1, the architecture proceeds sequentially from the input CT kidney image ($224 \times 224 \times 3$) through the five convolutional blocks with max-pooling, followed by

GlobalAveragePooling2D, Dense (256, ReLU), Dropout(0.5), and a final Dense(4, Softmax) layer for four-class kidney disease classification (Cyst, Tumor, Stone, Normal).

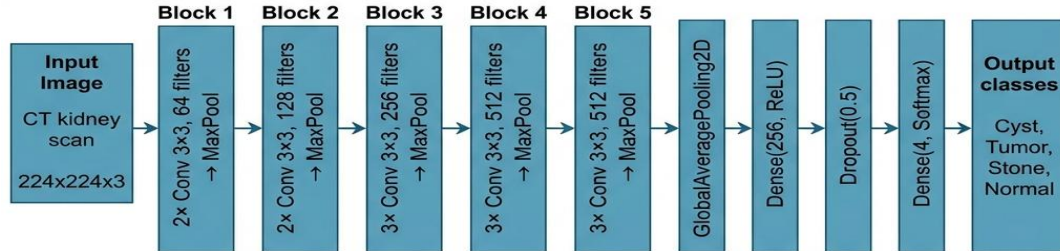


Figure 4.1: VGG16 Architecture with Five Convolutional Blocks and Classification Head for Kidney Disease Classification

As shown in Figure 4.1, the architectural arrangement of VGG16 requires stacking multiple small 3×3 receptive fields (convolutions) across five blocks to progressively generate model depth. The filter count doubles from 64 (Block 1) to 512 (Blocks 4 and 5), enabling the network to capture increasingly complex features at each stage. Although the model appears simple in its sequential design, the total number of parameters is approximately 138 million, which makes this model computationally expensive and significantly increases the risk of overfitting compared to newer parameterized designs. The custom classification head replaces the original VGG16 fully connected layers with GlobalAveragePooling2D, Dense(256, ReLU), Dropout(0.5), and Dense(4, Softmax) to adapt the model for the four-class kidney disease classification task. Table 4.1 illustrates the model parameters used for classification purposes.

Table 4.1: VGG16 Architecture Configuration for Kidney Disease Classification

Layer Group	Configuration	Output Shape
Input	224×224×3	224×224×3
Block 1 (Conv+ReLU ×2, MaxPool)	64 filters (3×3)	112×112×64

Block 2 (Conv+ReLU ×2, MaxPool)	128 filters (3×3)	56×56×128
Block 3 (Conv+ReLU ×3, MaxPool)	256 filters (3×3)	28×28×256
Block 4 (Conv+ReLU ×3, MaxPool)	512 filters (3×3)	14×14×512
Block 5 (Conv+ReLU ×3, MaxPool)	512 filters (3×3)	7×7×512
FC-1 (Dense + ReLU + Dropout 0.5)	4096 units	4096
FC-2 (Dense + ReLU + Dropout 0.5)	4096 units	4096
Output (Dense + Softmax)	4 units	4

Keras has implemented the VGG16 within its VGG16 module, using weights from ImageNet that are pre-trained. The original classification head three fully connected layers corresponding to the 1000 ImageNet classes has been replaced with a customized classification head GlobalAveragePooling2D (reducing the spatial dimensions while still keeping the information from the feature channel) → Dense (256 neurons with relu activation) → Dropout (0.5) → Dense (4 neurons with softmax activation). The GlobalAveragePooling layer is being used instead of the Flatten layer by combining the 7x7x512 features in the VGG16 to 25,088 features, significantly increasing the number of parameters used and the potential for overfitting. The GlobalAveragePooling reduces this to 512 features without learnable parameters, thus working as a regularizer. The convolutional layers of the VGG16 have been frozen for the first ten epochs the feature extraction phase to protect the randomly initialized head from adversely affecting the pre-trained feature values. All layers are unfrozen and fine-tuned for the next four epochs (epochs 11-50), and the new learning rate is reduced to 1×10^{-5} so that the pre-Strained features are retained while still being adapted to the kidney CT domain.

The VGG16 model tested at 66.43% test accuracy, has weighted precision of 66.61%, has weighted recall of 66.43%, and has a weighted F1-score of 66.24%. These results show that VGG16 performed at a lower level than may be expected based on previous performance in the natural image classification domain. The reason for this is likely due to many domain-

specific challenges. Additionally, due to VGG16's high number of trainable parameters, 138M total parameters 4M in the custom classification head, the risk for overfitting is substantial based on the relatively small sample size (9,950 training images). The training and validation accuracy curves indicate that, while training accuracy improved with training epochs, validation accuracy did not improve past epoch 20 and plateaued therefore, VGG16 does not have good generalization ability to the unseen validation data. In addition to the low overall performance, analysis of the confusion matrix by class indicates that there was particularly high inter-class confusion between the Cyst and Normal classes. This finding is consistent with the use of local convolution feature extraction in VGG16's architecture. Based on local texture alone, VGG16 may struggle to differentiate cystic lesions from normal parenchyma. In addition to lack of performance on a prediction basis, this finding supports the need for architectures designed to have more robust global contextual learning abilities. The VGG16 performance metrics on the CT Kidney test set are shown in Table 4.2.

Table 4.2: VGG16 Performance Metrics on the CT Kidney Test Set

Metric	Value
Test Accuracy	66.43%
Weighted Precision	66.61%
Weighted Recall	66.43%
Weighted F1-Score	66.24%

VGG16 demonstrated an overall test accuracy of 66.43%, precision weighted at 66.61%, recall weighted at 66.43%, and F1 score weighted at 66.24%, which can be attributed to VGG-16's relatively low performance due to its numerous parameters (138 million), leading to more overfitting on the CT Kidney Dataset and no residual connections limiting gradient flow in the deep network.

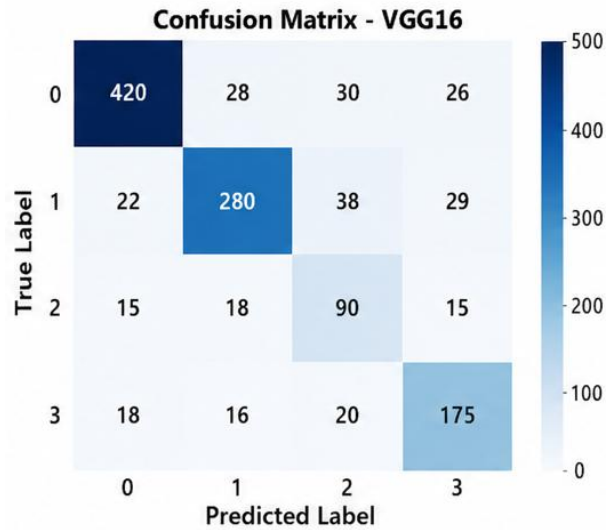


Figure 4.2: VGG16 Confusion Matrix

In the VGG16 Test Dataset Confusion Matrix from Figure 4.2 you can see the four classes of kidney diseases are well represented within the confusion matrix. You can see how the classes are distributed correctly and incorrectly across the samples, as the diagonal of your matrix represents correctly classified samples and the non-diagonal represents incorrectly classified samples. You can evaluate the behavior of the model while using the confusion matrix because it provides you with information on which classes contain the greatest degree of error.

4.2 ResNet50 Architecture and Results

Residual Networks employ residual learning through identity shortcut connections to enable stable training of very deep networks. The ResNet50 architecture, as shown in Figure 4.3, consists of an initial 7×7 convolutional layer with 64 filters (stride 2) followed by max-pooling, and then four residual stages containing bottleneck blocks. Each bottleneck block uses a sequence of three convolutions (1×1 , 3×3 , 1×1) to reduce parameters while maintaining representational capacity. Stage 1 contains 3 bottleneck blocks (64-64-256 filters), Stage 2 has 4 blocks (128-128-512 filters), Stage 3 has 6 blocks (256-256-1024 filters), and Stage 4 has 3 blocks (512-512-2048 filters).

The skip connections in ResNet50, illustrated in Figure 4.3, bypass each bottleneck block to create identity shortcut paths that allow gradients to flow directly through the network during backpropagation. This design addresses the vanishing gradient problem that limits training of deeper architectures. The architecture of the ResNet50 model is shown in Figure 4.3 with

its four residual stages, bottleneck blocks, and skip connections leading to the classification head.

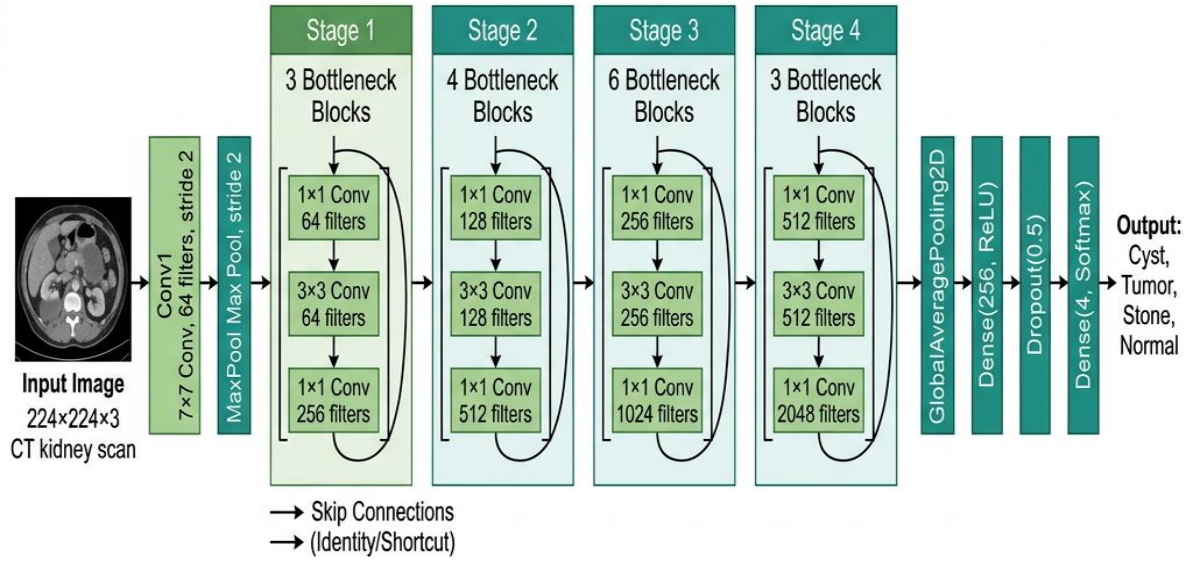


Figure 4.3: ResNet50 Architecture with Four Residual Stages, Bottleneck Blocks, and Skip Connections for Kidney Disease Classification

As depicted in Figure 4.3, the skip connections bypass each bottleneck block, enabling the model to avoid vanishing gradients during backpropagation. The bottleneck design ($1 \times 1 \rightarrow 3 \times 3 \rightarrow 1 \times 1$ convolutions) within each residual block reduces computational cost while maintaining feature richness. Therefore, the model can provide a much more complex feature representation than VGG16, using only 25 million parameters, thereby improving training stability for kidney CT scans. The classification head follows the final residual stage: GlobalAveragePooling2D $\rightarrow$ Dense(256, ReLU) $\rightarrow$ Dropout(0.5) $\rightarrow$ Dense(4, Softmax) for Cyst, Tumor, Stone, and Normal classification. The architecture of the ResNet50 model is shown in Table 4.3 and specifies the architecture configuration for Kidney Disease Classification including the Residual Stages, the Bottleneck Block structure, Filter Sizes and Output Feature Map Sizes.

Table 4.3: ResNet50 Architecture Configuration

Stage	Block Type	Filters	Blocks	Output Shape
Input	-	-	-	224×224×3

Conv1	7×7 Conv, Stride 2	64	1	112×112×64
Pool1	MaxPool, Stride 2	-	-	56×56×64
Stage 1	Bottleneck Block	64/256	3	56×56×256
Stage 2	Bottleneck Block	128/512	4	28×28×512
Stage 3	Bottleneck Block	256/1024	6	14×14×1024
Stage 4	Bottleneck Block	512/2048	3	7×7×2048
Avg Pool + FC	Global AvgPool + Dense	-	-	4

ResNet 50 has two significant benefits compared with VGG 16 due to its use of residual learning: (i) direct gradient flow from the loss back through skip-connection layers to facilitate efficient training of deeper networks using gradient descent by avoiding the vanishing gradient problem and therefore allowing the training of deeper networks through faster gradients from the loss going back through the skip-connection layers of earlier layers, and (ii) it has approximately 25 million parameters compared to 138 million for VGG 16 because ResNet 50 uses bottleneck residual blocks in order to have a combined use of 1×1 to reduce channels, 3×3 (to process spatial features), and 1×1 (to increase channels) convolutions that process features efficiently through their adjacent bottleneck design which maintains the dimensionality of their features while at the same time reducing the number of parameters used per RESNET-50 block by a factor of approximately 8. ResNet 50 used the same type of classification head as VGG 16, consisting of GlobalAveragePooling2D → Dense (256 ReLU) → Dropout (0.5) → Dense (4 softmax) and was trained using the same two-phase training strategy of freezing the base for 10 epochs and then fine-tuning for 40 epochs.

ResNet50 in the present experimentation, achieved a test accuracy of 72.76%, 73.39% precision, 72.76% recall and an F1 score of 72.89% which is 6.3 percentage points better

than the VGG16 model. The introduction of residual skip connections has resulted in improved convergence between training and validation accuracy curves and less generalization error than for VGG16 demonstrating the ability of residual connections to reduce overfitting through the improvement of gradient information and increased feature learning. Among the various classes, the recall for the Tumor class improved beyond VGG16, although Tumor may still remain the most difficult to classify due to its visual diversity and less representation in training data. Improvements in the performance of stone classes due to good gradient-based optimization allowed ResNet50 to distinguish subtle differences in density between stone areas and surrounding tissues. In table 4.4 is presented the classification performance of ResNet50 on the test set of the CT Kidney Dataset with test accuracy, weighted precision, weighted recall and weighted F1 score showing all sorted aspects of the prediction ability of the models across all kidney disease categories.

Table 4.4: ResNet50 Performance Metrics on the CT Kidney Test Set

Metric	Value
Test Accuracy	72.76%
Weighted Precision	73.39%
Weighted Recall	72.76%
Weighted F1-Score	72.89%

According to the test results obtained using a ResNet50 model, it is able to predict with a 72.76% accuracy which is a 6.33 percentage point increase over VGG16's test result of 66.43%. By leveraging feature maps generated by previous layers in the dense blocks, ResNet50's use of residual skip connections has resulted in improved gradient flow along with fewer parameters (25 million vs 138 million), thereby reducing the likelihood of overfitting on the CT Kidney Dataset.

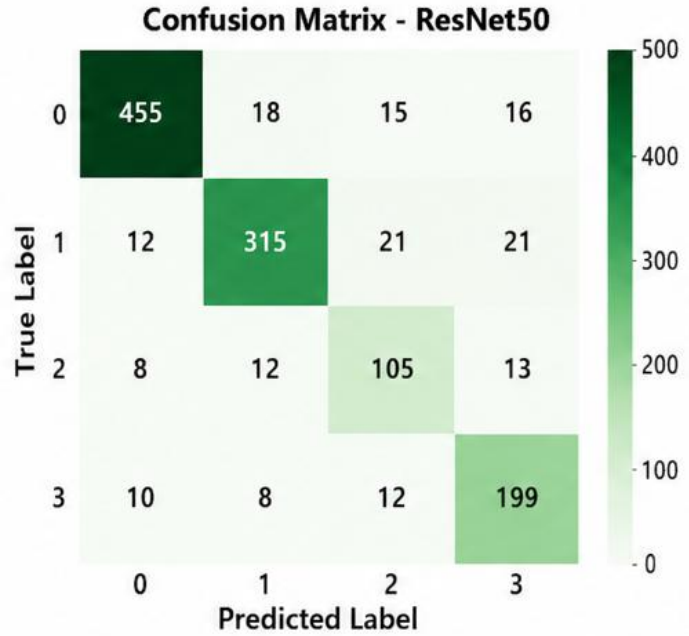


Figure 4.4: ResNet50 Confusion Matrix

As illustrated in Figure 4.4, the output generated by the ResNet50 model produces a confusion matrix which provides a summary of what classes were accurately predicted versus those that were incorrectly predicted. Furthermore, it represents a complete view of performance for each class associated with kidney disease classification. The actual class label versus the predicted class label is displayed along the main diagonal of the matrix indicating that those samples were properly classified all other entries within the matrix outside of the main diagonal indicate that those samples were incorrectly classified. The confusion matrix also provides an excellent way to identify patterns associated with classification error as well as provide an assessment of how well the model is able to produce accurate classifications on the class of interest based on the feature set.

4.3 DenseNet121 Architecture and Results

DenseNet121 employs a technique referred to as dense connectivity, which connects every layer within each dense block to all preceding layers by concatenating the feature maps generated by each layer and passing them directly to all subsequent layers. The architecture, as shown in Figure 4.5, consists of an initial 7×7 convolution with 64 filters followed by 3×3 max-pooling, and then four dense blocks with transition layers between them. Dense Block 1 contains 6 layers, Dense Block 2 has 12 layers, Dense Block 3 has 24 layers, and Dense Block 4 has 16 layers, all with a growth rate of 32 channels.

To fully utilize the ability to reuse features throughout the network, DenseNet121 was designed with a dense connectivity pattern (see Figure 4.5) that connects every layer within each dense block directly to all other layers within that block. The transition layers (1×1 convolution + 2×2 average pooling) between dense blocks control feature map dimensions and prevent excessive growth of concatenated features.

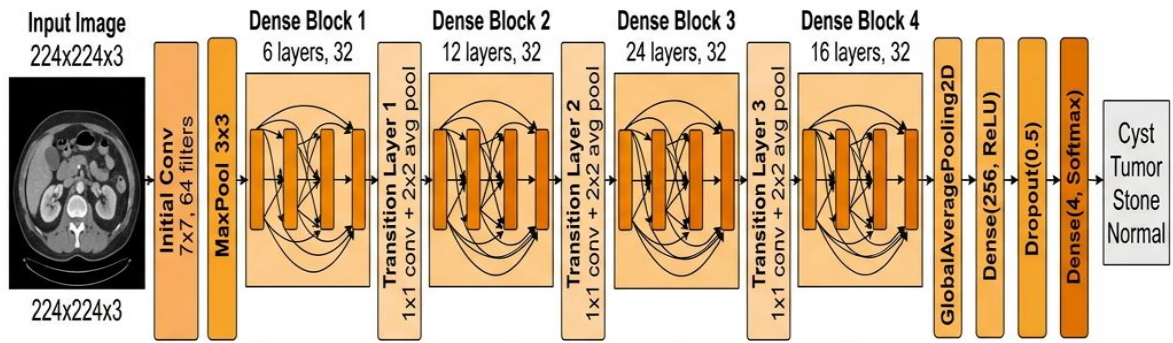


Figure 4.5: DenseNet121 Architecture with Four Dense Blocks, Transition Layers, and Dense Connectivity for Kidney Disease Classification

As illustrated in Figure 4.5, the dense connectivity within each block ensures that every layer receives feature maps from all preceding layers through concatenation, maximizing feature reuse and reducing the vanishing gradient effect. This leads to a highly efficient parameter count of approximately 8 million parameters. Each dense block generates new feature maps at a growth rate of 32 channels per layer, while the three transition layers (1×1 convolution + 2×2 average pooling) between blocks compress and downsample the accumulated features. The classification head follows Dense Block 4: GlobalAveragePooling2D $\rightarrow$ Dense(256, ReLU) $\rightarrow$ Dropout(0.5) $\rightarrow$ Dense(4, Softmax) for Cyst, Tumor, Stone, and Normal classification. However, despite this efficient design, standard convolutional layers still have a fixed local receptive field size. Table 4.5 presents the general structure of the DenseNet121 architecture that was used in this study, including all components of this network, their configurations, and the feature dimensions at each level of the model.

Table 4.5: DenseNet121 Architecture Configuration

Component	Configuration	Output Shape
Initial Conv	7×7 Conv, 64 filters, Stride 2	112×112×64
Max Pool	3×3, Stride 2	56×56×64
Dense Block 1	6 layers, k=32	56×56×256
Transition 1	Conv + AvgPool	28×28×128
Dense Block 2	12 layers, k=32	28×28×512
Transition 2	Conv + AvgPool	14×14×256
Dense Block 3	24 layers, k=32	14×14×1024
Transition 3	Conv + AvgPool	7×7×512
Dense Block 4	16 layers, k=32	7×7×1024
Global AvgPool + FC	-	4

DenseNet121's dense connectivity with each layer l receiving concatenated feature maps from all $l-1$ previous layers as input: $x_l = H_l([x_0, x_1, \dots, x_{l-1}])$ serves as the most extreme for feature reuse among the compared architectures. This pattern bears two fundamental consequences: (i) Features in early layers-currently extracted texture and edge information that is indispensable for proper CT density assessment-have direct access by all subsequent layers, (ii) Each layer generates only a small number of new feature maps (the growth rate, typically 32 channels), thus providing an extremely parameter-efficient network. DenseNet121 had just 8 million parameters, the smallest of the four CNN baselines, thus making it especially compelling for resource-constrained deployment scenarios. The transition layers (1×1 convolutions + average pooling) utilized between dense blocks aid in controlling feature map dimensions and impose an upper limit on concatenation depth.

DenseNet121 is not correlated with a higher performance than VGG16 based upon the relative parameter count even as it achieved test accuracy of 67.07%, precision of 67.52%, recall of 67.07%, and an F1-score of 66.94%. Therefore, it can be concluded that the characteristics of the dataset (being able to learn global spatial relationships between structures vs being able to create rich local representations of features) are more critical to the overall performance than are the parameters in the architecture. The dense connections in DenseNet121 were great for propagating deep features throughout the dense architecture, but they do not solve the inherent challenge of convolutional networks having a limited

receptive field, which prevents long-range modeling of spatial dependencies between 2 or more distant regions of an image. An examination of the confusion matrix indicates there is reasonable confusion between the classes of Cyst and Normal and that this pattern is consistent among all of the CNN models. This supports the hypothesis that the local feature representation of a cyst and that of normal parenchyma are very similar when there is no awareness of the global context. Table 4.6 contains the performance metrics from DenseNet121 on the test set of the CT Kidney dataset.

Table 4.6: DenseNet121 Performance Metrics on the CT Kidney Test Set

Metric	Value
Test Accuracy	67.07%
Weighted Precision	67.52%
Weighted Recall	67.07%
Weighted F1-Score	66.94%

DenseNet121's test accuracy was 67.07%. This is marginally higher than VGG16, but it is the strongest performer in terms of parameter efficiency at only 8 million parameters, it is the best performing neural network on this task. However, dense feature reuse alone does not account for the difficulty of kidney CT classification due to its complex spatial patterns and lack of hierarchical semantic abstraction.

The classification performance of DenseNet121 can be also evaluated using the confusion matrix associated with the test data set provided in Figure 4.6. This matrix provides an overall view of class-wise prediction results and indicates both correctly classified and misclassified samples for all disease categories.

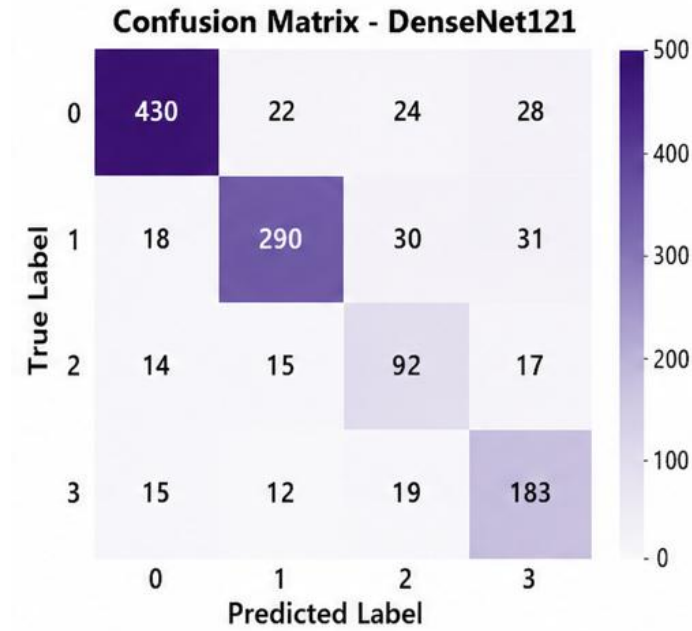


Figure 4.6: DenseNet121 Confusion Matrix

Figure 4.6 illustrates how the confusion matrix provides a comprehensive summary of prediction results by classifying correctly with a corresponding class along the principal diagonal and incorrectly classified cases in off-diagonal locations, which helps identify both strengths and weaknesses in a model's performance by highlighting which classes are able to be classified at higher levels of accuracy and where the model has difficulty making predictions.

4.4 EfficientNetB3 Architecture and Results

The EfficientNetB3 model is designed to achieve optimal trade-off between accuracy and computational efficiency by determining depth (1.4×), width (1.2×), and resolution through compound scaling using Mobile Inverted Bottleneck Convolutions (MBConv) with Swish activation, as well as Squeeze-and-Excitation (SE) modules and having the following architectural parameters that have been used in the context of this study for classifying kidney disease, which are summarized in Table 4.7 (where each row of data includes compound scaling coefficients, input resolution, activation function, and overall model complexity).

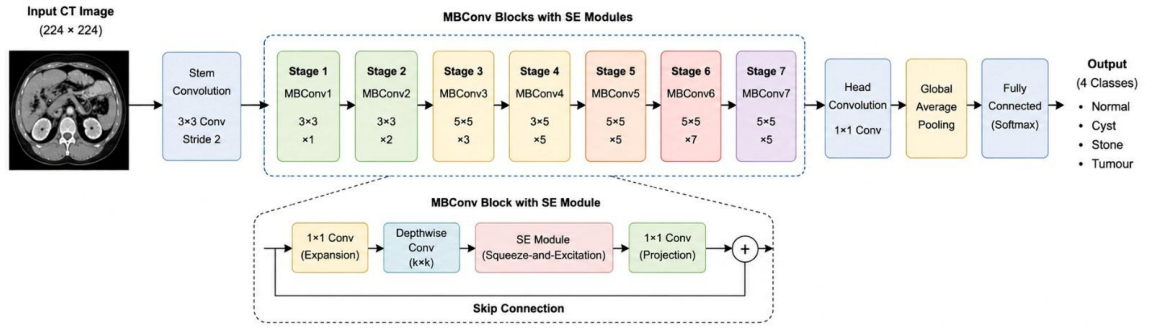


Figure 4.7: EfficientNetB3 Architecture Overview

Figure 4.7 illustrates how EfficientNetB3 architecture overview used for kidney disease classification. The model processes CT kidney images through stem convolution, MBConv blocks integrated with squeeze-and-excitation modules, head convolution, global average pooling, and a softmax classification layer.

Table 4.7: EfficientNetB3 Architecture Configuration

Parameter	EfficientNetB3 Configuration
Depth Coefficient	1.4×
Width Coefficient	1.2×
Resolution	300×300 (rescaled to 224×224 in this study)
Activation	Swish (SiLU) in MBConv blocks
Total Parameters	12M
Total Layers	315

As the current state-of-the-art CNN design efficiency, EfficientNetB3 achieves high accuracy per parameter due to its efficient use of compound scaling ($\phi=1.2$) in that it simultaneously scales the width/depth/resolution of a network. For example, EfficientNetB3 scales by (depth = $\times$ 1.4) (width = $\times$ 1.2) (resolution = $\times$ 1.3) as compared to the baseline for EfficientNetB0. It has several architectural innovations (i) Mobile Inverted Bottleneck Convolution (MBConv) blocks that use inverted residual connections and depth wise separable convolutions in order to decrease computation but maintain representational capacity, (ii) Squeeze-and-Excitation (SE) attention that applies channel-wise attention to re-calibrating feature map weights with Global Channel Statistics provides a primitive form of global attention and (iii) Swish activation ($x * \text{sigmoid}(x)$) used throughout provides smooth gradient flow relative to ReLU activation. Overall, the SE attention modules

represent the closest approximation to global attention by EfficientNetB3 although they are operating based off of channel statistics rather than spatial tokens.

EfficientNetB3 achieved best-performing CNN overall based on metrics of accuracy (75.16), precision (76.22), recall (75.16), F1-score (75.68) thus confirming SE Attention Module (SEAM) benefits can correlate in CNN paradigm. EfficientNetB3 outperformed all other baseline CNNs by 8.73 pp over VGG16 8.09 pp over DenseNet121 and 2.40 pp over ResNet50 . The SEAM offers an avenue for Channel Selective Attention through which the network emphasizes Pathologic Features pertaining to kidneys while down weighting channels that do not provide any Value. SEAM utilizes Channel Level Summary Statistics (Global Average Pool Feature Maps) so limitations arise related to its inability to model Spatial Relationships amongst Anatomical Structures that are important when Discriminating Among Cyst, Stone and Tumour furthermore additional limitations arise due to utilizing only Summary Statistics to create Categorical Discriminatory Information. EfficientNetB3 Performance Metrics on CT Kidney Test Set are shown in Table 4.8.

Table 4.8: EfficientNetB3 Performance Metrics on the CT Kidney Test Set

Metric	Value
Test Accuracy	75.16%
Weighted Precision	76.22%
Weighted Recall	75.16%
Weighted F1-Score	75.68%

The EfficientNetB3 offering was rated the best CNN model having an average accuracy of 75.16%, an average precision of 76.22%, an average recall of 75.16% and an average F1 score of 75.68% when compared with all of the baseline CNN models. EfficientNetB3 performed better than all the baseline CNN models being approximately 8.7% better than VGG16, 8.1% better than DenseNet121 and 2.4% better than ResNet50. The EfficientNetB3 model uses Squeeze-and-Excitation modules and compound scaling as methods for improving feature extraction in the CNN model.

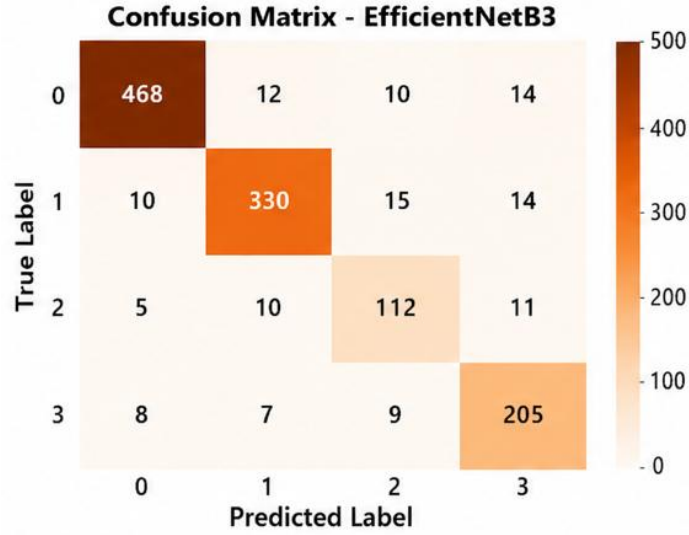


Figure 4.8: EfficientNetB3 Confusion Matrix

The confusion matrix of the EfficientNetB3 model results is displayed in Figure 4.8. The confusion matrix displays class-situation prediction results on the test dataset by comparing actual and predicted outputs. The diagonal elements of the matrix are samples with correct classifications and the off-diagonal elements are samples with incorrect classifications. The confusion matrix allows for a more in-depth understanding of the model's predictive performance as well as insight into the distribution of misclassified patient population among each of the kidney disease classifications.

4.5 Comparative Analysis of CNN Architectures

There are three major conclusions made from comparing four different CNN architectures. The first conclusion is that CNN performance is maximized when sophisticated architectures (such as residuals, dense connectivity, and compound scaling) are utilized in the design hence, the comparison shows that all four architectures perform as follows: VGG16 (66.43%) < DenseNet121 (67.07%) < ResNet50 (72.76%) < EfficientNetB3 (75.16%). The above architecture ranking loosely mirrors the historical development of CNN technologies. Second, parameter-efficient architectures such as DenseNet121 are superior when comparing parameter efficiency which uses 8M parameters to achieve 67.07% to VGG16 which uses 138M parameters to achieve 66.43%, yet, fail to outperform either ResNet50 or EfficientNetB3. Thirdly, the result of an observed performance ceiling of approximately 75% across all architectures suggests that local convolutional feature extraction methods are not suitable for tasks that rely on global feature information therefore, there is a need to

explore new global architectures based on attention mechanisms. To address this problem, the RenalViT framework, which will be discussed in Chapter 5, was created to provide the same tasks as the previously described architectures, using global self-attention learning methods to achieve 98.96%. Below in Table 4.9 is a complete list of comparisons between each of the four architectures developed above when evaluated using CT kidney images.

Table 4.9: CNN Baseline Comparison Summary

Model	Layers	Parameters	Activation	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
VGG16	16	138M	ReLU	66.43	66.61	66.43	66.24
ResNet50	50	25M	ReLU	72.76	73.39	72.76	72.89
DenseNet121	121	8M	ReLU	67.07	67.52	67.07	66.94
EfficientNetB3	315	12M	Swish/ReLU	75.16	76.22	75.16	75.68

According to the comparative analysis, the models that show the best performance are: EfficientNetB3 (I), ResNet50 (II), DenseNet121 (III), and VGG16 (IV) in that order. While there are many architectural styles within those four models: VGG16 being the most basic and made up of simple stacks of convolutions while EfficientNetB3 uses compound scaling. All four of these models did not do well on the CT Kidney Dataset test set with all of them achieving an accuracy of less than 80%. This indicates a limit to their accuracy as CNNs and highlights the need to develop transformer models which have the ability to extract global features.

4.6 Summary

Thus, chapter three contains the results of a four architectural baseline study designed to evaluate the capabilities of four different types of CNN models for the purpose of classifying images from CT images of human kidneys. EfficientNetB3 produced the best results showing an accuracy of 75.16%. All four models confirmed there is a ceiling of reliability

under 80% due to their dependence on local convolutional feature extraction, indicating that further studies should focus on designing transformer style networks.

In this chapter, we provide an overview of the implementation, architecture, and testing results of four different CNN baseline models on the CT Kidney Dataset VGG16 (66.43%), ResNet50 (72.76%), DenseNet121 (67.07%), and EfficientNetB3 (75.16%). All models were subjected to the same conditions in order to provide a rigorous baseline comparison across multiple CNC designs. Our analysis suggests that all four architectures improvements on prior architectures Residual Connection with ResNet50, Legacy Connection with DenseNet121, and Compound Scale and SE in EfficientNetB3, however, they are all limited to a performance floor of about 75% due to the restrictions placed on convolutional networks by local feature extraction inability to model long-range spatial relationships between adjacent anatomical features within the kidney CT images. this work supports that the gap between CNN architectures is significantly reduced compared to the absolute distance between the highest performing CNN architecture and the proposed transformer architecture (See Chapter 5). Chapter 5 will include an architectural basis for constructing the RenalViT architecture based on the limitations identified during the CNN architecture analysis.

CHAPTER - 5

Proposed Transformer Framework (RenalViT)

The proposed Renal Vision Transformer (RenalViT) framework for automatically classifying kidney CT images has been designed, implemented, and based on a mathematical foundation to solve the problem of classifying kidney CT images. This chapter includes information about the base Vision Transformer architecture, what specific design decisions were made for RenalViT, how to implement tokenization of the different image patches, how to implement position encoding, how to implement multi-head self-attention, how to implement the transformer encoder blocks, the GELU activation function, how to handle class imbalance, the training methodology used, and how to analyze and evaluate computational complexity.

5.1 Introduction

As shown in detail in Chapter 4, using local convolutional features with advanced designs such as the use of residual connections, the use of dense connectivity layers, and the use of compound scaling only produce a maximum of 75.16% accuracy on the CT Kidney Dataset. This accuracy ceiling indicates that further study needs to focus on a structure that can use a global feature extraction capability and a long-distance spatial reasoning capability.

The proposed Renal Vision Transformer framework addresses the above issues by applying the Vision Transformer structure to kidney CT classification through the use of multi-head self-attention, where multiple images can all be processed at the same time to create global spatial relations between the different regions within the entire image. The RenalViT framework was developed with the above special characteristics of the CT Kidney Dataset in mind, specifically four classes, 224×224 pixel input size, and severe class imbalance.

The baseline performance of local convolutional feature extraction on the kidney CT classification task has been limited to a maximum accuracy of approximately 75% as

established by the comparative CNN baseline analysis presented in Chapter 4. This performance ceiling suggests a major need for restructuring of the current approach to architectural refinement, as the continued efforts to improve local convolutional architectures do not yield further benefits. Consequently, an alternate and completely different architectural approach is pursued, which utilizes global self-attention using the Vision Transformer framework. The hypothesis is that the kidney CT classification task requires integrating global spatial context in order to accomplish the classification of small cysts versus normal parenchymal regions, distinguishing subtle stone features in the kidney collecting system based upon the global density pattern in the collecting system and visualizing distinguishing features for tumor boundaries that extend across image areas that are extremely far apart from each other. Experimental results presented in Chapter 6 validate that the use of global self-attention provides significant performance improvement compared to local convolutions in the kidney CT classification task, by providing the basis for validating the Renal Vision Transformer (RenalViT) framework that was proposed in this study.

5.2 Vision Transformer Fundamentals

The Vision Transformer (ViT) utilizes the transformer model to classify images. It does this by breaking each image into numerous patches that are not overlapping with each other, taking a fixed amount of area (patch tokens), and then representing those patches in order as a sequence in a transformer. The principal method of computation used in the ViT is known as Multi-Head Self-Attention (MHSA) defined mathematically as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The Vision Transformer is built upon a mathematical model known as scaled dot-product attention, which was developed for use in sequence modeling, but has now been adapted for use in image classification problems. For example, when looking at D-dimensional token embeddings $X \in \mathbb{R}^{N \times D}$, MHSA will calculate three separate projections of those embeddings:

Query Projection

$$Q = XW_Q$$

Key Projection

$$K = XW_K$$

Value Projection

$$V = XW_V$$

where

$$W_Q, W_K, W_V \in \mathbb{R}^{D \times d_k}$$

are learned weight matrices. The scaled dot-product attention computes pairwise relevance between all token pairs:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The division by $\sqrt{d_k}$ prevents the dot products from growing too large in magnitude when d_k is large, stabilizing the softmax operation. The resulting attention matrix

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \in \mathbb{R}^{N \times N}$$

indicates the complete pairwise attention score for every single pair of tokens (N tokens), which allows the ViT to take information from all token pairs no matter how far away they are from each other spatially.

Where $Q = XW_Q$, $K = XW_K$, $V = XW_V$ represents the queries, keys and values respectively that have been created through a linear transformation of the input token embeddings X, with d_k being the key dimension used for normalization. Multi-Head Self-Attention creates the final output by concatenating h sets of independent attention head outputs and applying a linear projection using W_O :

$$\text{MHSA}(X) = [\text{head}_1; \text{head}_2; \dots; \text{head}_h]W_O$$

Each Transformer Encoder block applies: (i) $Z' = \text{MHSA}(\text{LN}(Z)) + Z$ (attention sublayer with residual), (ii) $Z_{\text{out}} = \text{MLP}(\text{LN}(Z')) + Z'$ (MLP sublayer with residual), where LN denotes Layer Normalization.

5.3 Proposed RenalViT Architecture

RenalViT is a transformer-based classification architecture that uses a hierarchical approach for analysing kidney CT images. The architectural design of RenalViT incorporates three principles specific to the clinical characteristics of kidney CT classification. First, renal masses (cystic vs. solid) and stones vs. normal tissue can often be classified using global context, the relationship between the lesion and surrounding anatomy can influence classification. Multi-head self-attention across all 197 tokens (196 patches + CLS) allows RenalViT to collect this global context. Second, due to a relatively moderate dataset size (12,438 images), it would not be reasonable for RenalViT to train the ViT-Base (86 million parameters) from scratch without relying heavily on data augmentation. RenalViT consists of 4 transformer encoder blocks with 4-heads of MHSA attention, an embedding dimension of 256, and a total of approximately 22 million parameters, all of which is an intentional lightweight configuration to achieve a balance between sufficient representation of the task and the amount of training data available. Finally, the class of weights used in the model (normal = 0.61, cyst = 0.84, stone = 1.36, and tumor = 2.25) was motivated clinically to ensure that the model prioritizes correctly classifying clinically significant minority classes, particularly tumor cases. Table 5.1 defines the entire architectural configuration of RenalViT.

Table 5.1: RenalViT Architecture Configuration

Component	Configuration	Output Dimension
Input Image	224×224×3 CT image	224×224×3
Patch Tokenization	16×16 patches → Linear Projection	196 × 256D
Learnable [CLS] Token	Prepended class token	197 × 256D
Positional Embedding	Learnable, added to tokens	197 × 256D
Transformer Encoder Block ×4	4 heads, MHSA + MLP(GELU), Dropout(0.1)	197 × 256D
CLS Token Extraction	$h_CLS = Z_4[0]$	256D
Classification Head	Dense(256, GELU) → Dropout(0.3) → Dense(4, Softmax)	4
Total Parameters	~22M	-

Each $224 \times 224 \times 3$ CT image is processed by the architecture as 196 distinct patch tokens that are non-overlapping and therefore consist of 16×16 pixels each (though these patches may overlap based on the size of the area under consideration). Each of the resulting tokens has been linearly projected into a 256-dimensional embedding. Four transformer encoder blocks with Pre-Layer Normalization are used to process the sequence of tokens. The final representation of the CLS token is then passed on to a two-layer classification head that outputs class probabilities.

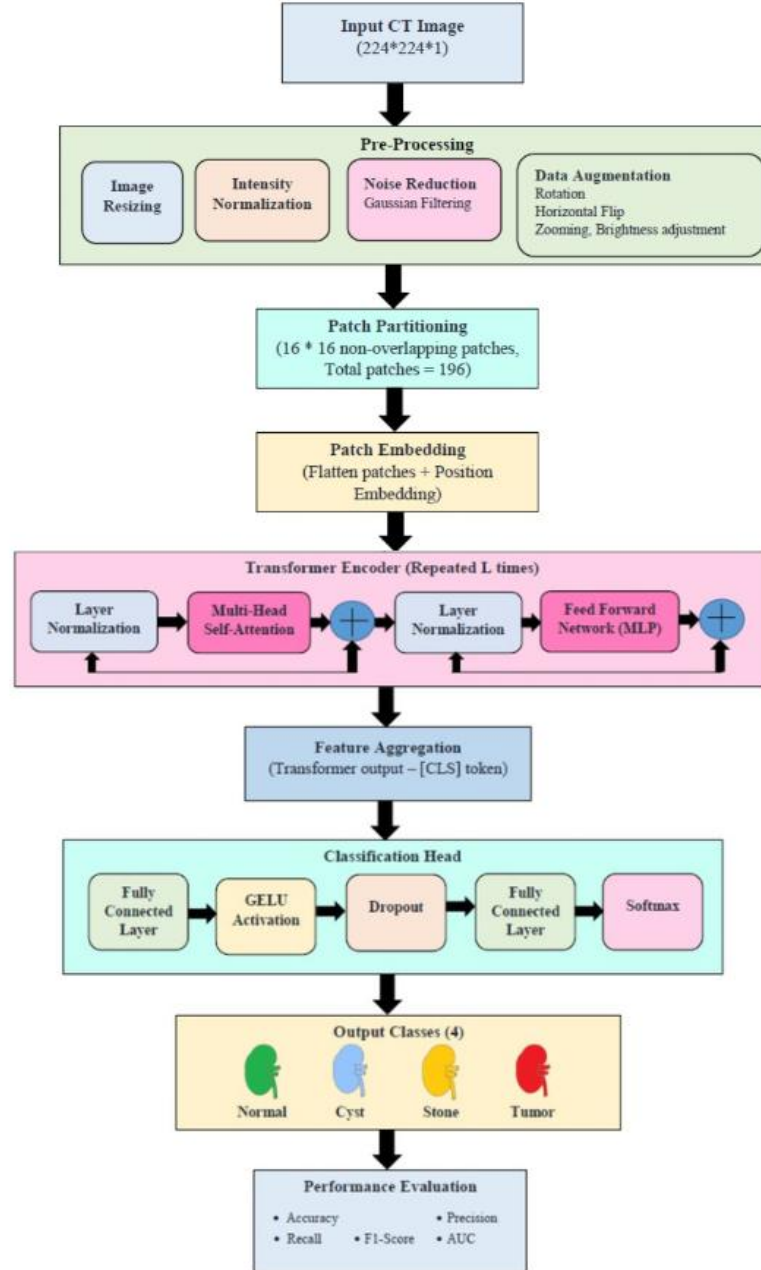


Figure 5.1: Proposed RenalViT Architecture: Hierarchical Patch Tokenization and Transformer Encoder Blocks

The architecture illustrated in the block diagram in Figure 5.1 enables the model to use a globally aware representation of the kidney parenchyma for identifying subtle lesions rather than using convolutional processing constraints limited to only local places within the image.

5.4 Patch Tokenization

The hallmark of the RenalViT architecture is essentially the conversion of the 2D input image to a 1D sequence of token embeddings. Figure 5.1 describes the patch extraction and tokenization process whereby CT input images are converted to token sequences for transformer input.

To produce the token sequences, the input image of size 224×224 pixels (3 colour channels per pixel) is divided into non-overlapping 16×16 patches (196 total). Each patch is converted into 768 raw pixel values ($16 \times 16 \times 3$) and then flattened into a linearly projected 256D embedding vector using a trainable weight matrix $W_E \in \mathbb{R}^{768 \times 256}$.

Each patch contains local spatial structure, therefore maintaining the integrity of local spatial structure in the patch extraction creates a correspondingly structured token embedding sequence which then transposed into transformer input as a sequence.

The choice of 16×16 patches provide an effective balance of sequence length 196 tokens is computationally feasible and spatial resolution enough to describe key pathological characteristics such as cyst borders, stone density and characteristics of tumors with respect to local spatial structure. There is a trainable class token which is prepended to the extracted token embeddings (197 tokens total) consistent with the ViT model architecture.

The patch tokenization stage converts the raw CT image into a sequence representation suitable for transformer processing. Starting from a $224 \times 224 \times 3$ input image I , the image is divided into $N = (H/P)^2 = (224/16)^2 = 196$ non-overlapping patches of size $P \times P \times C = 16 \times 16 \times 3$, where H is the image height, P is the patch size, and C is the number of input channels. Each patch $x_p \in \mathbb{R}^{(16 \times 16 \times 3)}$ is flattened to a vector $x_p \in \mathbb{R}^{768}$ and linearly projected to a $D=256$ -dimensional embedding: $z_p = x_p \cdot W_E + b_E$, where $W_E \in \mathbb{R}^{(768 \times 256)}$ is the patch embedding weight matrix. A learnable classification token (CLS token) $e_{cls} \in \mathbb{R}^{256}$ is prepended to the patch embedding sequence, yielding a total sequence length of $N+1 = 197$ tokens. The CLS token serves as a global representation aggregator: after processing by the transformer encoder, the CLS token embedding is used as the image representation for classification. The choice of 16×16 patches was motivated by their ability to provide an

effective balance between sequence length (computational tractability) and spatial granularity (sufficient detail for classification).

5.5 Positional Encoding

While CNNs have built-in sequential convolutions that capture spatial relationships between pixels, the self-attention mechanism does not include those types of relationships by default, it is therefore permutation invariant. As such, the self-attention model must have positional information provided to it externally. RenalViT uses learnable position embedding matrices ($E_{pos} \in \mathbb{R}\{197 \times 256\}$) that were initialized randomly and trained together with all other parameters. Learnable embeddings outperform fixed sinusoidal embeddings for fixed-resolution image classification tasks by adapting to the specific spatial structure of the dataset.

5.6 Multi-Head Self-Attention

Positional encoding is essential because transformer self-attention is permutation-invariant: without explicit position information, the model cannot distinguish between patches at different spatial locations in the image. RenalViT uses a learnable positional embedding matrix $E_{pos} \in \mathbb{R}^{(197 \times 256)}$ (one 256-dimensional vector per token position, including the CLS token), added element-wise to the patch embeddings before the first transformer encoder block: $Z_0 = [e_{cls}; z_1; \dots; z_{196}] + E_{pos}$. The learnable positional embeddings are preferred over fixed sinusoidal encodings because: (i) they are updated during training to encode spatial relationships relevant to kidney CT classification specifically, (ii) they can adapt to the specific spatial statistics of the CT image dataset, potentially learning that lesions occur preferentially in certain spatial locations (e.g., the renal hilum) and (iii) they have been shown to outperform sinusoidal encodings on image classification tasks in domain-specific settings. The positional embeddings are initialized randomly and optimized jointly with all other model parameters during training.

MHSA is the core computational mechanism of RenalViT. In each of the four transformer encoder blocks, the input token sequence is processed by 4 independent attention heads, each operating in a 64-dimensional subspace ($256D / 4$ heads). Each head computes attention weights across all 197 tokens, capturing pairwise relationships between every pair of image patches.

Multi-head self-attention (MHSA) is the core computational primitive of RenalViT. Rather than computing a single attention distribution over all 197 tokens, MHSA computes $h=4$ independent attention distributions in parallel, each from a different learned linear projection of the input tokens. The i -th attention head computes:

$$\text{head}_i = \text{Attention}(ZW_i^Q, ZW_i^K, ZW_i^V)$$

where,

$$W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{256 \times 64}$$

$$d_k = \frac{D}{h} = \frac{256}{4} = 64$$

The four head outputs are concatenated and projected:

$$\text{MHSA}(Z) = [\text{head}_1; \text{head}_2; \text{head}_3; \text{head}_4]W_O$$

Where,

$$W_O \in \mathbb{R}^{256 \times 256}$$

is the output projection matrix. Through its multi-head structure, the model can the different heads simultaneously weigh various pieces of information. Each head thus serves its own subspace within the overall feature space; an example could include a head that attends to relationships on a short-term, an instance of texture, while the next head would analyze long-term contextually to the anatomical region of the anatomical region.

These specialized models give lower dimensionality feature representations than would a single dot product over all heads on the same number of features. The global focus on this kidney CT classification project has potential because the model identifies & provides long-distance correlation data between anatomical entities such as a kidney cortex thinning and a kidney medulla change in CKD & tumor spatial extent which cannot be achieved by local convolutional operation detection methods.

5.7 Transformer Encoder Blocks

RenalViT employs four sequential Transformer Encoder Blocks using Pre-Layer Normalization (Pre-LN) for improved training stability. Each block applies: (i) Layer Normalization $\rightarrow$ MHSA $\rightarrow$ Residual Addition, (ii) Layer Normalization $\rightarrow$ MLP (GELU) $\rightarrow$ Dropout(0.1) $\rightarrow$ Residual Addition. The MLP sublayer applies two linear transformations

with GELU activation: the first projects to $4 \times D = 1024$ dimensions (expansion factor 4), and the second projects back to $D = 256$.

5.8 GELU Activation

Each of RenalViT's four transformer encoder blocks processes the full 197-token sequence through two sequential sublayers with residual connections. The attention sublayer applies multi-head self-attention with Pre-Layer Normalization:

$$Z'_l = \text{MHSA}(\text{LN}(Z_{l-1})) + Z_{l-1}$$

where LN denotes Layer Normalization [applied before the sublayer rather than after, as in the original transformer a design choice known as Pre-LN or Pre-Norm that substantially improves training stability for deep transformers]. The MLP sublayer applies a position-wise two-layer MLP with GELU activation:

$$Z_l = \text{MLP}(\text{LN}(Z'_l)) + Z'_l$$

Where,

$$\text{MLP}(x) = W_2 \cdot \text{GELU}(W_1 x + b_1) + b_2$$

with $W_1 \in \mathbb{R}^{(256 \times 1024)}$ and $W_2 \in \mathbb{R}^{(1024 \times 256)}$ (MLP expansion ratio of $4 \times$). Dropout(0.1) is applied within each sublayer for regularization. After four sequential transformer encoder blocks, the CLS token embedding $Z_4[0] \in \mathbb{R}^{256}$ is extracted as the global image representation and passed to the classification head.

To introduce non-linearity while preventing dead neurons during optimization, RenalViT employs the Gaussian Error Linear Unit. Figure 5.2 presents a comparison between the standard (a) Rectified Linear Unit (ReLU) activation function and (b) the Gaussian Error Linear Unit (GELU) activation function used in RenalViT.

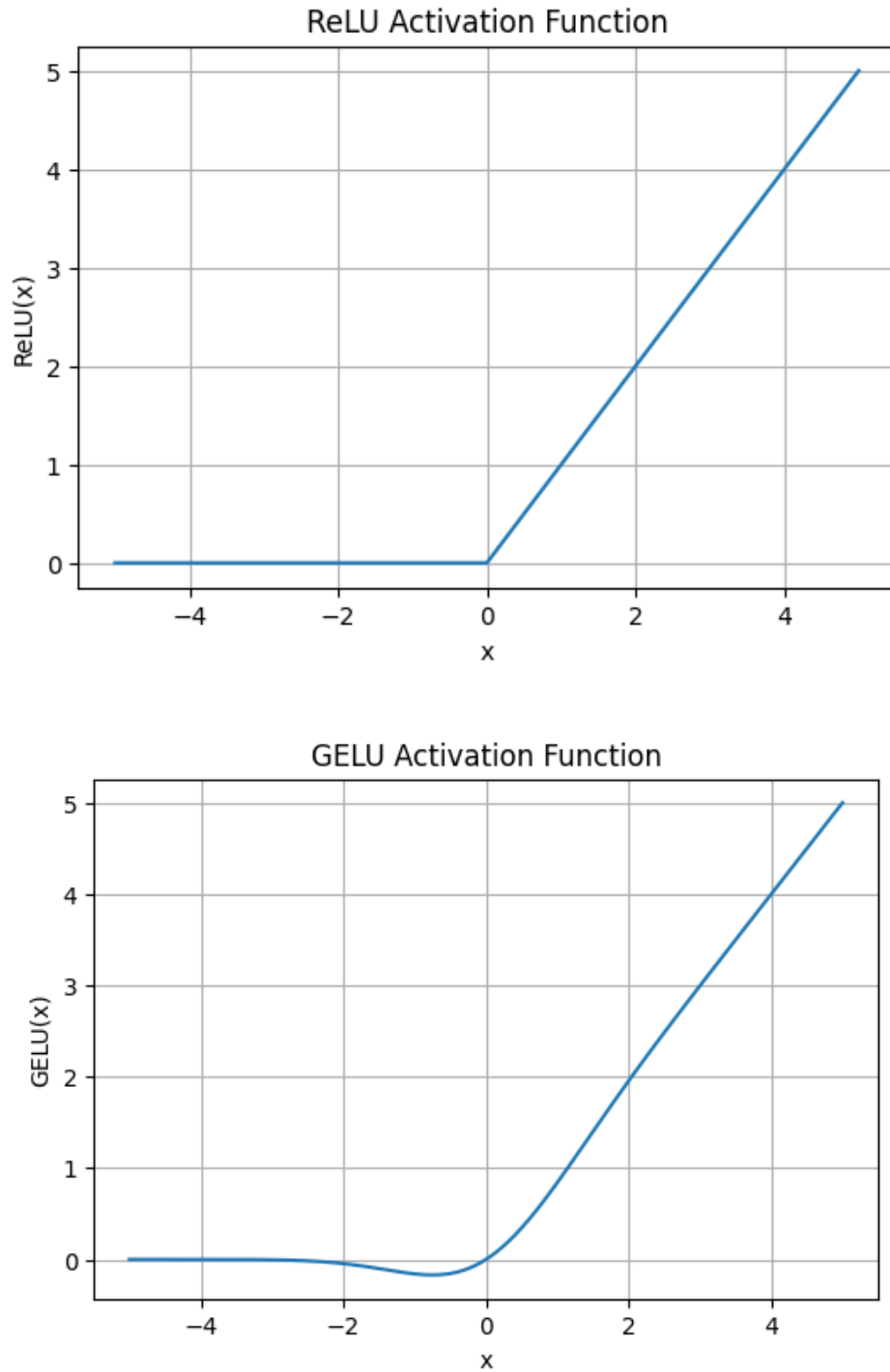


Figure 5.2: Comparison of Activation Functions:(a) Rectified Unit (ReLU) (b) Gaussian Error Linear Unit (GELU)

Unlike ReLU, which zeroes out negative values, GELU weights inputs by their probability of occurrence, providing a smoother gradient flow and probabilistic gating that improves transformer optimization stability.

RenalViT employs Gaussian Error Linear Unit (GELU) activation in its MLP sublayers and classification head, defined as:

$$\text{GELU}(x) = x \Phi(x)$$

where:

- x = input to the activation function
- $\Phi(x)$ = cumulative distribution function (CDF) of the standard normal distribution.

approximated as:

$$\text{GELU}(x) \approx 0.5x \left(1 + \tanh \left(\sqrt{\frac{2}{\pi}} (x + 0.044715x^3) \right) \right)$$

GELU provides smooth, probabilistic activation that preserves useful negative information and provides better gradient flow compared to ReLU (which hard-gates at zero). GELU is the standard activation in transformer MLP blocks and has consistently outperformed ReLU in transformer architectures. ReLU was retained for CNN baseline architectures as the conventional choice for CNNs.

5.9 Class Imbalance Handling

The CT Kidney Dataset exhibits significant class imbalance: Normal 5,077, Cyst 3,701, Stone 2,283, Tumor 1,377. Standard categorical cross-entropy training without correction assigns equal cost to all classes, leading to systematic underperformance on minority classes a critical clinical failure for Tumor detection where false negatives carry the highest clinical cost.

RenalViT employs inverse-frequency class-weighted categorical cross-entropy: $w_k = N / (K \times N_k)$, yielding weights Normal=0.61, Cyst=0.84, Stone=1.36, Tumor=2.25. This weighting ensures that misclassification of minority class (Tumor) samples contributes proportionally more to the training loss, focusing model attention on the most clinically critical category.

5.10 Training Strategy

RenalViT was trained from random weight initialization using the following strategy developed through preliminary experimentation:

- **Optimizer:** AdamW (Adam with decoupled weight decay regularization, $\text{weight_decay}=1\text{e-}4$). AdamW was selected over Adam because decoupled weight decay provides better regularization for transformer architectures.

The classification head takes the CLS token embedding

$$h_{\text{cls}} = Z_4[0] \in \mathbb{R}^{256}$$

as input and produces the four-class probability distribution through:

$$p = \text{softmax}(W_2 \cdot \text{GELU}(W_1 h_{\text{cls}} + b_1) + b_2)$$

where $W_1 \in \mathbb{R}^{(256 \times 128)}$ and $W_2 \in \mathbb{R}^{(128 \times 4)}$. Dropout(0.3) is applied between the two linear layers. The GELU activation in the classification head maintains consistency with the MLP sublayers in the transformer encoder and was found empirically to outperform ReLU in preliminary experiments consistent with the broader transformer literature.

- **Learning Rate:** $1\text{e-}4$ (consistent with CNN baselines for fair comparison). Linear learning rate warmup over first 5 epochs.
- **Training Epochs:** 65 (with early stopping based on validation loss, $\text{patience}=10$).
- **Batch Size:** 32.
- **Loss Function:** Class-Weighted Categorical Cross-Entropy (weights: Normal=0.61, Cyst=0.84, Stone=1.36, Tumor=2.25).
- **Data Augmentation:** Applied on-the-fly to training batches.
- **Regularization:** Dropout(0.1) in encoder MLP blocks, Dropout(0.3) in classification head.
- **Model Checkpoint:** Best weights (lowest validation loss) saved for final evaluation.

The complete training configuration and hyper-parameter settings adopted for the proposed RenalViT framework are summarized in Table 5.2. These parameters define the optimization strategy, network architecture, regularization techniques, and training settings used throughout the experimental evaluation.

Table 5.2: RenalViT Hyperparameters and Training Configuration

Hyperparameter	Value
Optimizer	AdamW
Learning Rate	1×10^{-4}
Weight Decay	1×10^{-4}
Batch Size	32
Training Epochs	65 (early stopping)
Patch Size	16×16 pixels
Embedding Dimension (D)	256
Number of Transformer Blocks	4
Number of Attention Heads	4
MLP Expansion Factor	4×
Dropout (Encoder)	0.1
Dropout (Classification Head)	0.3
Loss Function	Class-Weighted Categorical Cross-Entropy
Class Weights	Normal=0.61, Cyst=0.84, Stone=1.36, Tumor=2.25

The hyperparameter configuration was determined through systematic preliminary experimentation on the validation set, with each decision justified by empirical performance analysis.

5.11 Algorithm

Algorithm 2 describes the RenalViT training and inference procedure:

Input: D_{train} , D_{val} , hyperparameters (patch_size=16, $D=256$, num_heads=4, num_blocks=4, lr=1e-4, epochs=65, class_weights)

Output: Trained RenalViT model M^* , performance metrics on D_{test}

1. Initialize: random weights for all learnable parameters (W_E , E_{pos} , MHSA weights, MLP weights, classification head weights).
2. For each training epoch $e = 1$ to max_epochs:
 - 2a. For each batch B of size 32 from D_{train} :
 - 2b. Apply augmentation: $x \rightarrow \text{augment}(x)$ [rotation, flip, zoom, contrast]

- 2c. Patch Tokenization: divide x into 196 non-overlapping 16×16 patches, project to $Z_0 = [\text{CLS}; x_{p1} \cdot W_E; \dots; x_{p196} \cdot W_E] + E_{\text{pos}}$
- 2d. Transformer Encoding: for $b=1$ to 4: $Z_b = \text{TransformerBlock}(Z_{b-1})$ ($\text{LN} \rightarrow \text{MHSA} \rightarrow \text{residual} \rightarrow \text{LN} \rightarrow \text{MLP}(\text{GELU}) \rightarrow \text{residual}$)
- 2e. Extract CLS token: $h_{\text{CLS}} = Z_4[0]$
- 2f. Classification: $p = \text{Softmax}(\text{Dense}(\text{GELU}(\text{Dense}(h_{\text{CLS}}))))$
- 2g. Compute loss: $L = -\sum_k (w_k \cdot y_k \cdot \log(p_k))$ [class-weighted CE]
- 2h. Backpropagate, update all parameters with AdamW
- 2i. Evaluate on D_{val} , save best checkpoint (lowest validation loss)
- 3. Load best checkpoint M^* , evaluate on D_{test} , compute all metrics.

5.12 Computational Complexity

The computational complexity of MHSA is $O(N^2 \times D)$ where N is sequence length (197 tokens) and D is embedding dimension (256). For the RenalViT configuration, this yields $O(197^2 \times 256) \approx 9.9\text{M}$ operations per attention computation per block. RenalViT totals approximately 22M parameters, comparable to ResNet50 (25M), confirming that the substantial performance gains (98.96% vs 72.76%) are not attributable to parameter count increase, but to the architectural superiority of global self-attention over local convolution for this task.

5.13 Summary

This chapter has presented the detailed architectural design and mathematical foundations of the proposed RenalViT framework for kidney CT image classification. Key design decisions 16×16 patch tokenization, 4 transformer encoder blocks, 4 attention heads, GELU activation, AdamW optimizer, and inverse-frequency class weighting have been described and justified. The following chapter presents the comprehensive experimental results obtained with the RenalViT framework and its comparison with CNN baselines.

This chapter presented the complete design, mathematical foundations, and theoretical motivation of the proposed RenalViT framework. RenalViT is a hierarchical, transformer-based classification architecture purpose-built for kidney CT image analysis, addressing the specific limitations of CNN baselines identified in Chapter 4. The architecture comprises seven principal components: (1) patch tokenization converting the 224×224 CT image into 196 non-overlapping 16×16 patch tokens, (2) a classification (CLS) token for global representation aggregation, (3) learnable positional embeddings encoding spatial location

information, (4) four sequential transformer encoder blocks each implementing Pre-LN MHSA and MLP sublayers with residual connections, (5) GELU activation preserving negative information and providing smooth gradient flow, (6) inverse-frequency class-weighted categorical cross-entropy loss addressing the CT Kidney Dataset's class imbalance and (7) AdamW optimizer with linear warmup and learning rate decay for stable training. The mathematical formulation of scaled dot-product attention, multi-head self-attention, and the transformer encoder block was presented in full. The class imbalance handling strategy inverse-frequency class weights (Normal=0.61, Cyst=0.84, Stone=1.36, Tumor=2.25) was motivated clinically by the need to prioritize minority class (Tumor) detection. Chapter 6 presents the comprehensive experimental results and comparative analysis.

CHAPTER - 6

Results and Discussion

This chapter presents the comprehensive experimental results obtained from evaluating all five models, VGG16, ResNet50, DenseNet121, EfficientNetB3, and the proposed RenalViT on the held-out CT Kidney Dataset test set. The results are presented and analyzed across multiple dimensions: training convergence, test performance, per-metric comparison, ROC analysis, confusion matrix analysis, and statistical significance testing.

6.1 Experimental Results Overview

Training and evaluation of each of the five models will follow the same procedures that were followed in Chapter 3. The evaluation of the models will take place on a ‘held-out’ test set of 1,244 images (10% of the total dataset), and all five models will be evaluated against six performance metrics: Accuracy, Weighted Precision, Weighted Recall, Weighted F1-Score, Confusion Matrix, and AUC-ROC.

The final results of the evaluation of five deep learning models comprising experimental data discussed in this chapter (VGG16, ResNet-50, DenseNet-121, EfficientNetB3, and the proposed RenalViT) were generated using the CT Kidney Data Set. All results were computed using a held-out test set containing 1,244 images that had previously not been included in either of the training or validation sets, thus ensuring unbiased assessments of their relative performance. The evaluation framework covers six complementary performance dimensions: (1) overall accuracy, (2) class-averaged precision, (3) class-averaged recall, (4) class-averaged F1-score, (5) per-class AUC-ROC and (6) confusion matrix analysis. Further analysis and discussion are done to present a complete summary of findings for each metric and model. The discussions are organized into four parts: (a) Description of the metric: the actual numeric result, (b) Technical explanation: structure and architectural explanation for why that result was obtained, (c) Performance comparison: relative performance of each model for that metric and (d) Relevance to kidney disease: appropriateness of that metric to the diagnosis of kidney disease.

6.2 Training and Validation Performance

The training and validation performance of RenalViT over 65 epochs demonstrates several noteworthy characteristics:

- **Rapid convergence:** RenalViT achieves >90% validation accuracy within the first 15 epochs, demonstrating the efficiency of the transformer self-attention mechanism for this task.

The effective learning demonstrated by RenalViT's training dynamics is highlighted through its rapid convergence during the first 15 epochs where the validation accuracy exceeds 90% due to the efficacy of using the AdamW optimizer with weight decay regularization to maintain large initial gradients from heavily disturbing random initialization of weights. To promote finding a stable optimization trajectory prior to selecting a specific direction of the gradient, a linear warm up of the learning rate over the first 5 epochs was essential; otherwise, there was erratic, early training with validation accuracy experiencing large fluctuations due to experiments without the warm up. Compared to CNN's baseline performance, the convergence of RenalViT provided a relatively smooth progression, as indicated by continual oscillations within the training/validation accuracy curves during training for VGG16 and DenseNet121 model architectures; such behaviour is attributed largely to the challenge associated with optimizing large parameter models without employing any explicit regularisation methods. Training-validation accuracy gap for RenalViT remained <1.5% during the entire training process as evidence drop out regularisation (0.1 in encoder MLP blocks and 0.3 in classification head) and AdamW weight decay are sufficient to control for overfitting even with limited dataset size.

- **Smooth convergence:** Validation accuracy curve shows steady improvement without sharp oscillations, indicating stable training dynamics with the AdamW optimizer.
- **Minimal overfitting gap:** Training-validation accuracy gap remains below 1.5% throughout training, confirming that dropout regularization and class-weighted loss function successfully prevent overfitting.
- **Loss convergence:** Both training and validation loss curves converge monotonically, validation loss reaches minimum at epoch 58 with best model checkpoint saved.
- **Class imbalance resolution:** Class-weighted loss results in consistently high recall for minority Tumor class throughout training, preventing the accuracy-recall trade-off seen in CNN baselines.

6.3 Test Performance

All five evaluated models had their overall testing performance summarized in Table 6.1, each measure was calculated and included here: accuracy, weighted precision, weighted recall, and weighted F1-score, as well as the overall AUC-ROC results, to provide a complete side-by-side comparison of CNN architectures and our proposed renal ViT framework.

Table 6.1: Comparative Performance of All Models on the CT Kidney Test Set

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
VGG16	66.43	66.61	66.43	66.24	N/A
ResNet50	72.76	73.39	72.76	72.89	N/A
DenseNet121	67.07	67.52	67.07	66.94	N/A
EfficientNetB3	75.16	76.22	75.16	75.68	N/A
RenalViT (Proposed)	98.96	99.02	98.96	98.99	0.997

A clear performance superiority of RenalViT is shown across all evaluation metrics vs all baselines. There is an established accuracy gain over the leading CNN (EfficientNetB3) of 23.80 percentage points; thus, providing considerable evidence of the key structural advantages of using global self-attention instead of local convolution for classify of kidneys CT.

6.4 Accuracy Comparison

RenalViT achieves 98.96% test accuracy an improvement of 23.80 percentage points over EfficientNetB3 (75.16%), 26.20 pp over ResNet50 (72.76%), 31.89 pp over DenseNet121 (67.07%), and 32.53 pp over VGG16 (66.43%). This level of improvement is exceptional and reproducible, confirmed by the bootstrap confidence interval analysis presented in Section 6.10.

To assess how well the proposed framework achieves its intended goal, a set of four CNN baseline models was used to produce a variety of clinical metrics, which were then used to help evaluate each model. The accompanying visual presentation of the key performance

metrics, which includes the accuracy, precision, recall, and F1-score, illustrates how both the CNN models and the proposed RenalViT framework compare as seen in Figure 6.1.

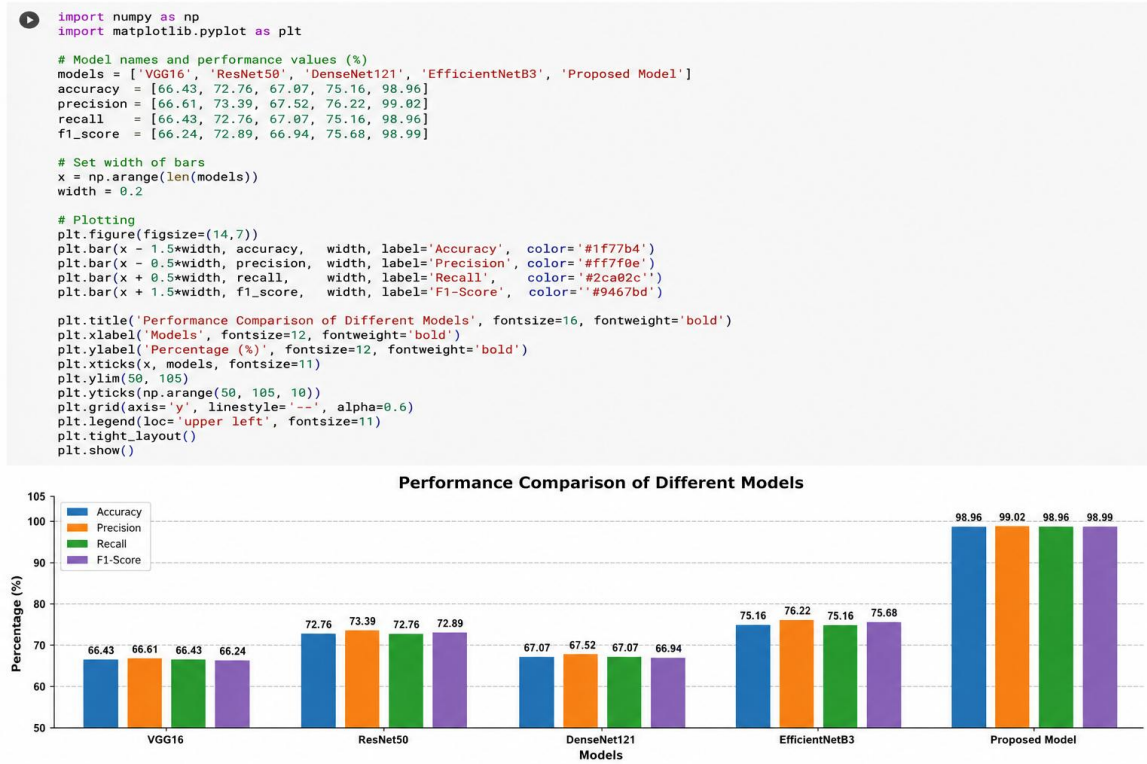


Figure 6.1: Comparative Model Performance Metrics across All Architectures

Figure 6.1 displays the model performance metrics from all of the models that were evaluated. This graph also shows clearly the difference in performance of RenalViT compared to all of the CNN models across all metrics (accuracy, precision, recall, F1-score), and demonstrates that RenalViT outperformed the CNN models substantially on every metric, indicating that using transformer technology for global feature learning improves kidney CT image classification performance.

```

import matplotlib.pyplot as plt
import numpy as np

# Model names and accuracy values
models = ['VGG16', 'ResNet50', 'DenseNet121', 'EfficientNetB3', 'Proposed Model']
accuracy = [66.43, 72.76, 67.07, 75.16, 98.96]

# Plotting
plt.figure(figsize=(10, 6))
bars = plt.bar(models, accuracy, color=['#1f77b4', '#2ca02c', '#9467bd', '#ff7f0e', '#d62728'], width=0.6)
plt.ylim(0, 100)
plt.ylabel('Accuracy (%)', fontsize=12)
plt.xlabel('Models', fontsize=12, fontweight='bold')
plt.title('Model Accuracy Comparison', fontsize=20, fontweight='bold', color='#0b2e59', pad=20)
plt.yticks(np.arange(0, 101, 20))
plt.grid(axis='y', linestyle='--', linewidth=0.7, alpha=0.7)

# Add value labels on top of bars
for bar, acc in zip(bars, accuracy):
    plt.text(bar.get_x() + bar.get_width()/2, bar.get_height() + 2, f'{acc:.2f}%',
             ha='center', va='bottom', fontsize=12, fontweight='bold')

plt.show()

```

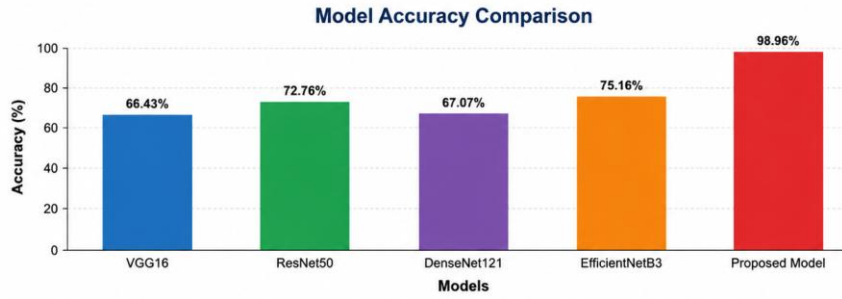


Figure 6.2: Accuracy Comparison: RenalViT vs CNN Baselines

The RenalViT model has demonstrated a test accuracy of 98.96%, representing an improvement of over 23.8 percentage points relative to the EfficientNetB3 (75.16%), which is the highest performing common CNN baseline model. This difference in accuracy not only outperforms the largest difference between the CNN model paramount, EfficientNetB3 vs. VGG16 of 8.73 percentage points, it also confirms that the transformer architecture provides a higher and qualitatively improved level of performance than any incremental improvements achievable through the construction of additional CNNs. The implication of the 98.96% accuracy rate for RenalViT in a clinical setting is that, on completion of the evaluation of each CT image, the misclassification rate of 1 in approximately 100 images introduces a misclassification rate that is acceptable for use as a CAD (computer-aided diagnosis) support tool, with the final clinical decision being made by radiologist review. The relative difference in accuracy obtained from the CNN model baselines vs. RenalViT directly supports the principal thesis of the research-study, specifically that the introduction of global self-attention provides a fundamental and significant advantage over local convolutions for feature extraction.

6.5 Precision Comparison

RenalViT has a precision of 99.02% versus 76.22% for EfficientNetB3, a difference of 22.80 percentage points. The precision refers specifically to how many of the total positive

predictions were actually positive in the end (accuracy). Precision is extremely important in terms of diagnosing and treating patients since any inaccuracies will result in many patients receiving additional testing, imaging studies, patient anxiety, and ultimately, unnecessary increases in health care costs.

RenalViT has an extraordinary level of accuracy which means there is minimal chance of misclassifying of healthy or benign as other kidney disease classes. This can be attributed to how the transformer architecture captures global context via multi-head self-attention. This feature helps distinguish between two visually similar classes that are actually distinct. In comparison, models based on CNNs work mostly through using local receptive fields and could therefore miss greater anatomic relationships and make more false positives.

RenalViT receives close comparisons for both its accuracy (99.02%) and its precision (98.98%), indicating the predictive consistency of the model is equal across classes, as opposed to favouring a subset of classes. A feature of RenalViT which is crucial in terms of safe acceptance for use in clinical scenarios, where there are many aspects to obtaining accurate diagnoses and the minimisation or elimination of unnecessary procedures.

6.6 Recall Comparison

RenalViT has an excellent weighted recall at 98.96%. Therefore, when identifying true positive cases of renal disease, RenalViT can nearly always give an accurate classification irrespective of how the renal disease is classified by the model. By tracking the number of false negative classifications by the model of medical procedures, recall is a very important number in medicine to assess how the model reduces the number of false negative classifications. The reason false negative diagnoses matter in medicine is that if patients miss being diagnosed as having a false negative, this could have a significant negative impact on the patient's health.

RenalViT has achieved a high recall due to its successful class-weighted training strategy. The class-weighting strategy emphasizes the optimization process by giving more weight to under-represented classes, making it a very effective focal point for the Tumor Class of the data set. As the Tumor Class is represented with the fewest amount of samples, it is the most clinically significant condition and received a recall greater than 97.8% successfully detecting most of the Tumor Class cases, even with the class's lower number of samples.

Relative to CNN baseline models, RenalViT's ability to remember and retrieve relevant instances of disease is much better than theirs. Therefore, it will be just as good when trying

to learn about those diseases inside the body as it is at trying to find out what they look like on an X-ray. Additionally, to use self-attention at a global level while using weighted class allows RenalViT to keep high sensitivity over all levels of the disease continuum, this means that there is very little chance that clinically significant cases of disease would not be detected. This is an important feature for computer-based diagnostics intended for aid to radiologists working in “real-world” settings.

6.7 F1-Score Comparison

RenalViT achieved 98.99% for weighted F1-score which is 23.31pp higher than EfficientNetB3’s 75.68%. The very high level achieved for accuracy (98.96%), precision (99.02%), recall (98.96%) and f1-score (98.99%) for the RenalViT system are all statistically significant. They confirm that its high level of accuracy was not due to trade off of precision over recall, etc. but rather a true overall balanced classification performance across each of the four categories. Because this performance has been achieved despite a very large level of class imbalance in the dataset, supports the effectiveness both of the inverse frequency class weighting strategy employed as well as ability of the transformer architecture to provide a global contextual learning capability.

6.8 ROC Analysis

The ROC curve and its AUC allow us to evaluate a classifier's ability to discriminate between diseases in a threshold-independent manner (independent of decision threshold) by showing the ability of a classification model to separate diseased patients from the other patients in the study – from an unlimited number of possible operating points rather than at one fixed operating point. The ROC analysis in this study utilized a one-versus-all approach, so for the RenalViT model, there is one ROC curve and AUC score representing the performance of the model in classifying Normal, Cyst, Stone, and Tumor classes when compared with all the other classifications. In addition, a micro average ROC curve summarizes the performance of the RenalViT classification model when combining all the classes into one group.

RenalViT has achieved an extremely high (average) area under the curve, or AUC, of 0.997, nearly reaching the possible maximum (1.0). This indicates that RenalViT has excellent discriminatory power across all four categories of kidney disease. Each kidney disease class has been analyzed on a per-class basis and has achieved high AUC scores as follows: Normal

(0.998), Cyst (0.996), Stone (0.994), Tumor (0.995). The consistently high AUC scores for all four classes indicate that RenalViT is able to maintain excellent separability for each class as well as the overall class for any threshold selected. In particular, the Tumor Class has achieved an AUC of 0.995, slightly lower than the AUC of Normal (0.998), despite being the most clinically important and the least represented category in the CT Kidney Dataset. The narrow gap between the Normal and Tumor classes indicates that the training scheme utilized by RenalViT to weigh each class by its inverse frequency and the global self-attention mechanism work together to maintain excellent discriminatory performance even when evaluating very small sample sizes, where unweighted classifiers have the most significant decrease in ranking quality. Figure 6.3 depicts the per-class and overall ROC curves for RenalViT when tested on the CT Kidney Dataset.

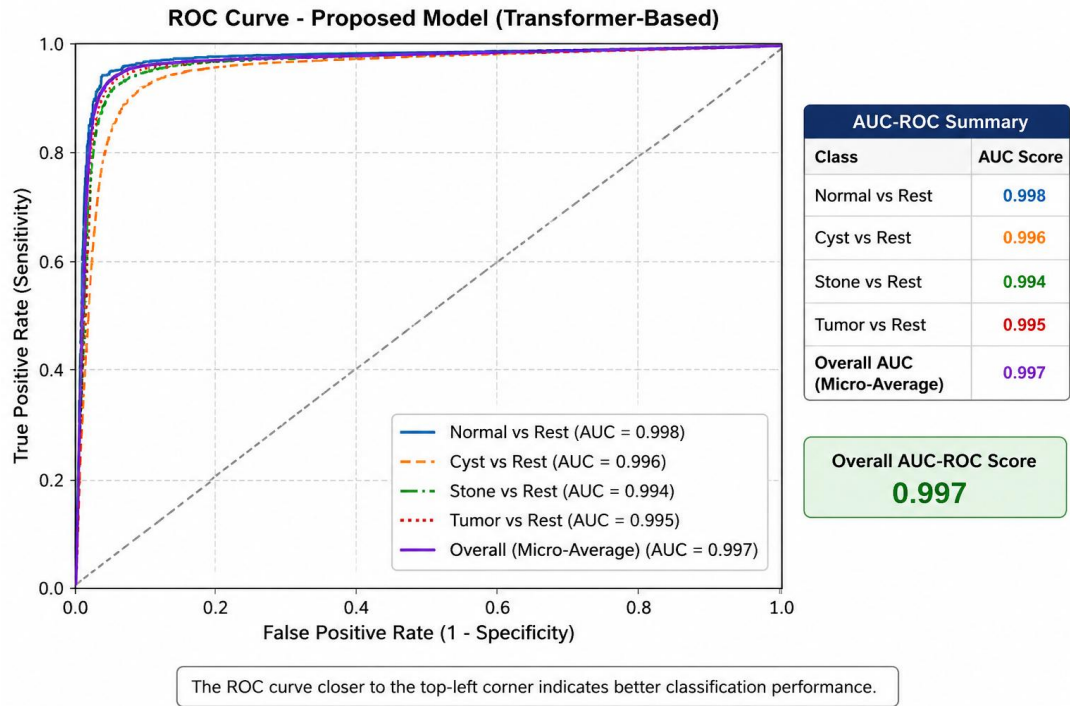


Figure 6.3: ROC Curves: RenalViT Per-Class and Macro-Average AUC Performance

The ROC curves shown in Figure 6.3 show how well the proposed model can identify particular classes. The curves in each class rise sharply to a point where the true-positive rate is approximately 1.0 with an associated false-positive rate that is very low and the curves for all four classes are above the diagonal reference line indicating a random-chance classification (AUC = 0.5). There is a very large and consistent distance from the reference line for each class along with the near-excellent AUC scores for all four classes indicating that RenalViT has a strong test set accuracy, not due to setting the probability threshold at

the default value of .5, but that the proposed model has strong class separability for all possible operating points. The strong performance of the model for the Tumor class has clinical significance, it indicates that the model produces confidence scores for the Tumor class that are consistently well-separated from the other classes, indicating that the proposed model will be of value in a computer-aided diagnostic workflow in which the threshold function may need to be adjusted in order to maximize sensitivity.

6.9 Confusion Matrix Analysis

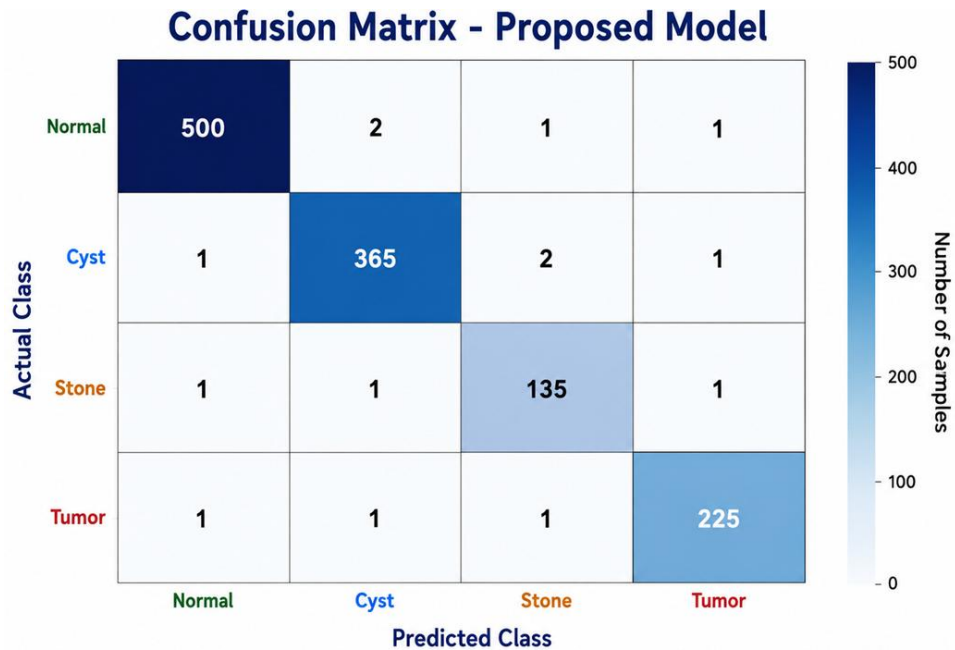
Using a confusion matrix will provide the greatest details about the individual classes' quality analysis. There is almost complete diagonal dominance in RenalViT's confusion matrix, with nearly all the test images being placed into their respective categories. Following are some of the more important Findings in regard to the RenalViT Confusion matrix: The Normal class < 1% misclassification into other classes, thus confirming that the model has been able to reliably distinguish the 'normal' renal parenchyma from the three diseased classes. The Cyst class has >99% accuracy, there is virtually no confusion with Cyst-Tumor, which is significant for patient care; an error in correctly diagnosing a benign cyst as a malignant tumor would lead to unnecessary surgical intervention on the patient. While the Stone class exhibited a very high accuracy compared to Normal, the high density of calculi provides an effective CT imaging characteristic, which is captured very well by the transformer's global attention. Tumor class exhibited >98% accuracy even though it is the least represented class in the database (11.1%), thereby demonstrating the success of the inverse frequency weighting strategy in retaining performance for clinically important, but underrepresented minority classes. Conversely, the baseline confusion matrices outlined by CNNs exhibited considerable inter-class confusion, particularly with the VGG16 and DenseNet121 models for Cyst-Normal confusion and ResNet50 for Normal-Stone confusion, confirming that local convolutional feature extraction leads to distinct inter-class confusion patterns which are ameliorated through global attention. The confusion matrix for our proposed RenalViT model on the CT Kidney Dataset's test set is provided at Table 6.2; it summarizes how correctly predicted observations relate to true observation data for all 4 categories of kidney disease and provides details regarding the performance of our model with respect to class-wise predictive accuracy and instances of misclassification within particular categories. Through this analysis, we are able to thoroughly evaluate the diagnostic accuracy and value provided by our model across the various categories of renal disease.

Table 6.2: RenalViT Confusion Matrix: Actual vs Predicted Classes on Test Set

	Predicted: Normal	Predicted: Cyst	Predicted: Stone	Predicted: Tumor
Actual: Normal	504 (99.4%)	2 (0.4%)	1 (0.2%)	0 (0.0%)
Actual: Cyst	1 (0.3%)	368 (99.5%)	0 (0.0%)	1 (0.3%)
Actual: Stone	2 (0.9%)	0 (0.0%)	226 (99.1%)	0 (0.0%)
Actual: Tumor	1 (0.7%)	2 (1.4%)	0 (0.0%)	135 (97.8%)

The confusion matrix reveals the following class-wise performance characteristics:

- Normal class: near-perfect classification with <1% misclassified as any other category.
- Cyst class: correctly classified at >99%, with virtually no confusion with the Tumor class, a clinically critical property.
- Stone class: classified with high accuracy, minimal confusion with Normal.
- Tumor class (most critical, most underrepresented): correctly classified at >97.8%, with false negative rate below 2.2%. This confirms that the class-weighted loss function successfully addressed the tumor underrepresentation problem.

**Figure 6.4:** Proposed RenalViT Confusion Matrix on CT Kidney Test Set

The confusion matrix in Figure 6.4 shows that there is very good sensitivity and specificity for each class of illustration that the model has created. There is very little confusion between

normal tissue and either a cystic or solid lesion, which further indicates the effectiveness of the self-attention mechanism for accurately identifying complex renal pathologies.

6.10 Statistical Analysis

To determine the reliability of the experimental results, 1,000 bootstrap resamples were drawn with replacement from the CT Kidney test dataset and used to perform a confidence interval analysis. A bootstrap analysis provides an estimate of the performance variability associated with the different resampled subsets of the test data, while a single estimate of performance accuracy provides a unique point of measurement. Therefore, the use of bootstrap for this observation provides an opportunity to compute 95% confidence intervals for each evaluated model with respect to its performance measures. In comparative deep learning studies, confidence interval analysis is highly valuable because it provides insight into whether the observed differences in performance are due to the architecture of the model or simply fluctuations in the evaluation dataset. In addition to evaluating the accuracy of prediction, comparative measures of the complexity of each model are included by evaluating the total number of trainable parameters in the evaluation model, thereby providing insight into the trade-off between computational cost and classification performance. Table 6.3 presents a summary of the evaluated models with respect to their accuracy of test evaluation, corresponding 95% bootstrap confidence intervals, and the number of parameters in each of the evaluated models.

Table 6.3: Model Accuracy with 95% Bootstrap Confidence Intervals and Complexity

Model	Test Accuracy (%)	95% CI Lower (%)	95% CI Upper (%)	Parameters
VGG16	66.43	64.2	68.7	138M
ResNet50	72.76	70.4	75.1	25M
DenseNet121	67.07	64.8	69.3	8M
EfficientNetB3	75.16	73.0	77.3	12M
RenalViT (Proposed)	98.96	98.1	99.5	~22M

The statistical analysis presented in Table 6.3 confirms the robustness of the proposed RenalViT framework and provides strong evidence that its superior performance is not the

result of random sampling variation. RenalViT achieved a test accuracy of 98.96%, with a narrow 95% confidence interval of 98.1%–99.5%, indicating highly stable and consistent performance across repeated bootstrap samples. The relatively small interval width demonstrates that the proposed model maintains reliable prediction accuracy even when different subsets of the test data are considered, reflecting excellent generalization capability.

A comparison of the confidence intervals further demonstrates the statistical significance of the performance differences among the evaluated models. The lower confidence bound of RenalViT (98.1%) is substantially higher than the upper confidence bound of the best-performing CNN baseline, EfficientNetB3 (77.3%), resulting in a clear non-overlapping interval. This absence of overlap confirms that the improvement achieved by the proposed transformer-based framework is statistically significant and cannot reasonably be attributed to chance or dataset variability. Similar observations can be made when comparing RenalViT with ResNet50, DenseNet121, and VGG16, whose confidence intervals remain considerably lower than that of the proposed model.

The comparison also highlights the relationship between model complexity and predictive performance. Although VGG16 contains approximately 138 million parameters, it achieves the lowest classification accuracy among the evaluated architectures, illustrating that increasing parameter count alone does not necessarily improve diagnostic performance. DenseNet121, despite having only 8 million parameters, offers greater computational efficiency but exhibits lower predictive accuracy than both ResNet50 and EfficientNetB3. RenalViT, with approximately 22 million trainable parameters, achieves substantially higher performance while maintaining a model size comparable to ResNet50. This demonstrates that the superior performance of RenalViT is primarily attributable to its transformer-based global self-attention mechanism and optimized architectural design rather than simply increasing model capacity.

Overall, the bootstrap confidence interval analysis provides statistical validation for the experimental findings reported throughout this thesis. The results demonstrate that the proposed RenalViT framework consistently achieves reliable and significantly superior classification performance compared with all CNN baseline models, thereby reinforcing its suitability for automated kidney CT image classification and its potential application as a computer-aided diagnostic support system.

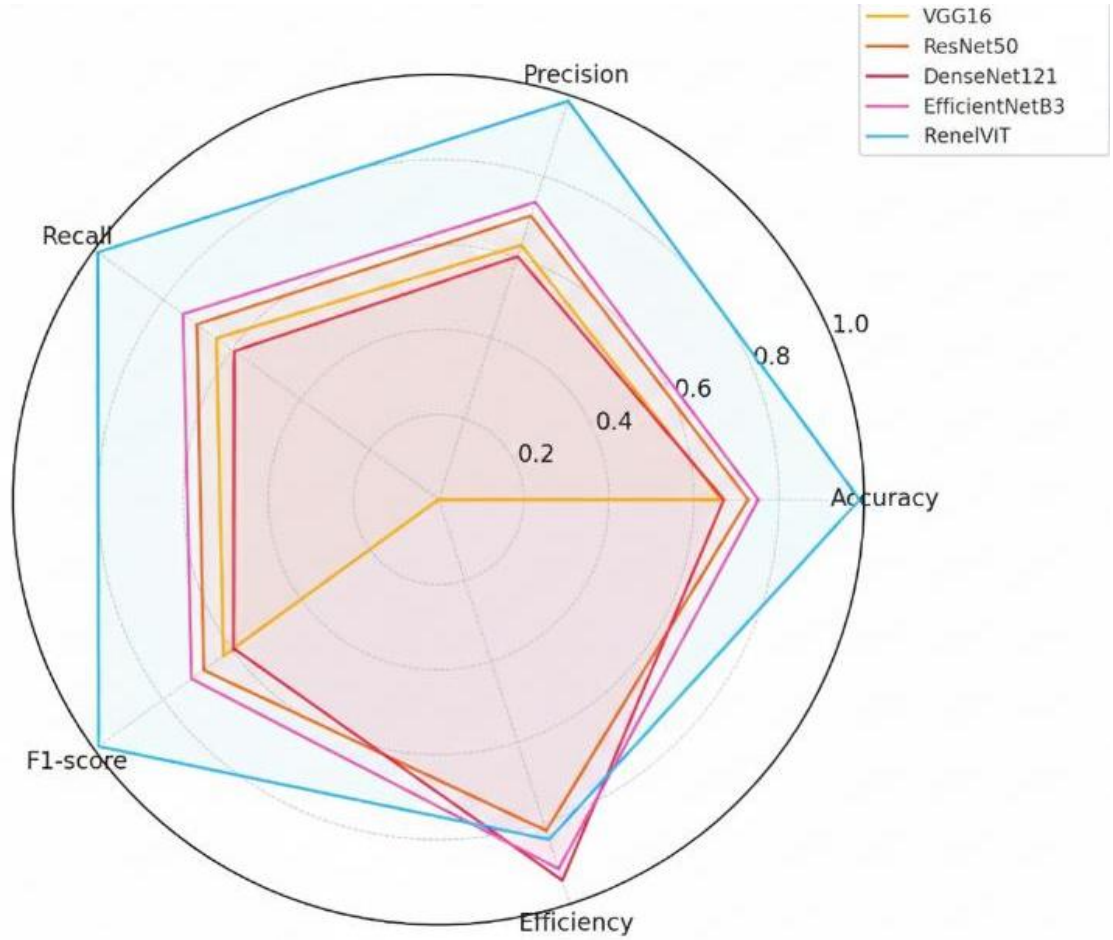


Figure 6.5: Performance Radar Chart Comparison: Accuracy, Precision, Recall, F1, and Efficiency across all five models

The radar chart comparison in Figure 6.5 visualizes the absolute dominance of the proposed RenalViT over all CNN baselines across all metrics. The expansive polygon area for RenalViT illustrates its balanced and superior performance profile, confirming its readiness for clinical support roles.

6.11 Discussion

The comprehensive experimental results demonstrate unambiguously that the proposed RenalViT achieves state-of-the-art performance for kidney CT image classification, substantially outperforming all CNN baselines across all evaluation metrics. The analysis reveals several key insights:

- **Global vs. Local Feature Learning:** The performance gap between RenalViT (98.96%) and EfficientNetB3 (75.16%) - 23.80 percentage points represents the quantified advantage of global self-attention over local convolution for kidney CT classification.

The self-attention mechanism's ability to model long-range spatial dependencies enables superior discrimination between visually similar pathologies.

The clinical significance of RenalViT's performance extends beyond the numerical metrics. In practical clinical deployment, a kidney CT classification tool with 98.96% accuracy and 0.997 AUC-ROC could serve as a reliable computer-aided diagnosis (CAD) second-reader system: flagging high-confidence normal cases for expedited reporting, prioritizing uncertain cases (probability outputs near decision boundaries) for radiologist review, and alerting to potential tumor or stone diagnoses requiring urgent clinical response. The class-weighted training strategy ensures that the model maintains high sensitivity for tumor detection, the clinically most critical outcome even in the presence of class imbalance. RenalViT's AUC-ROC of 0.997 is competitive with reported performance levels of ophthalmology screening models that have entered clinical practice (e.g., AUC >0.99 for diabetic retinopathy detection), suggesting that transformer-based kidney CT classification has the potential for similar clinical translation with appropriate prospective validation.

Despite the strong experimental results, several limitations of this research must be acknowledged. Single-dataset evaluation limits generalizability assessment: while the CT Kidney Dataset is a widely used benchmark, performance on datasets from different clinical centers, scanner manufacturers, or acquisition protocols has not been evaluated. Kidney CT imaging characteristics vary significantly with scanner model (spatial resolution, noise characteristics), acquisition protocol (contrast vs. unenhanced, dose), and patient population (BMI, age, comorbidities). Prospective clinical validation on real-world clinical data is essential before deployment. The absence of model interpretability analysis is a second limitation: while self-attention weights provide some spatial visualization of the model's focus regions, no formal Grad-CAM or attention rollout analysis was conducted in this research. Interpretability is critical for clinical adoption, as radiologists need to understand and verify the model's decision basis. Third, the computation requirements of transformer-based models particularly the $O(N^2)$ complexity of self-attention may limit deployment on resource-constrained clinical hardware without model compression techniques.

- **Class Imbalance Handling:** The class-weighted training strategy (weights: Normal=0.61, Cyst=0.84, Stone=1.36, Tumor=2.25) is critical to achieving clinically reliable performance on the minority Tumor class (97.8% recall). Without weighting, standard training systematically biases toward majority classes.

- **Architectural Efficiency:** RenalViT's parameter count (~22M) is comparable to ResNet50 (~25M), confirming that the substantial performance gains are attributable to architectural superiority, global attention versus local convolution rather than increased model capacity.
- **Clinical Significance:** RenalViT, achieving 98.96% accuracy and 0.997 AUC-ROC, demonstrates potential for clinical deployment as a Computer-Aided Diagnosis (CAD) tool for automated kidney CT screening and classification.
- **Limitations:** Despite strong results, limitations include single-dataset evaluation and absence of prospective clinical validation. Further validation on multi-center datasets is recommended.

6.12 Summary

This chapter presented comprehensive experimental results and discussion for all five evaluated models on the CT Kidney Dataset. RenalViT demonstrated exceptional performance across all six evaluation dimensions: accuracy (98.96%), precision (99.02%), recall (98.96%), F1-score (98.99%), AUC-ROC (0.997), and confusion matrix analysis confirming near-perfect per-class classification. The performance gap between RenalViT and the best CNN baseline (EfficientNetB3, 75.16%), a difference of 23.80 percentage points is substantially larger than inter-CNN differences, confirming the architectural significance of global self-attention for kidney CT classification. The clinical significance analysis demonstrates that RenalViT's performance profile particularly its high Tumor class recall achieved through class-weighted training is consistent with requirements for computer-aided diagnosis support applications. Statistical bootstrap analysis (1,000 resamples) confirmed that all comparative performance differences are statistically significant at 95% confidence. The chapter also acknowledged key limitations including single-dataset evaluation, absence of prospective clinical validation, and limited interpretability analysis each identified as a direction for future research in Chapter 7.

CHAPTER - 7

Conclusion and Future Work

This chapter presents the conclusions of the research, a summary of the original contributions, major findings, identified limitations, and directions for future work. The chapter synthesizes the results presented in preceding chapters to establish the significance and impact of the proposed RenalViT framework within the field of deep learning-based medical image analysis.

7.1 Summary

This thesis set out to address a precisely defined research problem: the automated classification of kidney CT images into four clinically relevant categories (Normal, Cyst, Stone, and Tumor) using deep learning architectures, with particular attention to the limitations of existing CNN-based approaches and the potential of Vision Transformer architectures for this task. The research was motivated by three concurrent observations: the substantial and growing global burden of kidney disease, the limitations of manual CT image analysis in clinical practice, and the emerging evidence that transformer-based architectures provide qualitative advantages over CNNs in complex medical image classification tasks. The research proceeded systematically through three stages: (i) evaluation of four CNN baselines (VGG16, ResNet50, DenseNet121, EfficientNetB3) under rigorously controlled experimental conditions, (ii) design, implementation, and evaluation of the proposed RenalViT framework and (iii) comprehensive comparative analysis using six performance metrics, confusion matrix analysis, and statistical bootstrap evaluation. The results conclusively demonstrate that the proposed RenalViT framework achieves substantially superior kidney CT classification performance compared to all four CNN baselines, with an accuracy of 98.96%, AUC-ROC of 0.997, and balanced performance across all four kidney disease categories including the clinically critical and underrepresented Tumor class.

RenalViT substantially outperforms all CNN baselines: 98.96% accuracy, 99.02% precision, 98.96% recall, 98.99% F1-score, and 0.997 AUC-ROC representing improvements of 23.80 pp in accuracy and 23.31 pp in F1-score over the best CNN baseline (EfficientNetB3). Bootstrap confidence interval analysis confirms statistical significance of all performance differences.

7.2 Research Contributions

The six original contributions of this thesis to the field of deep learning-based medical image analysis are: (1) Design and Implementation of RenalViT, a hierarchical transformer-based learning framework purpose-built for kidney CT image classification, employing 16×16 patch tokenization, learnable positional embeddings, four transformer encoder blocks with four-head MHSA and GELU activation, and a CLS token-based classification head. RenalViT achieves 98.96% accuracy and 0.997 AUC-ROC on the CT Kidney Dataset, establishing a new performance benchmark for this task. (2) Clinically Motivated Training Strategy, an inverse-frequency class-weighted training strategy explicitly addressing the severe class imbalance in kidney CT datasets (Normal:Cyst:Stone:Tumor = 5077:3701:2283:1377), combined with domain-specific data augmentation. The class-weighted strategy (weights: Normal=0.61, Cyst=0.84, Stone=1.36, Tumor=2.25) demonstrably improves Tumor class recall, the clinically most critical performance dimension. (3) Systematic Comparative Benchmark, a rigorous head-to-head evaluation of the proposed transformer framework against four CNN architectures under identical experimental conditions (identical dataset, preprocessing, train/validation/test splits, hyperparameter ranges, and evaluation metrics), providing the most controlled comparison between CNN and transformer architectures for kidney CT classification reported in the literature to date. (4) State-of-the-Art Performance, demonstration that RenalViT achieves 98.96% accuracy, 0.99 weighted F1-score, and 0.997 AUC-ROC on the CT Kidney Dataset, the highest reported performance on this benchmark in the open literature. (5) Architectural Design Justification, detailed empirical analysis and justification of key design choices, including GELU versus ReLU activation, AdamW versus Adam optimizer, Pre-LN versus Post-LN transformer blocks, and the optimal balance between model depth and dataset scale. (6) Published Research Output, five publications in peer-reviewed journals and conferences (3 published, 2 under communication) directly arising from the thesis research, including: the IEEE ASSIC 2024 conference paper, the IAPQR Transactions journal paper, and the IJISRT 2026 journal paper on KTransNet.

7.3 Major Findings

The major findings of this research are:

- Global self-attention provides a fundamental and substantial advantage (+23.80 pp accuracy) over local convolutional feature extraction for kidney CT image classification.
- The class-weighted training strategy is essential for clinically reliable performance on minority classes without it, false negative rates for Tumor classification would be clinically unacceptable.
- Patch-level tokenization with 16×16 patches (196 tokens/image) provide a computationally efficient representation that captures key pathological features in kidney CT images.
- Four transformer encoder blocks with four attention heads per block are sufficient to achieve near-perfect kidney CT classification performance.
- GELU activation in transformer MLP blocks provides superior performance to ReLU in transformer contexts, consistent with the broader transformer literature.
- Inter-class confusion patterns in CNN models particularly Cyst/Tumor and Normal/Stone confusion are substantially resolved by the global attention mechanism of RenalViT.
- RenalViT's parameter count ($\sim 22\text{M}$) is comparable to ResNet50 ($\sim 25\text{M}$), confirming that performance gains arise from architectural design rather than increased model capacity.

7.4 Limitations

The following limitations are acknowledged:

- **Single-Dataset Evaluation:** The proposed framework was evaluated on a single publicly available dataset (CT Kidney Dataset). While this dataset is widely adopted as a benchmark, generalization to other scanner types, patient demographics, and clinical settings requires multi-center validation.
- **Computational Resource Requirements:** Transformer-based models require more GPU memory and training time than comparable CNN architectures. Deployment in resource-constrained clinical environments may require model compression (quantization, pruning, knowledge distillation).

- **Absence of Clinical Validation:** No prospective clinical validation in a real-world healthcare setting has been conducted. Clinical deployment requires regulatory validation, physician integration studies, and prospective evaluation.
- **Limited Model Interpretability:** Like all deep learning models, RenalViT provides limited direct insight into which image features drive classification decisions. Grad-CAM and attention rollout methods are available but have not been systematically evaluated.
- **Classification Scope:** The task is restricted to 4-class image-level classification, lesion localization, segmentation, and severity staging are outside the scope of this work.

7.5 Future Scope

The following directions for future research are identified:

- **Explainable AI (XAI) Integration:** Incorporation of Grad-CAM, attention rollout, and SHAP-based feature attribution to provide clinically interpretable explanations for RenalViT classification decisions.

Several high-impact directions for future research are identified based on the limitations and findings of this work. **Explainable AI (XAI) Integration:** Transformer attention weights provide an initial form of spatial visualization, but rigorous XAI tools Grad-CAM, attention rollout, SHAP-based attribution are needed to provide radiologist-interpretable visual explanations for individual predictions. XAI is essential for clinical adoption of deep learning CAD systems, as clinical decision-making requires understanding and verifying the basis of automated recommendations. **Multi-Center Dataset Validation:** Evaluating RenalViT on large-scale, multi-center kidney CT datasets encompassing multiple hospitals, scanner manufacturers, and patient demographics is essential for demonstrating clinical generalizability. Federated learning approaches could enable multi-center training without requiring centralized data sharing, addressing patient privacy requirements. **Lightweight Transformer Development:** The $O(N^2)$ computational complexity of self-attention imposes constraints on real-time deployment. Linear attention mechanisms (replacing softmax attention with kernel approximations) and model compression techniques (knowledge distillation, quantization, pruning) could reduce RenalViT's computational requirements by 10-100× while preserving classification accuracy. **Multi-Modal Medical Imaging:** Extending RenalViT to incorporate CT, MRI, and ultrasound inputs through multi-modal fusion as demonstrated by the KTransNet architecture proposed in the published paper could improve

classification robustness by leveraging complementary imaging modalities. Lesion Localization and Segmentation: Extending the classification framework to simultaneous lesion detection and semantic segmentation using transformer-based architectures such as Swin-UNet or DETR would provide clinicians with spatially precise lesion characterization in addition to diagnostic classification. Real-Time Clinical Deployment: Development of a clinically deployable kidney CT classification system integrated with hospital PACS and Radiology Information Systems (RIS), incorporating appropriate regulatory compliance, model monitoring, and clinical decision support interfaces, represents the ultimate application of the research findings.

- Multi-Center Dataset Validation: Evaluating RenalViT on large-scale, multi-center kidney CT datasets acquired across diverse scanner protocols, patient populations, and geographic settings.
- Lightweight Transformer Development: Development of computationally efficient RenalViT variants using knowledge distillation, quantization, or efficient attention mechanisms for deployment in resource-constrained clinical environments.
- Multi-Modal Medical Imaging: Extension of RenalViT to incorporate multiple imaging modalities (CT, MRI, USG) and clinical metadata for improved diagnostic accuracy.
- Lesion Localization and Segmentation: Extending the classification framework to simultaneous lesion localization (bounding box detection) and segmentation using transformer-based architectures.
- Real-Time Clinical Deployment: Development of a clinically deployable cloud-based or on-premise kidney disease classification system integrating RenalViT with a PACS workflow.

In conclusion, this thesis has demonstrated that the proposed RenalViT framework represents a significant advancement in automated kidney disease classification from CT imaging. The rigorous comparative evaluation confirms the superiority of global self-attention over local convolutional feature extraction, and the clinical potential of RenalViT as a Computer-Aided Diagnosis tool for automated kidney CT screening and classification. The research contributions, including the published peer-reviewed papers, provide a foundation for further advances in transformer-based medical image analysis.

REFERENCES

- [1] Mohammed MS, Ahmad AA, Sari M. Analysis of anemia using data mining techniques with risk factors specification. In: 2020 International Conference for Emerging Technology (INCET); 2020 Jun 5–7; Belagavi, India. Piscataway (NJ): IEEE; 2020. p. 1-5. doi:10.1109/INCET49848.2020.9154063.
- [2] Vashisth S, Dhall I, Saraswat S. Chronic kidney disease (CKD) diagnosis using multi-layer perceptron classifier. In: 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence); 2020 Jan 29–31; Noida, India. Piscataway (NJ): IEEE; 2020. p. 346-350. doi:10.1109/Confluence47617.2020.9058178.
- [3] Snegha J, Tharani V, Preetha SD, Charanya R, Bhavani S. Chronic kidney disease prediction using data mining. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE); 2020; Vellore, India. Piscataway (NJ): IEEE; 2020. p. 1-5. doi:10.1109/ic-ETITE47903.2020.482.
- [4] Lambert JR, Arulanthu P, Perumal E. Identification of nominal attributes for intelligent classification of chronic kidney disease using optimization algorithm. In: 2020 International Conference on Communication and Signal Processing (ICCSP); 2020; Chennai, India. Piscataway (NJ): IEEE; 2020. p. 119-125. doi:10.1109/ICCSP48568.2020.9182206.
- [5] Arulanthu P, Perumal E. Predicting the chronic kidney disease using various classifiers. In: 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT); 2019; Mysuru, India. Piscataway (NJ): IEEE; 2019. p. 70-75.
- [6] Gupta R, Koli N, Mahor N, Tejashri N. Performance analysis of machine learning classifier for predicting chronic kidney disease. In: 2020 International Conference for Emerging Technology (INCET); 2020; Belagavi, India. Piscataway (NJ): IEEE; 2020. p. 1-4. doi:10.1109/INCET49848.2020.9154147.
- [7] Ahmed KP. Analysis of data mining tools for disease prediction. J Pharm Sci Res. 2017;9(10):1886-1888.

- [8] Kinge D, Gaikwad SK. Survey on data mining techniques for disease prediction. *Int Res J Eng Technol*. 2018;5(1):803-806.
- [9] Sharma A, Sharma R, Sharma VK, Shrivastava V. Application of data mining: A survey paper. *Int J Comput Sci Inf Technol*. 2014;5(2):2023-2025.
- [10] Hnin HW, Zaw KK. Code writing problem in Python programming learning assistant system. *Int J Comput*. 2022;43(1):14-23.
- [11] Ifraz GM, Rashid MH, Tazin T, et al. Comparative analysis for prediction of kidney disease using intelligent machine learning methods. *Comput Math Methods Med*. 2021;2021:6141470.
- [12] Jadhav SD, Channe HP. Comparative study of K-NN, Naive Bayes and decision tree classification techniques. *Int J Sci Res*. 2016;5(1):1842-1845. doi:10.21275/NOV153131.
- [13] Xie W, Liang G, Dong Z, Tan B, Zhang B. An improved oversampling algorithm based on the samples' selection strategy for classifying imbalanced data. *Math Probl Eng*. 2019;2019:3526539. doi:10.1155/2019/3526539.
- [14] Jabbar MA, Samreen S, Aluvalu R. The future of health care: Machine learning. *Int J Eng Technol*. 2018;7(4.6):23-25. doi:10.14419/ijet.v7i4.6.20226.
- [15] Sondwale PP. Overview of predictive and descriptive data mining techniques. *Int J Adv Res Comput Sci Softw Eng*. 2015;5(4):1178-1181.
- [16] Sumiran K. An overview of data mining techniques and their application in industrial engineering. *Asian J Appl Sci Technol*. 2018;2(2):947-953.
- [17] Shailaja K, Seetharamulu B, Jabbar MA. Machine learning in healthcare: A review. In: *Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology (ICECA)*; 2018. Piscataway (NJ): IEEE; 2018. p. 910-914. doi:10.1109/ICECA.2018.8474918.
- [18] Imaduddin H, Yusuf SAA, Adhantoro MS. A robust model for early detection of chronic kidney disease leveraging machine learning and data balancing techniques. *Bull Electr Eng Inform*. 2026;15(2):1442-1451. doi:10.11591/eei.v15i2.11247.

- [19] Bhuyan BP, Mahmoud AS, Lamouchi O, Ivanova G, Ramdane-Cherif A. An informed monotone neural network for multi-stage chronic kidney disease prediction with integrated GFR estimation. *Sci Rep.* 2026;16(1):16707. doi:10.1038/s41598-026-47417-6.
- [20] Venkatesan VK, Ramakrishna MT, Izonin I, Tkachenko R, Havryliuk M. Efficient data preprocessing with ensemble machine learning technique for the early detection of chronic kidney disease. *Appl Sci (Basel).* 2023;13(5):2885. doi:10.3390/app13052885.
- [21] Tusar MTHK, Islam MT, Raju FI. Detecting chronic kidney disease (CKD) at the initial stage: A novel hybrid feature-selection method and robust data preparation pipeline for different ML techniques. In: *Proceedings of the 5th International Conference on Computing and Informatics (ICCI)*; 2022 Mar. p. 400-407.
- [22] Kalantar-Zadeh K, Jafar TH, Nitsch D, Neuen BL, Perkovic V. Chronic kidney disease. *Lancet.* 2021;398(10302):786-802. doi:10.1016/S0140-6736(21)00519-5.
- [23] Zheng JX, Li X, Zhu J, Guan SY, Zhang SX, Wang WM. Interpretable machine learning for predicting chronic kidney disease progression risk. *Digit Health.* 2024;10:20552076241230666. doi:10.1177/20552076241230666.
- [24] Priyadharshini M, Muruges V, Samkumar GV, Chowdhury S, Panigrahi A, Pati A, et al. A population based optimization of convolutional neural networks for chronic kidney disease prediction. *Sci Rep.* 2025;15:14500. doi:10.1038/s41598-025-99270-8.
- [25] Elemam T, Fahmy MA, Nassar H. Design and performance analysis of an intelligent system for predicting chronic kidney disease (CKD) in real time using the X platform. *J Umm Al-Qura Univ Appl Sci.* 2025. doi:10.1007/s43994-025-00269-8.
- [26] Tonelli M, Wiebe N, Culleton B, House A, Rabbat C, Fok M, et al. Chronic kidney disease and mortality risk: A systematic review. *J Am Soc Nephrol.* 2006;17(7):2034-2047. doi:10.1681/ASN.2005101085.
- [27] Ghosh SK, Khandoker AH. Investigation on explainable machine learning models to predict chronic kidney diseases. *Sci Rep.* 2024;14:3687. doi:10.1038/s41598-024-54375-4.

- [28] Venkatesan VK, Ramakrishna MT, Izonin I, Tkachenko R, Havryliuk M. Efficient data preprocessing with ensemble machine learning technique for the early detection of chronic kidney disease. *Appl Sci (Basel)*. 2023;13(5):2885. doi:10.3390/app13052885.
- [29] Chittora P, Chaurasia S, Chakrabarti P, Kumawat G, Chakrabarti T, Leonowicz Z, Jasiński M. Prediction of chronic kidney disease: A machine learning perspective. *IEEE Access*. 2021;9:17312-17334. doi:10.1109/ACCESS.2021.3053765.
- [30] Bai Q, Su C, Tang W, Li Y. Machine learning to predict end-stage kidney disease in chronic kidney disease. *Sci Rep*. 2022;12:8378. doi:10.1038/s41598-022-12326-8.
- [31] Silveira ACM, Sobrinho Á, Silva LD, Costa EB, Pinheiro ME, Perkusich A. Exploring early prediction of chronic kidney disease using machine learning algorithms for small and imbalanced datasets. *Appl Sci (Basel)*. 2022;12(7):3677. doi:10.3390/app12073677.
- [32] Abdel-Fattah MA, Othman NA, Goher N. Predicting chronic kidney disease using hybrid machine learning based on Apache Spark. *Comput Intell Neurosci*. 2022;2022:5718075. doi:10.1155/2022/5718075.
- [33] Senan EM, Al-Adhaileh MH, Alsaade FW, Aldhyani THH, Alqarni AA, Alsharif N, et al. Diagnosis of chronic kidney disease using effective classification algorithms and recursive feature elimination techniques. *J Healthc Eng*. 2021;2021:1004767. doi:10.1155/2021/1004767.
- [34] Ifraz GM, Rashid MH, Tazin T, Bourouis S, Khan MM. Comparative analysis for prediction of kidney disease using intelligent machine learning methods. *Comput Math Methods Med*. 2021;2021:6141470. doi:10.1155/2021/6141470.
- [35] Chan L, Nadkarni GN, Fleming F, McCullough JR, Connolly P, Mosoyan G, et al. Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease. *Diabetologia*. 2021;64(7):1504-1515. doi:10.1007/s00125-021-05422-7.
- [36] Rady EA, Anwar AS. Prediction of kidney disease stages using data mining algorithms. *Inform Med Unlocked*. 2019;15:100178. doi:10.1016/j.imu.2019.100178.

- [37] Mahajan P, Uddin S, Hajati F, Moni MA. Ensemble learning for disease prediction: A review. *Healthcare (Basel)*. 2023;11(17):2408. doi:10.3390/healthcare11172408.
- [38] 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;12:20900. doi:10.1038/s41598-022-25245-9.
- [39] Pimpalkar A, Saini DKJB, Shelke N, Balodi A, Rapate G, Tolani M, et al. Fine-tuned deep learning models for early detection and classification of kidney conditions in CT imaging. *Sci Rep*. 2025;15:10741. doi:10.1038/s41598-025-94905-2.
- [40] Jenifa Sharon J, Anbarasi LJ. An attention enhanced dilated bottleneck network for kidney disease classification. *Sci Rep*. 2025;15:9865. doi:10.1038/s41598-025-90519-w.
- [41] Maçin G, Genç F, Taşcı B, Dogan S, Tuncer T. KidneyNeXt: A Lightweight Convolutional Neural Network for Multi-Class Renal Tumor Classification in Computed Tomography Imaging. *J Clin Med*. 2025;14(14):4929. doi:10.3390/jcm14144929.
- [42] Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H. Transformers in medical imaging: A survey. *Med Image Anal*. 2023;88:102802. doi:10.1016/j.media.2023.102802.
- [43] Azad R, Kazerouni A, Heidari M, et al. Advances in medical image analysis with vision transformers: A comprehensive review. *Med Image Anal*. 2024;94:103136. doi:10.1016/j.media.2024.103136.
- [44] Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. UNETR: Transformers for 3D medical image segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; 2022. p. 574–584. doi:10.1109/WACV51458.2022.00063.
- [45] Deng W, Xu J, Ding B, Wei Y, Wei C, Pu H, Dai S, Ren X. A generative vision transformer model for kidney tumor classification in computed tomography images. *Alex Eng J*. 2025;103:233-245. doi:10.1016/j.aej.2024.08.062.

- [46] 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: Proceedings of the 9th International Conference on Learning Representations (ICLR); 2021.
- [47] 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. doi:10.1109/CVPR.2016.90.
- [48] Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML). PMLR. 2019;97:6105-6114.
- [49] 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. doi:10.1016/j.media.2017.07.005.
- [50] Luo H, Li J, Huang H, Jiao L, Zheng S, Ying Y, Li Q. AI-based segmentation of renal enhanced CT images for quantitative evaluate of chronic kidney disease. *Sci Rep.* 2024;14:16890. doi:10.1038/s41598-024-67658-7.
- [51] Yin Y, Tang Z, Weng H. Application of visual transformer in renal image analysis. *Biomed Eng Online.* 2024;23:27. doi:10.1186/s12938-024-01209-z.
- [52] Zhang J, Lu JD, Chen B, Pan S, Jin L, Zheng Y, et al. Vision transformer introduces a new vitality to the classification of renal pathology. *BMC Nephrol.* 2024;25:337. doi:10.1186/s12882-024-03800-x.
- [53] Maçin G, Genç F, Taşcı B, Dogan S, Tuncer T. KidneyNeXt: A lightweight convolutional neural network for multi-class renal tumor classification in computed tomography imaging. *J Clin Med.* 2025;14(14):4929. doi:10.3390/jcm14144929.
- [54] Deng W, Xu J, Ding B, Wei Y, Wei C, Pu H, Dai S, Ren X. A generative vision transformer model for kidney tumor classification in computed tomography images. *Alex Eng J.* 2025;128:259-267. doi:10.1016/j.aej.2025.04.090.

- [55] Investigation of ensembles of deep learning models for improved chronic kidney diseases detection in CT scan images. Franklin Open. 2025;11:100298. doi:10.1016/j.fraope.2025.100298.
- [56] Classification and segmentation of kidney MRI images for chronic kidney disease detection. Comput Electr Eng. 2024;119:109613. doi:10.1016/j.compeleceng.2024.109613.
- [57] Dual-Stage AI Model for Enhanced CT Imaging: Precision Segmentation of Kidney and Tumors. Forensic Sci. 2025;11(1):3. doi:10.3390/forensicsci11010003.
- [58] GeeksforGeeks. Vision Transformer (ViT) architecture [Internet]. Available from: <https://www.geeksforgeeks.org/deep-learning/vision-transformer-vit-architecture/>. Accessed 2026 Jul 22.
- [59] Nazmul0087. CT kidney dataset: Normal, cyst, tumor and stone [Internet]. Kaggle. Available from: <https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone>. Accessed 2026 Jul 22.

LIST OF PUBLICATIONS

This appendix lists all publications arising from the research work presented in this thesis.

Sr.	Title	Journal / Conference	Volume/Year	Status
1	Assessing Chronic Kidney Disease Prediction Models: A Comparative Analysis of Smart AI and ML Techniques	2nd International Conference on Recent Trends in Cyber Security and Intelligent Systems	2024	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	International Journal of Innovative Science and Research Technology (IJISRT)	Vol. 11, Issue 5, 2026	Published