ANALYSIS OF TRANSFORMER AND CNN BASED APPROACHES FOR CLASSIFYING RENAL ABNORMALITY FROM IMAGE DATA

by

S. M. Mushfiq Reza
20101254
Abu Bakar Hasnath
20301037
Ankita Roy
23141059
Amreen Rahman
20301479
Abdullah Bin Faruk
23341108

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
June 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

S. M. Mushfiq Reza
20101254

Abu Bakar Hasnath
20301037

Ankita Roy
23141059

Amreen Rahman
20301479

Abdullah Bin Faruk
23341108

Approval

The thesis/project titled "ANALYSIS OF TRANSFORMER AND CNN BASED APPROACHES FOR CLASSIFYING RENAL ABNORMALITY FROM IMAGE DATA" submitted by

1. S. M. Mushfiq Reza(20101254)
2. Abu Bakar Hasnath(20301037)
3. Ankita Roy(23141059)
4. Amreen Rahman(20301479)
5. Abdullah Bin Faruk(23341108)

Of Spring, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on June 30, 2024.

Examining Committee:

Supervisor:

____

Md. Tanzim Reza
Lecturer
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:

____

Dr. Farig Yousuf Sadeque
Associate Professor
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:

Dr. Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

There is a pressing need to revise the current diagnostic framework for renal abnormality due to the projected increase in its global prevalence as about 10% of people worldwide are suffering from renal diseases. Recognizing the escalating trends of renal disease, proactive measures are warranted to overcome upcoming challenges in accurate diagnosis and management. Renal abnormalities, often symptomless and hard to diagnose, can be dangerous but curable if detected early. Therefore, machine learning and deep learning techniques can be instrumental if implemented correctly to determine this anomaly early in this modern time. Our approach for renal abnormality detection from image data incorporates the topologies of Convolutional Neural Networks and transformer-based image classification topologies, as well as data augmentation methods and precise hyperparameter tuning (learning rate, batch size, dropout rate, regularization strength, etc.); additionally, we proposed CNN-based and transformer-based architectures for renal abnormality detection. Transformer-based deep learning methods are the latest trend in classifying diseases from medical images; for this reason, we analyzed the performance of CNN-based architectures and transformer-based architectures. We build a hybrid binary class dataset of Computed Tomography(CT) scan renal images using primary data collected from Kidney Foundation Hospital & Research Institute, Dhaka, Bangladesh and secondary data from publicly available online source. Our approach is a sequence of steps that allows for the abnormality detection using state-of-the-art classifiers ResNet50, Inception ResNetV2, InceptionV3 and VGG16 along with our proposed ResNet152 based custom model and ViT architecture-based custom model without manual intervention. Our experimental results showed that our proposed transformer-based model achieved the highest accuracy of 99.99% while our proposed CNN model achieved an accuracy of 99.97%. Among the four pre-trained CNN models, ResNet50 scored the highest accuracy of 99.95%, and VGG16 scored 99.92%, InceptionResNetV2 was able to score 98.87%, while the lowest performance was shown by the InceptionV3 model, which was 96.87%. All four pre-trained models have demonstrated acceptable performance, and our proposed model was able to perform better than state-of-the-art prepared models.

Keywords: Deep Learning; Convolutional Neural Networks; Transformers; Renal Abnormality; ResNet50; Inception ResNetV2; InceptionV3; VGG16; ViT; ResNet152;

Dedication

Firstly, all praise to the Great Allah for Whom our thesis have been completed without any major interruption.

Secondly, to our Supervisor Mr. Md. Tanzim Reza sir and our Co-Supervisor Mr. Dr. Farig Yousuf Sadeque sir for their kind support and advice in our work. They helped us whenever we needed help.

Thirdly, the whole judging panel helped us a lot in our later works.

And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Acknowledgment

I would like to express my deepest gratitude to the Kidney Foundation Hospital Research Institute for the collaboration with our thesis. And I am extremely grateful to those people who have helped us to achieve our desired dataset. Special thanks to Dr. Harun Ur Rashid, Professor and Chief Consultant, Nephrology, Kidney Foundation Hospital Research Institute, and Dr. Amit Bari, Assistant Professor, Nephrology, Kidney Foundation Hospital & Research Institute for allowing us to collect the 3D CT Scan Image of Lower Abdomen from their valuable patients.

I would like to extend my sincere thanks to my friends who helped me with the thesis work and motivated me to achieve the goal.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

ANN	Artificial Neural Network
AUC	Area Under Curve
ccRCC	clear cell Renal Cell Carcinoma
CKD	Chronic Kidney Disease
CNN	Convolutional Neural Network
CTScan	Computed Tomography Scan
DNA	Deoxyribonucleic Acid
DNN	Deep Neural Network
DSC	Dice Similarity Coefficient
DT	Decision Tree
eGFR	Estimated Glomerular Filtration Rate
FPN	Feature Pyramid Network
GELU	Gaussian Error Linear Unit
HE	Hematoxylin Eosin
HBP	High Blood Pressure
IOU	Intersections Over Union
KNN	K-Nearest Neighbors
KUB	Kidney, Ureter & Bladder X-ray
MLP	Multi Layer Perceptron
MRIScan	Magnetic Resonance Imaging Scan
PACSs	Picture Archiving and Sharing Systems
PCA	Principal Component Analysis

PKD	Polycystic Kidney Disease
PR	Precision-Recall
pRCC	papillary Renal Cell Carcinoma
RCC	Renal Cell Cancer
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
SLE	Systemic Lupus Erythematosus
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
VGG	Visual Geometry Group
ViT	Vision Transformer
WHO	World Health Organization

Chapter 1

Introduction

1.1 Background of Renal Abnormality

The kidney, a vital organ of the human body, helps to filter out waste products from blood from the body to make urine. The primary kidney filtration units are called nephrons. The portion of the nephron where electrolytes are balanced and water is absorbed is termed as the renal tubule. The term ”renal” describes the kidneys. The kidney might face any major or minor problems at any stage of life. People usually are not able to acknowledge or realize this crucial issue until they feel discomfort at its last stage. As a result, doctors are unable to diagnose the patient’s kidney. Renal abnormalities surround a wide range of conditions that impact kidney functions and structure. In different ways, these conditions can manifest and also affect the body’s ability to maintain fluid, regulate blood pressure, and urinate waste products. The kidney plays a vital role in our body. Its disability can lead to consequence systematic health problems which is why understanding renal abnormality is very important. For renal abnormalities, Preventative strategies focus on managing risk aspects for example hypertension, diabetes, and exposure to nephrotoxins. Lifestyle modifications, pharmacological interventions, and regular observation of kidney function in at-risk populations are necessary components of these strategies.

To avoid and mitigate kidney-related health problems, understanding the type of causes, pathophysiology, and management of these disorders is crucial. To diagnose and eventually enhance outcomes for individuals with renal abnormalities, research, and clinical advancements are still continuous to enhance our ability.

1.2 Types of Renal Abnormality

Genetically renal abnormality can affect anyone. Most of the time, renal abnormality does not pass from father to child. But if any family has a history of renal abnormality in their family, it can affect multiple generations[1]. There are different types of renal abnormality, but some are discussed bellow:

Chronic kidney disease - is formed when the kidney has been damaged for more than three months and the patients have a hard time doing all the work. This disease starts to develop in a very slow process with some minor symptoms like urine like foam, appetite loss, going to pee more often or less than normal cases, and itchy skins. Some patients live with chronic kidney disease without any types of symptoms until the disease develops to advanced stages. Chronic Kidney Disease(CKD) has 5 stages identified by the eGFR. In the last stage of CKD, the kidney totally fails and the patient needs to go through the dialysis process.

Renal Hydronephrosis - is a type of disease which develops due to problems with urination and gathering inside the kidney. Some symptoms might show up like urine with blood, painful urination, low urination, etc. Hydronephrosis refers to dilation of the urinary tract which blocks the urine to flow to the bladder from the kidney. When one kidney gets affected then it is called unilateral hydronephrosis and when both kidneys get affected then that is called bilateral hydronephrosis, if both kidneys get affection, it can lead to kidney functionality loss and even kidney failure.

Glomerulonephritis - is a disease which develops when the glomerulus gets damaged. Glomerulus is a part of kidney which helps to filter out toxic, excess fluid from the body. Due to glomerulonephritis, toxic fluid and wastes can not be filtered out, rather it gathers up in the kidney and the kidney swells up. Some symptoms can be seen like, foamy urine and color of the urine become pink, the face of the patient swells up and blood pressure rises.

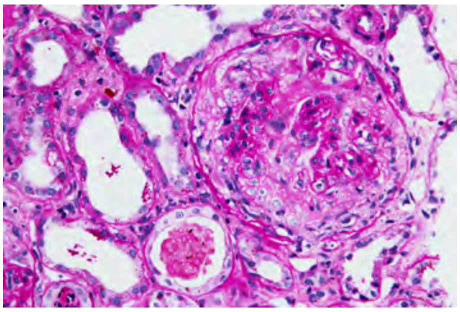

Figure 1.1: Glomerulonephritis

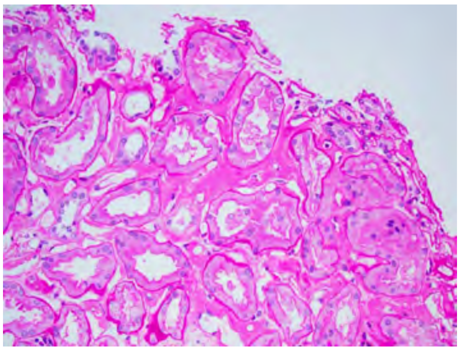

Figure 1.2: Normal Image

Kidney stones - is a form of hard stone like material which develops due to excess deposit of salt and minerals inside the kidney. Kidney stone does not do any permanent damage to the kidney but it can create problems in urination due to the crystals sticking together and it can damage the urinary tract from kidney to bladder.

Kidney cancer - The malignant cell or abnormal growth of cells that forms in these tubules is commonly called renal cancer or kidney cancer [19]. The malignant cells form due to mutation of DNA in between cells. However, the cells develop and spread rapidly so that cells can break free and affect other cells of different parts of the body. Most of the cells of the bones and lungs are mostly affected. There is no specific reason why these malignant cells are formed in renal tubules, but doctors researched some of the causes of renal cancer. If someone has a bad habit of chain smoking, then he has a lot of risk for renal cancer. It depends on how much one smokes. If someone is a non-smoker, the risk is reduced a lot. Patients with obesity or HBP are at the highest possibility of having renal cancer. If any patient is already suffering from kidney failure and going through regular dialysis has a possibility of renal cancer.

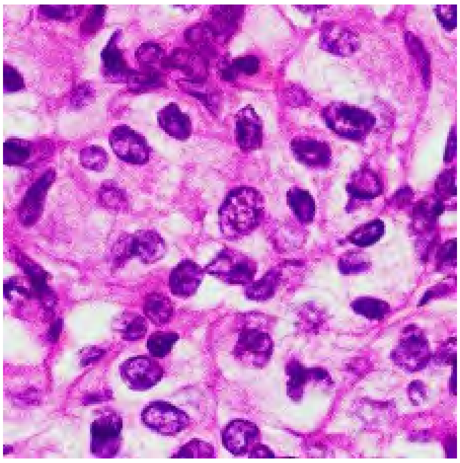

Figure 1.3: Cancer Cell

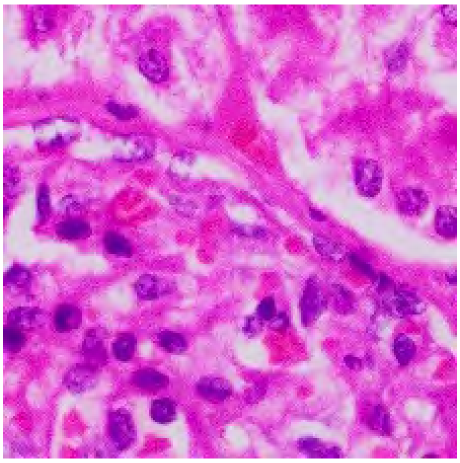

Figure 1.4: Normal Image

Lupus Nephritis - is a type of kidney disease which develops due to SLE or lupus. SLE is a type of disorder of the immune system in which a patient's immune system attacks its own cell and organ. More than 90% of the cases of SLE are of women and children. The lupus nephritis gets worse over time and leads to kidney failure.

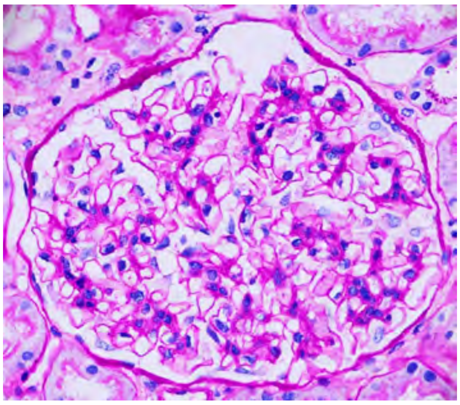

Figure 1.5: Lupus Nephritis

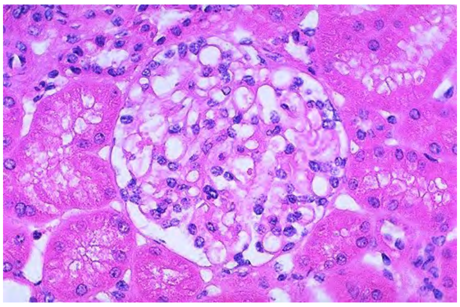

Figure 1.6: Normal Image

Polycystic Kidney Disease - is a form of CKD which gradually reduces the functionality of the kidney and leads to kidney failure. PKD can also develop from other diseases like HBP, liver cysts, narrowing of blood vessels, etc. The symptoms are similar to other renal diseases, like, blood in urine, swelled up kidney.

Kidney Cysts - is formed around and inside the kidney like a pouch full of fluid. Kidney cysts can form due to any abnormality in the kidney. These cysts rarely cause problems to the kidney and do not convert to cancer. The kidney cysts are like simply cysts formed on the body or inside the body.

1.3 Diagnosis & Treatment

As there are a lot of identified signs or symptoms of renal abnormality. Still, it is difficult to locate kidney-related diseases firsthand. But doctors initially mentioned some problems that most patients face like unexplained weight loss, nausea, malaise, vomiting, loss of appetite, pain in the backside, and high fever. In clinical manifestations, the tumor or stones grow enough and spread enough to physically obstruct the urinary flow which will make critical conditions in the ureter. To avoid progression to severe kidney damage, early identification and management of renal abnormalities are very crucial.

Diagnostic approaches include blood tests to assess renal function, and imaging studies to figure out underlying pathology. However, kidney diseases can be diagnosed through lab studies, imaging tests, and renal biopsy. In a lab test, Complete Blood Count (CBC), Kidney Function Tests, and Urinalysis are conducted. In an imaging test, Pelvic or Abdominal CT scans, MRI scans, Renal Ultrasound, KUB (Kidney, Ureter & Bladder X-ray), and radiomics are conducted by which stage tumors or stones or cancer can be classified[19]. However, deep residual learning has been presented forth in recent years to identify kidney abnormality by image identification.

Once suspicious parts are identified, a renal biopsy is performed and confirms the diagnosis. When small localized tumors or kidney stones are diagnosed, surgery takes place to remove the affected part which is known as a partial nephrectomy. When a large place is affected, a patient has to go through a radical nephrectomy in which the whole kidney may be removed or nearby lymph nodes are removed and chemotherapy and radiation therapy to terminate the cancer cells.

1.4 Statistics of Renal Abnormality

About 85 million people in the world are suffering from various types of kidney diseases. Every year 2.4 million people die of chronic kidney disease. On the other hand, about 1.3 million people suffer from sudden renal abnormality every year. Out of these, about 1.7 million patients die prematurely. Chronic kidney disease in the elderly is a hot topic these days. But there is no public awareness about this. The number of people suffering from some kind of kidney disease in Bangladesh is 38 millions. 40 to 50 thousand people suffer from kidney failure resulting continuous dialysis for surviving.

According to a report of WHO, Professor Dr. MA Samad said, by 2040, nearly 5 million kidney failure patients will die due to lack of treatment. At present, more than 850 million people are suffering from chronic kidney disease. The sad truth is that 750 million of these patients are unaware that malignant kidney disease is silently destroying their kidneys. 1.3 million people suffer from sudden kidney failure every year, 85% of them in developing countries like Bangladesh. According to various studies, the rate of chronic kidney disease among adults in Bangladesh is 16%-18%. In the United States, more than 5 hundred thousand people are affected by renal diseases.

1.5 Problem Statement

The present modern era has brought many scientific successes in the field of medicine. Now people do not have to die for lack of treatment and no one has to suffer. Although, in this scientific age, there are still some diseases, for which treatment and cure have yet to be wholly found. Only early detection and treatment of life-threatening illnesses can preserve a person's life. Renal disease, encompassing a variety of kidney conditions, is a significant health issue worldwide. Chronic kidney disease (CKD), the most prevalent form of renal disease, affects approximately 10% of the global population. This disease often progresses silently, making early detection challenging but vital for effective treatment. In 2020, kidney disease was responsible for a substantial number of health complications and fatalities globally. In the United States alone, it was anticipated that over 43 million people would be affected by CKD, with the mortality rate from this condition remaining a serious concern.

Globally, renal disease contributes to a significant number of deaths each year. For instance, the 5-year survival rate for patients diagnosed with early-stage kidney disease is significantly higher compared to those whose disease has spread to adjacent tissues or organs. While these statistics provide valuable insights into trends and outcomes, they are estimates and cannot predict individual prognoses with certainty. These figures illustrate patterns observed in the population of patients who have been diagnosed with renal disease at various stages. By leveraging advanced techniques such as deep learning and machine learning, there is potential to improve the early detection and management of renal disease, ultimately benefiting patient outcomes and quality of life.

1.6 Research Objectives

Every year more than 1500 cases are found in Bangladesh and more than 1000 death cases of renal disease. As the detection and diagnosis of kidney diseases are difficult at the early stage, they can not be detected and the death rate and new case rates are increasing. If we can detect these diseases at an early stage, then it will allow for timely treatment that can prevent severe outcomes. However, kidney disease is hard to deal with at the last stage, which can result in kidney transplantation. If renal disease can be detected much earlier, small surgeries can save someone's life. However, the current diagnostic tests in Bangladesh are costly and time-consuming, which most of the patients cannot afford. That is the reason, the death rate is high for kidney disease patients. Moreover, Every year diseases are mutating into more powerful diseases but we are lagging behind in our medical treatment and healthcare. Our technologies for medical science are not capable of detecting diseases at an early stage. Therefore, some early adjustments should be made to testing processes, spreading awareness of the value of early detection, and ease of access to medical treatments. Besides, research is concentrated on creating more effective diagnostic techniques to identify kidney abnormalities in their earliest stages.

This research focuses on improving the diagnostic framework for kidney abnormality using advanced deep-learning techniques. Specifically, we aim to develop and evaluate a convolutional Neural Network (CNN) and transformer-based architecture for classifying renal abnormality from CT scan images. Our objectives are:

1. To design and implement a CNN-based model using topologies like ResNet50, InceptionResnetV2, Inceptionv3, and VGG16 and assess their performance.

2. To propose and develop a transformer-based model for the same task and compare its performance with CNN-based models.

3. To employ data augmentation techniques and precise hyperparameter tuning, including learning rate, batch size, dropout rate, and regularization strength, to optimize models.

4. Collecting and pre-processing a normal and abnormal kidney image's hybrid dataset.

5. Conducting a complete evaluation of the models, assuring periodic assessments to meet our purposes effectively.

This research targets to create an authentic, automated diagnostic tool for medical professionals to enhance initial identification of kidney abnormalities.

Chapter 2

Literature Review

Abnormality classification has gained a lot of interest from researchers in the computer vision field as a critical challenge in digital pathology investigation. Recently, a number of CNN-based models for various abnormality of various organs, such as breast [15], lung [14], and kidney [21], have been developed. The main goal of these models is to categorize these diseases into subgroups based on histological characteristics that CNN can readily extract, such as tumor architecture.

2.1 Clinical Process

Iodine mapping and dual-energy CT were utilized by Mileto et al. [6] to help radiologists distinguish clear cell renal cell carcinoma (ccRCC) from papillary renal cell carcinoma (pRCC) visually. Using color-coded iodine maps, five readers who were blind to the pathological diagnosis independently assessed each case's lesion iodine concentration. To determine the best threshold for separating clear cells from papillary renal cell cancer, the researchers used receiver operating characteristic curve analysis. Leave-one-out cross-validation was used to confirm the correctness of the results. An intraclass correlation coefficient helped to find the inter-observer agreement. Investigators looked at the relationship between tumor iodine content and tumor grade. Tumor iodine concentrations of 0.9 mg/mL were found to be the ideal threshold for differentiating between ccRCC and pRCC, with the following findings: sensitivity of 98.2%, specificity of 86.3%, a predictive value of 95.8%, the negative predictive value of 93.7%, and overall accuracy of 95.3% with an area under the curve of 0.923. The measured tumor iodine content was found to be very well agreed upon by the five readers (intraclass correlation coefficient, 0.9990). A strong relationship between tumor iodine content and tumor grade was discovered for both clear cell and malignant tumors. The high-throughput extraction of many picture characteristics from radiographic images is known as radiomics. It has been suggested to use radiomics to extract quantitative information from radiographic pictures and create models that link image aspects to outcomes [2]. As solid cancers are diverse in both time and space. This restricts the use of expensive biopsy-based molecular tests but opens up a world of possibilities for medical imaging that can non-invasively capture intra-tumoral heterogeneity. Some radiomics models have been put out in recent years to categorize kidney abnormality.

2.2 Machine Learning Algorithms

In an earlier study on the classification of renal cancer subtypes, Kocaka et al. [17] used SVM and ANN to classify the subtypes of renal cancer. Three radiologists performed a reproducibility study as the initial step in feature selection, followed by a wrapper-based classifier-specific method. The model was optimized and features were chosen using layered cross-validation. Artificial Neural Networks (ANN) and Support Vector Machines were the main classifiers (SVM). To enhance generalizability performance, base classifiers were additionally merged with three additional methods. The following categories were used for classification: (i) clear cell RCC (cc-RCC) vs papillary cell RCC (pc-RCC) versus chromophobe cell RCC; and (ii) non-clear cell RCC (non-cc-RCC) versus clear cell RCC (cc-RCC) (chc-RCC). The Matthews correlation coefficient served as the primary performance parameter for comparisons (MCC). The best method for differentiating non-cc-RCCs from cc-RCCs using corticomedullary phase pictures was an ANN with an adaptive boosting algorithm (MCC = 0.728), which had external validation accuracy, sensitivity, and specificity of 84.6%, 69.2%, and 100%, respectively. However, the effectiveness of QCT-TA is pretty subpar when it comes to differentiating the three main subtypes. For differentiating pc-RCC from other RCC subtypes, the SVM with bagging algorithm performed best (MCC = 0.804), with accuracy, sensitivity, and specificity for external validation of 69.2%, 71.4%, and 100%, respectively [25].

Texture analysis was utilized by Hodgdon et al. [9] to distinguish angiomyolipoma (AMLs) from renal cell cancer (RCC). The DeLong approach helped to differentiate a receiver operating characteristic curve's regions under(AUC) between subjective heterogeneity evaluations and textural attributes. An AUC of 0.89 was obtained using a model that used many texture characteristics. SVM accuracy for textural characteristics varied from 83% to 91% on average (10-fold cross-validation).
A random forest was used by Raman et al. [7] to forecast the pathophysiology of kidney malignancies. External validation of the model was performed on a different group of 19 unidentified cases. Oncocytomas and clear cell RCCs were properly classified by the random forest model in 89% (sensitivity = 89%, specificity = 99%) and 91% (sensitivity = 91%, specificity = 97%) of the cases, respectively.
To distinguish between several kinds of tiny renal tumors, Feng et al. [16] used quantitative texture analysis based on machine learning on CT images. However, the characteristics used in these investigations, such as their form, intensity, texture, and wavelet textures, were specifically developed or made by hand [3]. In preoperative three-phase CT scans, texture characteristics were manually segmented from the biggest tumorous areas of interest (ROIs). A preliminary selection of characteristics was made using the Mann-Whitney U test and inter-observer reliability. Then, using support vector machines with recursive feature elimination (SVM-RFE) and the synthetic minority oversampling method (SMOTE), discriminative classifiers were created, and their performance was assessed. The SVM-RFE+SMOTE classifier showed the best performance by differentiating between microscopic angiomyolipoma without visible fat (AMLwvf) and RCC with the highest accuracy, sensitivity, specificity, and AUC of 93.9%, 87.8%, 100%, and 0.955, respectively. The potential of radiomics may be limited as a result of the selection of these low-throughput traits based on the professional knowledge of radiologists.

In the classification of kidney diseases, a variety of deep-learning techniques are used. In the paper [12], They begin by using the median filter, Gaussian filter, and un-sharp masking to improve the image. To analyze kidney stone images, they first employed morphological procedures including erosion and dilation. Then, to determine the region of interest, they used entropy-based segmentation. For both the original picture and the segmented image, they computed many metrics, including the standard deviation, entropy, thresholding, energy, and homogeneity. Finally, they used KNN and SVM classification approaches. Subsequently, The K-nearest neighbor (KNN) classifier and principal component analysis (PCA) are used to extract information from the pictures. They proposed two types of classification here. KNN was found to be 89% accurate, whereas SVM was shown to be 84% accurate.

2.3 Convolutional Neural Networks

Convolutional neural networks (CNNs) have recently made significant advances in computer vision capabilities as a result of the introduction of graphics processing units and massive training datasets [5]. A CNN may automatically extract high-throughput features and forgo the laborious artificial feature extraction method when a sizable training dataset is provided [4]. CNN has performed admirably in the medical domains.

By using 2000 dermatological photos and the associated pathological findings to train a CNN model, Esteva et al. [11] were able to differentiate between benign and malignant skin malignancies utilizing only the inputs of pixels and disease labels. By contrasting the results of the procedure with those reached by 21 board-certified dermatologists, the method's effectiveness was evaluated. The dermatologists concentrated on two crucial categorization tasks: identifying keratino cyte carcinomas from benign seborrheic keratoses and separating malignant melanomas from benign cell. They combined biopsy data with photos that had been clinically verified. The outcomes showed that some convolutional neural networks (CNNs) are capable of classifying skin cancer with a degree of accuracy comparable to dermatologists. On each task, the CNN fared just as well as the tested pros.

Additionally, Arevalo et al. [8] used a CNN to categorize mammography mass lesions. They employed a hybrid strategy in which the representation was supervised and learned using CNNs. In other words, they directed the feature learning process using the annotations and had outstanding results with values ranging from 79.9% to 86.0% as measured by the ROC and AUC curve. Feature learning for mammography mass lesions using convolutional neural networks is evaluated here before being fed to a classification step. It has not yet been thoroughly investigated if such a method may help distinguish properly between benign and malignant kidney abnormality based on CT scans.

This study [23] uses convolutional neural networks and other machine learning techniques to categorize individuals as either healthy or patients based on the presence or absence of kidney stones in medical photographs (CNN). The automated categorization of B-mode renal ultrasound pictures is suggested in this study [22] and is based on a group of deep neural networks (DNNs) that use transfer learning. The quality selection in ultrasound pictures is based on the perception-based image quality assessor score, and speckle noise often affects the images. The support vector machine is used for classification after the pre-trained DNN models extract features from three different datasets. The majority voting method is used with multiple pre-trained DNNs, including ResNet-101, ShuffleNet, and MobileNet-v2, to produce final recommendations. By combining the predictions from several DNNs, the ensemble model outperforms the individual models in classification[26]. When contrasted with traditional and DNN-based categorization approaches, the given method clearly demonstrated its advantages. The established ensemble model divides the normal, cyst, stone, and tumor classes for the kidney ultrasound pictures. The authors achieved the maximum accuracy of 95.58% using ultrasound pictures there for the categorization challenge.

In order to reliably and quantitatively detect chronic kidney illnesses, this research [24] concentrated on using deep learning techniques for the division of CT images. First, renal cysts in CT images were automatically segmented using the residual dual-attention module (RDA module). As research participants, 79 individuals with renal cysts were chosen, of whom 52 instances served as the training group and 27 cases served as the test group. The segmentation results for the test group were evaluated using the Dice similarity coefficient, recall, and precision (DSC). The experimental results demonstrated that the RDA-UNET model's loss function value quickly converged and declined, Additionally, the segmentation outcomes of the study's model were nearly identical to those of hand labeling, confirming the model's high level of picture segmentation accuracy as well as its capacity to precisely segment the kidney's shape. By obtaining 96.25% DSC, 96.34% precision, and 96.88% recall for the left kidney and 94.22% DSC, 95.34% precision, and 94.61% recall for the right kidney, the RDA-UNET model beat earlier methods. The results showed that the algorithm model utilized in this study outscored other algorithms in each assessment index. The authors in [18] suggested a lesion identification algorithm based on morphological cascaded convolutional neural networks that use multiple intersections over union (IOU) thresholds (CNNs). To improve the detection of small lesions (1–5 mm) and boost network stability, we proposed two morphological convolution layers, updated feature pyramid networks (FPNs), and four IOU threshold cascade RCNNs. PyTorch was used to train the modified CNN for this lesion detection task. The research was done using DeepLesion kidney CT pictures supplied by hospital picture archiving and sharing systems (PACSs). The findings showed that our suggested detector is an excellent tool for detecting lesions in CT and outperformed the dataset, with our technique achieving an AP of 0.840 and AUC of 0.871.

In [13], the scientists built a fully automated system for detecting renal cysts that are backed by reliable kidney segmentation done by a fully convolutional neural network. Initial candidates for cysts are provided by an integrated 3D fluid and kidney distance map around them. The final step is to classify the candidate's status as cysts or non-cyst objects as a second convolutional neural network. 52 abdomen CT images with more than 70 cysts that were randomly picked from a genuine radiological workflow and annotated by a skilled radiologist were used to evaluate performance. at the time When the minimum cyst diameter was set to 10 mm, the system detected 59/70 cysts with a true-positive rate of 84.3% and an average of 1.6 false positives per case.

2.4 Transformer Based Approach

In the study [27] , Nazmul et al. , aimed to develop an AI-based diagnostic system for kidney diseases, focusing on kidney stones, cysts, and tumors. Using a dataset of 12,446 CT images, six machine learning models were evaluated: three based on Vision transformer (EANet, CCT, and Swin transformers) and three on deep learning models (ResNet, VGG16, and Inception v3). Among these, the Swin transformer achieved the highest accuracy of 99.30%, outperforming all other models in terms of F1 score, precision, and recall. Additionally, it was the quickest to train. VGG16 also showed superior performance compared to ResNet50 and Inceptionv3 in monitoring anatomical abnormalities.

Yang et al. [29], in his paper proposed a Transformer-based learning algorithm to upgrad the diagnostic accuracy of grading clear cell renal cell carcinoma (ccRCC) using CT images. Experiments were conducted on a dataset of 759 patients, and the model's performance was evaluated using average classification accuracy, sensitivity, specificity, and Area Under Curve (AUC). The transformer based model outperformed traditional CNNs, achieving a mean accuracy of 87.1%, sensitivity of 91.35, specificity of 85.3%, and an AUC of 90.3%. The integrated model, which combined different training models, showed further improvement with an accuracy of 86.5% and an AUC of 91.2%. The result indicate that the Transformer-based network is more effective than traditional deep learning algorithms for ccRCC grading, and it demonstrates robustness in handling noise in CT images, suggesting potential applications in other tasks.

Kushol ei al., in his study [28] introduces a Vision Transformer based approach which will automatically detect Alzheimer's disease (AD) from healthy controls using MRI data. The model leverages both frequency and image domain features, incorporating coronal 2D slices pre-trained on ImageNet, and applied majority voting for final classification. Evaluated on the ADNI dataset, the proposed method shows superior performance in compare to state-of-the-art techniques. Specifically, it achieves an accuracy (ACC) of 0.882, sensitivity (SEN) of 0.956, and specificity (SPE) of 0.774, outperforming popular CNN-based models and other advanced methods in terms of ACC and SEN while ranking second in SPE. The robustness of the approach is further validated by attention maps highlighting crucial regions near the hippocampus. An ablation study confirms the efficacy of fusing frequency and image domain features, showing significant accuracy improvement with the proposed architecture.

From the above studies we can see, CNN and Transformer-based approaches have outperformed in classifying diseases in medical images, surpassing traditional ML models. In Esteva et al.'s [11] work CNN has proven efficacy in identifying skin and kidney diseases. Transformers, like those in Nazmul et al.'s [27] study showed 99.30% accuracy in renal diseases. Despite their triumph, these studies often focus on specific abnormality or imaging types, leaving gaps. Our research focuses to build a comprehensive diagnostic system utilizing CNN and transformer strengths to address these limitations.

Citations	**Used Models**	**Dataset Details**	**Best Performing Model**
[15] - [21]	CNN Based & Transformer Based Models	Various Organs(Breast, Lung, Kidney)	-
[6]	ROC Curve Analysis	Iodine Mapping & Dual Energy CT for ccRCC vs. pRCC	-
[17]	SVM, ANN with Adaptive Boosting	Radiologists' reproducibility work & classifier specific wrapper technique	ANN with Adaptive Boosting (MCC 0.728)
[9]	Texture Analysis, SVM	Texture features for differentiating AML from RCC	SVM (Accuracy: 83% - 91%)
[7]	Random Forest	Pathophysiology prediction of kidney malignancies	Random Forest (Accuracy: 91%)
[16]	SVM-RFE, SMOTE	Quantitative texture analysis on CT scan images	SVM-RFE with SMOTE (Accuracy: 93.9%)
[3]	Machine Learning Algorithms	Various studies	-
[12]	KNN, SVM	Gaussian Filters, Median, un-sharp Masking, Morphological techniques	KNN (Accuracy: 89%)
[5] - [4]	CNN	Medical Domains	-
[11]	CNN	2000 Dermatological Images	CNN(Performance as good as Dermatologists)

[8]	CNN, Hybrid Approach	Mammography Mass Lesions	Hybrid CNN (ROC and AUC: 79.9% - 86%)
[23] - [22]	CNN, Ensemble Neural Network(DNNs)	B-mode Kidney Ultrasound Images	Ensemble DNNs (Accuracy: 95.58%)
[24]	Deep Learning(RDA-UNET)	CT scan Images for Chronic Kidney Diseases	RDA-UNET (DSC: 96.25%, 94.22%)
[18]	Morphological Cascaded CNNs	Deep-Lesion Renal CT scan Images	CNN (AP: 0.84, AUC: 87.1%)
[18]	Fully Automated System with CNN	52 Abdomen CT scan Images with 70 Cysts	CNN (True-positive Rate: 84.3%)
[27]	Vision Transformer, ResNet, VGG16, InceptionV3	12,446 CT scan Images	Vision Transformer (Accuracy: 99.30%)
[29]	Transformer	759 Patients with CT scan Images for Renal Cell Carcinoma(ccRCC)	Transformer (Accuracy: 87.1%)
[28]	Vision Transformer	MRI Data from ADNI Dataset for Alzheimer's Disease	Vision Transformer (Accuracy: 88.2%)

Table 2.1: Summarizing of Literature Review

Chapter 3

Background Studies

3.1 Convolutional Neural Network

Convolutional Neural Networks (CNN) are a class of deep learning models particularly effective for analyzing visual data. They excel at identifying patterns in images, such as objects and categories, and can also be effective for classifying audio, time-series, and signal data.

CNN consists of an input layer, an output layer, and multiple layers. Each hidden layer detects different features of an images by applying filters. Initially, these may identify simple features like edges, but as layers progress, they capture increasingly complex patterns unique to the object. Convolutional layers apply filters to input images to active certain features. Activation (ReLu) layers speed up and improve training by mapping negative values to zero and retaining positive values. Pooling layers reduce the dimensionality of feature maps through nonlinear down-sampling, simplifying the output and reducing the number of learnable parameters.

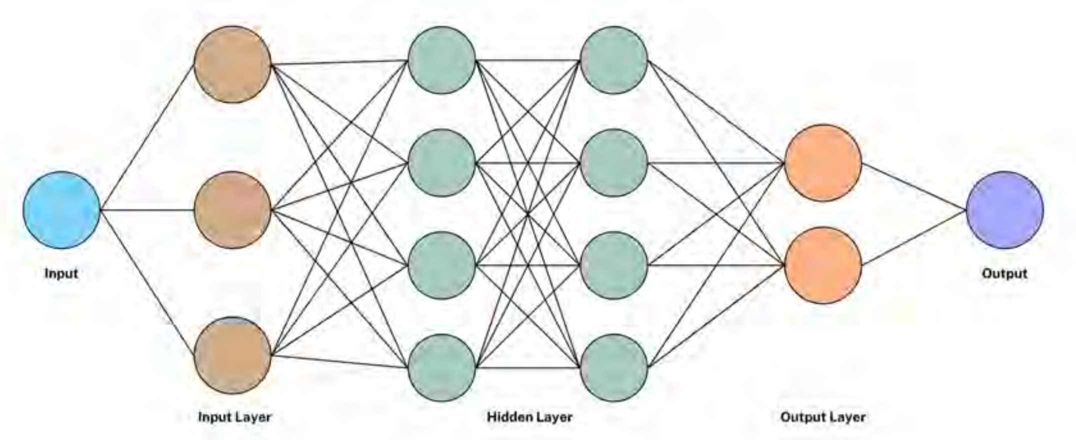

Figure 3.1: Convolutional Neural Network

This component repeated across many layers, each learning to identify various features. CNN use shared weights and biases across all neurons in a layer, meaning each neuron detects the same feature in different regions of the image.

3.1.1 ResNet50

ResNet50 is a variant of the Resnet model. ResNet, particularly the ResNet50, addresses the vanishing gradient problem in deep learning neural networks, which occurs when gradients become extremely small during back propagation, hindering the training of earlier layers. The problem is mitigated using residual networks, which employs skip connection. This allows the network to bypass certain layers. Skip connection, represented by

$$Y = F(X) + X$$

allows the network to learn residual function F(X) that approach zero, making Y approximately equal to X.

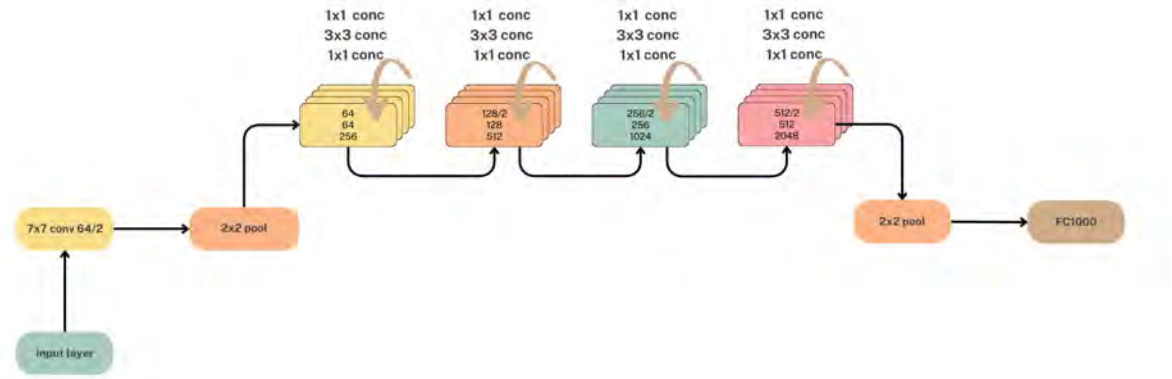

Figure 3.2: ResNet50 Architecture

In Residual block the calculation follows:

$$Y = F(X) + X$$

where F(x) is the convolutional operations. ResNet50 architecture consists of 50 layers, structured as follows: 1 + 9 + 12 + 18 + 9 +1 = 50 layers. It includes 48 convolutional layers, along with 1 MaxPooling and 1 average pooling layer. In First layer: Filter size: 7 × 7, Filters =64, Stride= 2, Padding= 3. Output Size Calculation:

$$(\frac{n + 2p - f}{8}) + 1 = (\frac{300 + 23 - 7}{2}) + 1 = 150 \times 150 \times 64$$

With MaxPooling 3×3 stride 2 and padding 1 the image size reduces to 75 × 75.

3.1.2 InceptionResNetV2

InceptionResNetV2 is a deep convolutional neural network architecture that combines the strengths of Inception networks and Residual networks (ResNets). This hybrid model integrates the efficient filter concatenation approach of inception modules with the identity mapping of residuals connections. InceptionResNetV2 architecture consists of stem module, InceptionResNet modules with continues from A to

C which are the core building blocks of the network, Reduction modules and Final layers. Stem process the initial input image. It involves a series of convolutional and pooling layers.The final layers often include average pooling, dropout, and a fully connected layer to produce the final classification.

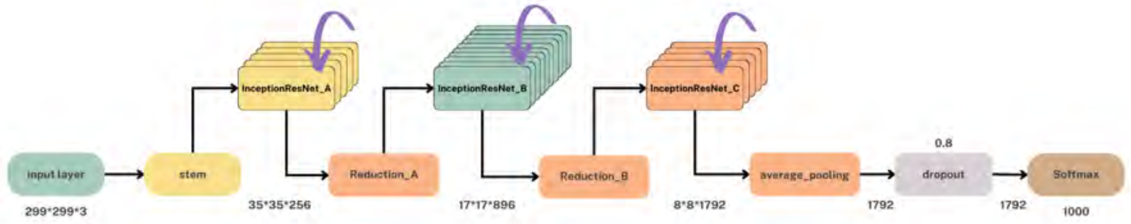

Figure 3.3: InceptionResNetV2 Architecture

The essence of the residual connection can be described by the following equation:

$$y = F(x, W_i) + x$$

Here, x is the input to the residual block, y is the output of it. An Inception module's output is a concatenation of several convolutional operations:

$$y = concat(Conv_1, Conv_2, ..., Conv_n)$$

InceptionResNetV2 leverages the multi-scale processing with the identity mappings of residual connections.

3.1.3 InceptionV3

InceptionV3 is a deep convolutional neural networks designed for image classification tasks. It is a successor to the InceptionV1 and InceptionV2 models, which was developed by Google.

Inception usually improves upon its predecessors by optimizing cost without compromising performance and making it suitable for practical implementation. The architecture of InceptionV3 is built on some concept: Factorization into smaller convolutions, asymmetric convolutions, auxiliary classifiers, reduction of grid size and batch normalization. The convolutional operation equation:

$$Y[i, j, k] = \sum_{m,n,l} X[i + m, j + n, l] \times W[m, n, l, k]$$

Here Y is the output feature map, X is the input, W is the convolution filter , and i,j,k,m,n,l are the spatial channel indices.

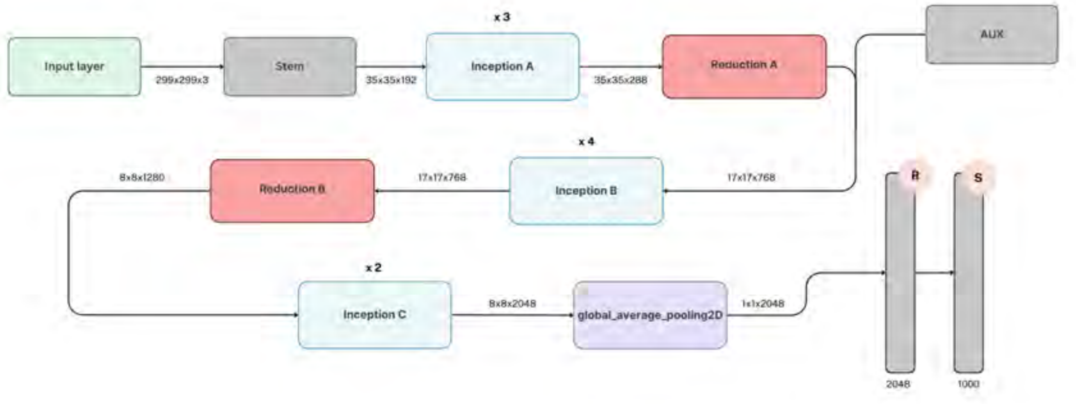

Figure 3.4: InceptionV3 Architecture

3.1.4 VGG16

The VGG16 model is a deep convolutional neural network (CNN) known for its simplicity and effectiveness. It's made of 16 layers, including 13 convolutional layers and 3 fully connected layers, which makes it capable of learning complex topological representations for image classification. The architecture features sequential stacks of convolutional layers followed by max-pooling layers, with increasing depth.

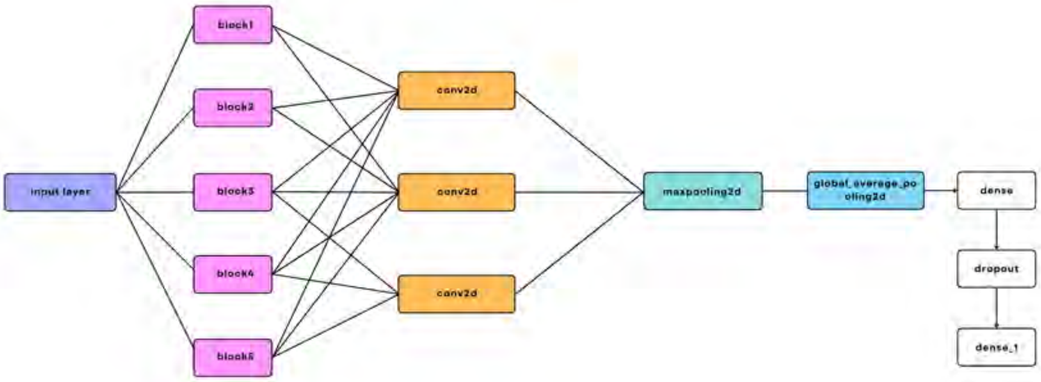

Figure 3.5: VGG16 Architecture

In the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), VGG16 got remarkable performance, ensuring a top-5 classification accuracy of 92.7% and showing best result in both classification and localization tasks. The VGG16 model process input image size of 224x224x3 and give a probability vector of 1000 class as output.The softmax function that is used to calculate the output vector:

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$$

Where $\hat{y}_i$ is the predicted probability for class i, and z_i is input score of i class. VGG16 adds two or three convolutional layers per block and each layer is followed by max-pooling. With fully connected final result leads to softmax output. Despite of simpleness, it is widely used for its robustness.

3.2 Transformer Based Models

Transformer models take input data through layers and process with self-attention mechanisms and feedforward neural networks. Firstly input embeddings and positional encoding to capture positional and semantic information. The given data is passed through multi head attention and feedforward layers, along with normalization of layers. After refining the multiple layer, an output layer is generated.

3.2.1 Vision Transformer (ViT)

The Vision Transformer is first produced in the paper [20] ”An Image is Worth 16x16 words” by Dosovitskiy et al., shows that an transformer model can be well trained on ImageNet for classifying without depending on convolutional neural networks. In the study, it is demonstrated that the model with achieve excellent result if a pure Transformer is applied directly to sequences of image patches.

3.3 Evaluation Metrics

For assessing the performance of any model evaluation matrices are very important [10]. They give insights on how well the model is performing. It makes easier to understand it effectiveness.

3.3.1 Accuracy

In the test set accuracy calculates the ratio of predicted instance with the total number of instances. We can get overall models predicted performance from the accuracy.

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Total\ number\ of\ prediction}$$

3.3.2 Precision

The ratio of correctly predicted positive instances to the number of predicted instances is calculated by precision. The accuracy of the positive predictions made by the model can be seen from precision.

$$Precision = \frac{True\ Positives}{True\ positive\ +\ False\ positives}$$

3.3.3 Recall

Recall, also known as sensitivity, calculates the ratio of correctly predicted positive instance with the total number of actual positives. Its shows that if the model could identify all the relevant instances.

$$Recall = \frac{True\ Positives}{True\ Positive\ +\ False\ Negatives}$$

3.3.4 F1 Score

The harmonic mean of precision and recall is F1 score. F1 score give a single matrix which balance recall and precision. It is specially useful in datasets that are not balanced.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Models overall performance is showed by these matrices collectively. Which highlights the area of improvement.

3.4 Description Of The Data

The Hybrid dataset we have utilized in this paper is composed of a collection of Abnormal CT scan renal images which is a Novel dataset prepared with the help of the Kidney Foundation Hospital & Research Institute and Normal CT scan images from Kaggle. The dataset is built specifically to facilitate the detection of kidney abnormalities using CNN and transformer-based models. Our dataset comprises two distinct classes: normal kidney images and abnormal kidney images.

3.4.1 Abnormal Dataset

We collected abnormal kidney images from the Kidney Foundation Hospital & Research Institute, Dhaka, Bangladesh. The abnormal images were 3D CT scan files of the whole abdomen and lower abdomen later processed in the pre-processing part. The images were produced during our affiliation with the Kidney Foundation Hospital & Research Institute from June 2023 to January 2024. The images were produced and collected following the 1946 Helsinki Declaration ethical standards, hence human participants were informed consent from all the patients was obtained during the whole collection procedure, and personal details of each individual were secured and protected. These image data encompass a diverse range of kidney images collected from patients representing various kidney abnormalities. The data were collected with informed consent and ethical approval in mind. Patient privacy was strictly maintained, and the whole process was done under the supervision of regulatory compliance. This Abnormal class of the dataset is a fully Novel dataset.

3.4.2 Normal Dataset

The images belonging to the Normal class utilized in our hybrid dataset were sourced, especially from the “Kidney Cancer Image” dataset on the popular machine learning dataset website Kaggle. This publicly available dataset on Kaggle contains CT scan images of normal kidney. These images were collected because they provide a comprehensive representation of healthy renal kidney CT scan images, the healthy renal anatomy represented by these images serves as a crucial reference point for distinguishing normal kidney images from pathological findings of abnormal kidney images. Moreover, this class of the dataset represents a diverse variety of normal kidney CT scan images taken under several different conditions, this ensures a diverse and robust set of examples that can be used to train, validate, and test machine learning classification models.

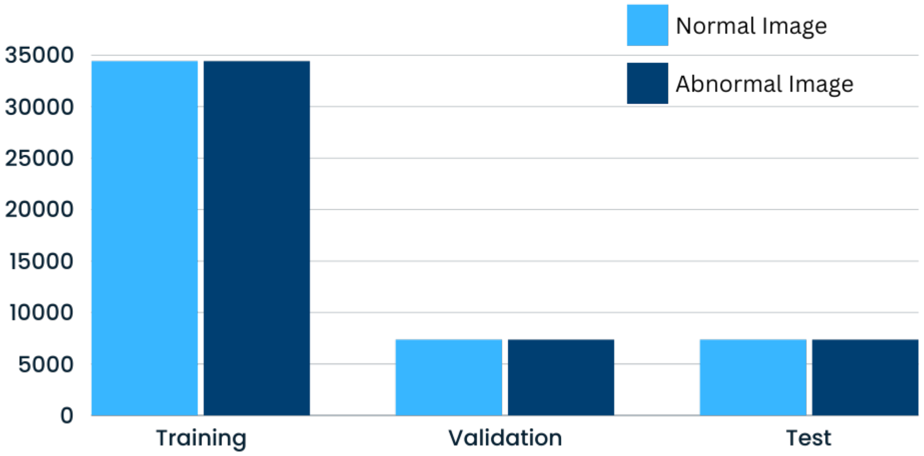

Figure 3.6: Graph of Dataset

Type	**Training Patch**	**Test Patch**	**Validation Patch**
Label 0	34444	7381	7381
Label 1	34444	7381	7381
Total	68888	14762	14762

Table 3.1: Proposed Dataset Label Distribution

Chapter 4

Proposed Methodology

4.1 Methodology

Many pre-trained and custom have been used for classifying the binary labeled image. Among them, vision transformers (ViT) are a type of neural network architecture that is a transformer-based neural network and were developed by Google back in 2017 and were mostly applied for NLP. However, in this present era, it has been proven to be quite effective in computer vision applications, especially in object detection and image segmentation. To use a vision transformer for renal abnormality image classification we need to follow some necessary steps. These steps or techniques are crucial to train the model well and ensure the expected outcome from our model. ResNet152 is used to make the training of deep networks significantly easier than usual neural networks. These residual networks have shown that they are not that difficult to enhance. On the other hand, ResNet50 is also a deep neural residual network that is known for its depth and efficiency in image classification tasks. This residual network works on 50 layers deep and trained on large datasets and achieves the desired outputs. Inception V3 is a convolutional neural network that makes several improvements to label information down the network. This model has high efficiency and a deeper network compared to other models. VGG16 is usually used for object detection and classification algorithms. This neural network works on 16 layers that have weights and huge hyper parameters. In a comparison to VGG16, both are the advanced versions of each model. However, the accuracy and better working between them depends on the dataset on what both models are working on. InceptionResNetV2 is a convolutional neural architecture that works on residual connections to improve its performance. It has maintained its effectiveness in image processing and pattern recognition along with low-cost maintenance.

4.2 Data Pre-processing

4.2.1 Pre-processing

During the data pre-processing phase, a series of crucial procedures were performed to ensure that the kidney computed tomography(CT) scan images were appropriately used to build the Hybrid Dataset that would represent real-world data.

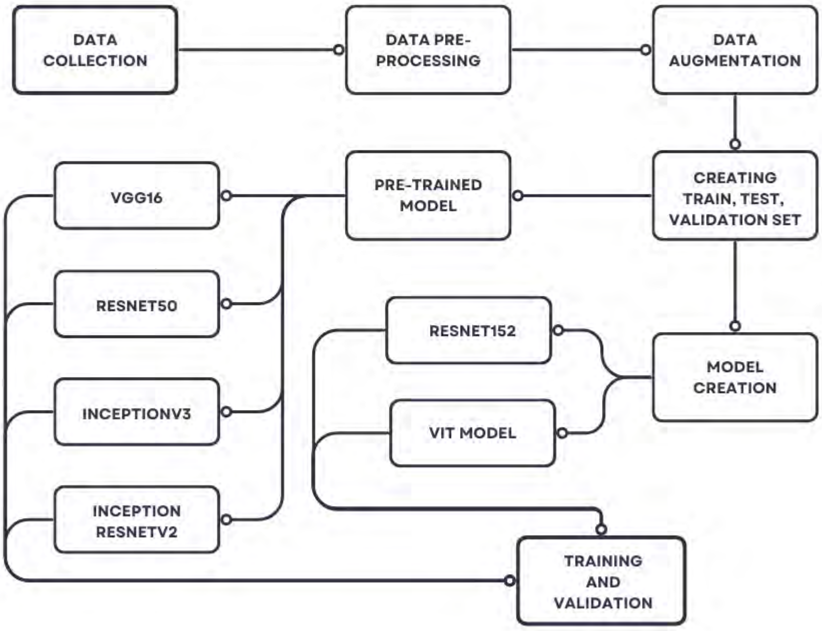

Figure 4.1: Work Plan

Data Collection

The used image dataset is a binary class dataset containing normal images of the kidney and abnormal images of the kidney. The dataset of abnormal images is collected from the Kidney Foundation Hospital & Research Institute, Dhaka, Bangladesh. The abnormal images were 3D CT scan files of the whole abdomen and lower abdomen later processed in the pre-processing part. All the images were collected from the patients of Kidney Foundation Hospital & Research Institute. The images were produced during our affiliation with Kidney Foundation Hospital and Research Institute from 15 June 2023 to January 2024. The images were produced by following the 1964 Helsinki Declaration ethical standards, hence human participants were informed consent from all the patients was obtained during the period of the whole experiment and the personal details of each individual were secured and protected.

Data Conversion

Firstly, the DICOM files were converted to JPG format as our models used pixel-to-figure patterns of images, the converted JPG images were later used to build the Hybrid Dataset. The following Python libraries were necessary to translate the DICOM files of 3D CT Scan image to usable JPG image files: pydicom, PIL, glob, and OS. MicroDicom software was also used to explore and study the DICOM files and convert them to JPG image files.

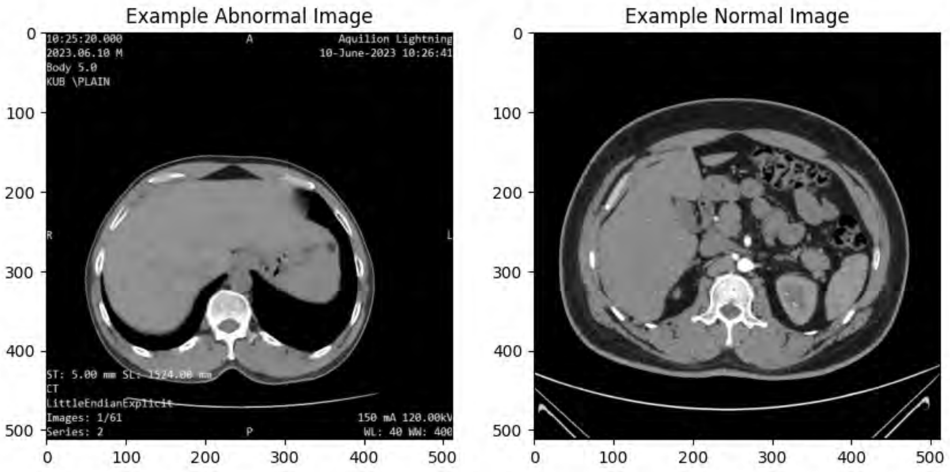

Figure 4.2: Normal Abnormal Image of Dataset

Selection of Kidney Images

The upper process of converting the DICOM field to JPG files produced many unnecessary images created during the CT scan process. To ensure that our hybrid dataset contained only renal images, we had to identify and select only renal images out of all the images produced in the previous process. Medical professionals were present to supervise our renal image selection.

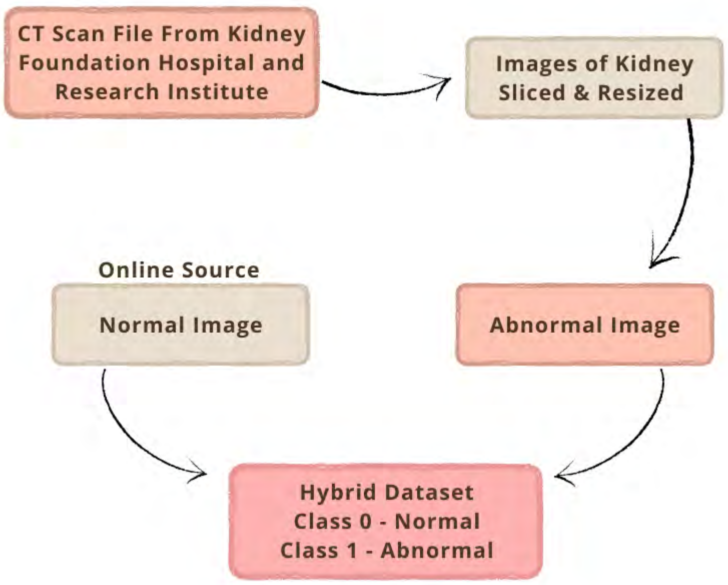
CT Scan File From Kidney Foundation Hospital and Research Institute
Images of Kidney Sliced & Resized
Online Source
Normal Image
Abnormal Image
Hybrid Dataset
Class 0 - Normal
Class 1 - Abnormal

Figure 4.3: Data Collection

4.2.2 Data Augmentation

In the field of medical image data, more precisely in the field of renal CT scan data, machine learning, and deep learning have significantly improved diagnosis and treatment frameworks. However, the capabilities of these machine learning models heavily rely on the quality and volume of the dataset used to train and evaluate them. Specific data augmentation methods are necessary to optimize the dataset and fine-tune the machine learning models to ensure that the models can deliver high performance and reliability while handling real-world CT scan images.

Limited data availability

The hybrid dataset used in this paper consists of two classes: the abnormal class and the normal class. The abnormal class of this dataset is novel, as it was built using a novel dataset approach by collecting data from Kidney Foundation Hospital & Research Institute, The normal class was built using data collected from a publicly available online dataset from Kaggle. The number of data collected from two different sources was not precisely the same. Hence, it was necessary to go through data augmentation to build an acceptable Hybrid dataset for our research work.

Improved model performance

Overfitting of machine learning classifier models is reduced by exposing the model to a variety of image data, data augmentation prevents the models from learning unnecessary patterns. Data augmentation is also necessary to stimulate the variations of the dataset so that robust real-world capable models can be built using the datasets, as Medical images like renal CT scan images can exhibit considerable variability due to the variations in patient anatomies, CT scanning procedure method and the noise produced while producing CT scan images.

Data augmentation method

We have used the following data augmentation technique to improve the variety and reliability of our renal CT scan hybrid dataset so that the dataset is representative of real-world data.

- **Rotation:** To simulate various orientations of renal images during the CT scanning process, a rotation of 10 degrees was used. Wide shift range: The renal images were randomly shifted horizontally by using a width shift range of 0.1, this process randomly shifted horizontally by 10% of the width of the image. The wide shifting was used to replicate misalignment that can occur during the CT scan process.

- **Height Shift Range:** The renal images were randomly shifted vertically using a wide shift range of 0.1, which means they were randomly displaced vertically by 10% of their vertical height. Again, just like before, this was done to replicate the misalignment that can occur during.

- **Shear Range:** The shear range of 0.1 was applied to repeat shear transformations that usually occur due to image angles. Zoom range: We decided to use a zoom range of 0.1, which randomly magnifies and shrinks images to about 10% of their original sizes. This was so that classification models could recognize kidneys at various scales.

- **Horizontal Flip:** We decided to use Horizontal flip to augment the hybrid dataset, which created horizontally flipped mirrored images, increasing the number of images available.
- **Vertical Flip:** We also decided to use vertical flip techniques to augment the hybrid dataset this ensured the number of images was increased using vertically flip images.

Parameter	**Changed Values**
Rotation	10%
Height Shift Range	10%
Shear Range	10%
Flip	Horizontal & Vertical

Table 4.1: Data Augmentation

4.2.3 Data Segmentation

After data augmentation was done we divided our dataset into three segments before using our hybrid dataset to build our model.

Training Set

The training segment contains 70% of the whole dataset, which results in 34,444 images being in the abnormal class and 34,444 images being in the normal class of our hybrid dataset. We decided to split our dataset so that the models are trained on a large number of image files and can pick up a significant pattern of the images.

Validation Set

The validation set contains 15% of the entire dataset, which resulted in 7,381 images in the abnormal class and 7,381 in the normal class. The validation set was created to ensure we can fine-tune our model and mitigate overfitting by accessing the performance of our proposed models and pre-trained model on previously unseen data while running the training phase.

Testing Set

Finally, we decided to include 15% of all the images available in the testing segment, this resulted in 7,381 images being in the abnormal image class and 7,381 on the normal image class. The resting dataset consists of the remaining available renal images to evaluate the model‘s performance.

Ensuring the model‘s ability to cope up with real-world data and ensuring the capability of working with unseen renal CT scan images were our key objectives. This method of dataset separation and augmentation technique helped our models to become more efficient and reliable.

4.3 Model Creation

We worked on some pre-trained models of CNN and made some custom proposed model of CNN and Transformer based architecture.

4.3.1 VGG16

The VGG16 pre-trained model was used as a feature extractor within the custom model architecture. Specially, the VGG16 model of ours is loaded with pre-trained weights of ImageNet and set to separate the top layer of classification. Which means that instead of utilizing VGG16 for its main task of image classification, it works as a strong feature extractor, capturing topological representation of the inputs. After extracting features and feeding it into additional layers, including globalaveragepooling, dense and dropout layers. These layers are responsible for learning how to classify images based on the extracted features done by VGG16. Some layers of VGG16 was unfrozen and fine-tuned for better suit. Which ultimately upgraded its performance.

4.3.2 ResNet50

In our model, ResNet50 architecture pre-trained was used on ImageNet to support its strong feature extraction capabilities. We got rid of the top layer and added some custom layers, including globalaveragepooling, a dense layer with an ReLU activation and dropout layer which helped to avert overfitting. Also a final dense layer with sigmoid function for the binary classification. This setup helped us to easily classify the kidney images from normal to abnormal. We used Adam optimizer and binary cross entropy loss. We also used some callbacks like early stopping and model checkpointing to make sure that we do optimal training.

4.3.3 InceptionV3

To create and implement the InceptionV3 model on our hybrid dataset, at first we had to import necessary Python libraries like numpy, pandas, train_test split, and TensorFlow. Keras, we had to import ImageDataGenerator. On top of that, we had to import preprocess_input from Tensorfow. Keras. Applications. Resnet. So that crucial operations like numerical computations, data manipulation, dataset splitting, and image pre-processing can be implemented. After that, we set the dataset paths and labeled the classes; labeled “1” represented our abnormal class, and labeled “0” represented our normal class. As discussed in the data pre-processing part, we split the dataset into training, validation, and testing. The first 70% of the dataset was used for training, 15% for validation, and the rest for testing. Necessary data augmentation was previously done to ensure the dataset is balanced between the two classes, using ImageDataGenerator. Data was generated from training, validation, and testing, these generators were used to convert the renal images and labels in the form of data frames into suitable batch sizes for work, the target size was 224 by 244, while we used a batch size of 32. Resnet50 was used as the base model, “ImageNet” weight was used, and input_shape was 244X244X3, the dropout rate was 0.5, and dense was 512 we used sigmoid as activation since we are doing binary classification. We implemented a learning rate of 0.001 binary_crossentropy, which was used as loss, and we used accuracy as matrices. Finally, we trained our model using necessary parameters like early stopping, and we used 30 epochs to collect the essential metrics that we used for our evaluation. We also added some final layers and used custom callbacks to prevent from overfitting.

4.3.4 InceptionResNetV2

In our model we used the InceptionResnetV2 architecture like the backbone of the feature extraction because of its powerful performance in maintaining complex image separation tasks. Combining the strength of Inception modules and residual connections, it was enable to learn complex patterns in data but it maintained computational productivity. Initializing the model with pretrained weights from ImageNet, providing a strong initializing point and helped in faster convergence. The mentioned backbone was then fine tuned on our dataset of normal and abnormal kidney images.

4.3.5 Proposed CNN Model

This section outlines an overview of the structure of our custom Convolutional Neural Network (CNN) which is specifically purposed for classifying renal images of our hybrid novel dataset that consists of two classes “abnormal” and “normal”, hence our proposed Convolutional Neural Network (CNN) is designed to handle binary classification efficiently. Our proposed custom model was built using ResNet152 as its base layer and we have utilized 128 layers from this base layer, additionally, we made further adjustments to better fit this model for our specific binary classification task.

Architecture Overview

Our proposed custom CNN architecture integrates powerful feature extraction capabilities. We implemented the ResNet152 model as our base model on our ImageNet, eliminating the fully connected layers. We loaded our proposed custom CNN with local weights, the input shape for the images was 224×224×3. Initially, All the layers of the base ResNet152 were frozen to prevent their use during the training phase, as we had added our custom weight for our purpose.

We have added a globalaveragepooling Layer right after the base ResNet152 model to reduce the spatial dimensions of our feature maps. We also applied a Batch Normalization Layer to the outputs of the Global Average Pooling layer, a dropout rate of 0.5 was added which prevented model overfitting. We decided to use 128 units of the fully connected dense layer with L2 Regularization with a factor of 0.01 from the ResNet152 base models, this step was necessary to penalise heavier with which as a result would reduce overfitting. After the dense layer three more layers were added they were the Batch Normalization layer, the Dropout layer with a rate of 0.5, and finally, the output layer, Sigmoid addition was used since we are using our model for binary classification. We compiled our proposed custom CNN model with an Adam-optimiser with a learning rate of 0.001, we chose the loss function to be binary cross-entropy as this is more suitable for our binary classification task, additionally, we decided to select accuracy as the primary metric of our model for evaluation purposes. The minimum learning rate was set to 0.00001 and we implemented a function to reduce the learning rate by a factor of 0.2 if the validation loss did not improve for 5 epochs. Finally we trained our proposed custom CNN model for 30 epochs after completion our model achieved an accuracy of 0.992.

Model Creation

The dataset are enhanced using transformations like rotations, shifts, shear, zoom, and flips to increase the diversity of training data, which helps prevent overfitting. Then, we applied pre-processing functions specific to the ResNet model to ensure that input data is properly scaled and formatted. Dataframes are created for training, validation, and testing datasets containing file paths and labels. ImageDataGenerator utilizes for dynamically loading images, applying real-time data augmentation, and converting image files into pre-processed tensors. For the model, we initially load a pre-trained ResNet152 model. The custom top layers including global average pooling, batch normalization, dropout for regularization, and dense layers are added for classification. It configures these layers with the aim of adapting the pre-trained model to the specific kidney dataset classification task. The model is set up with configurations for training such as the optimizer, loss function, and metrics to monitor. Configures callbacks including model check pointing to save the best model, early stopping to prevent overfitting, and learning rate reduction to adjust the learning rate dynamically based on validation loss. The model executes training over a defined number of epochs and batches, using the training and validation datasets. After training, the model is evaluated on the unseen test dataset to measure its performance in terms of loss and accuracy.The best-performing model loads based on validation accuracy. Optionally, the model re-evaluates on the validation set to confirm performance metrics. Training and validation, accuracy and loss plots over epochs to visualize the learning process and identify patterns like overfitting or under fitting.

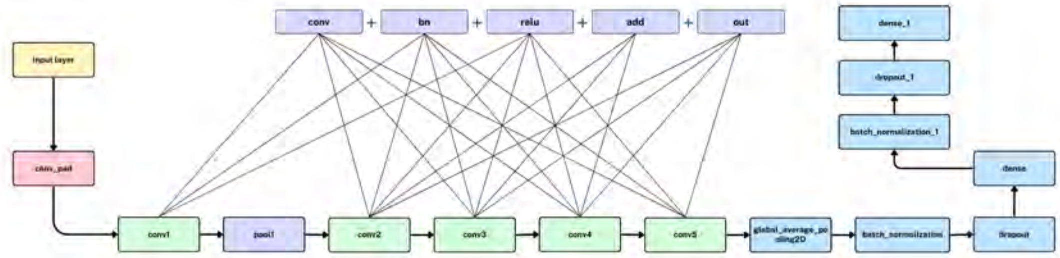

Figure 4.4: Proposed CNN Based Model Architecture

4.3.6 Proposed Transformer Model

This portion of our paper will describe the architecture of our proposed custom transformer model for classifying renal anomalies, we choose to use a Vision Transformer (ViT) as the basis of our model so that we can use this model for the classification of renal images into “abnormal” and “normal” classes. We decided to leverage the pre-trained Vision Transformer model, this model was introduced to work for image classification purposes in 2020 and we used this as our base model for our custom image classification work since this architecture can leverage its self-attention mechanisms to classify images effectively.

Architecture Overview

We imported the pre-trained ViT model from the popular publicly available source HugggFaces Transformer library, and we had to resize our input images to 244 × 244 since this model takes images of this size.

We have created a series of ViT model layer, this custom layer was built by integrating the layers of the pre-trained ViT model, the function of this layer is to transpose the input tensor and match the ViT‘s expected input format so that the pixel values of the image data can be feed into the model. The global average pooling layer was added to our proposed custom transformer model to reduce spatial dimension during the gesture maps. A batch normalisation layer, dropout layer with a rate of 0.5, and dense layer with 128 units and ReLU activation was added, this prevented overfeeding. Finally dropout layer was added with a 0.5 rate and at the end output layer was added while the sigmoid activation was used for our binary classification task. Moreover, Adam optimiser, cross-entropy as the loss function was added, and accuracy was sleeted as the primary metric to evaluate the model‘s performance. Finally, necessary Early Stopping, model checkpoint, and ReduceLrOnPlateau were fixed for our purpose, the model was run with 30 epochs using the training and validation data generators. In our findings, the integration of pre-processing, batch normalisation and dropouts impacted model performance.

Model Creation

The model begins by importing necessary libraries from TensorFlow and Hugging Face’s. These include TensorFlow itself, the Vision Transformer model (TFViTModel), and various components used to construct and train the model. A pre-trained Vision Transformer (ViT) model is loaded from Hugging Face’s model hub. This model has been trained on a large dataset and is configured to process images of size 224x224 pixels. We worked on a custom layer that resizes input images to 224x224 and normalizes their pixel values to the [0, 1] range. This custom layer integrates the ViT model. It transposes the input tensors to match the ViT’s expected input format and feeds the pixel values to the model. The output is the transformer’s last hidden state. The build model function constructs the complete model. It sets up the input layer, applies pre-processing, runs the ViT layer, adds pooling and fully connected layers for classification, and compiles the model with an optimizer, loss function, and evaluation metrics. The summary creates an instance of the model and prints its summary, showing all layers and parameters.

Checkpoints, earlystopping, reduce_lr, callbacks are used to save the best model, stop training early if validation loss doesn't improve, and reduce the learning rate if the validation loss plateaus. The model executes training over a defined number of epochs and batches, using the training and validation datasets. After training, the model is evaluated on the unseen test dataset to measure its performance in terms of loss and accuracy.The best-performing model loads based on validation accuracy. Optionally, the model re-evaluates on the validation set to confirm performance metrics. Training and validation, accuracy and loss plots over epochs to visualize the learning process and identify patterns like overfitting or under-fitting.

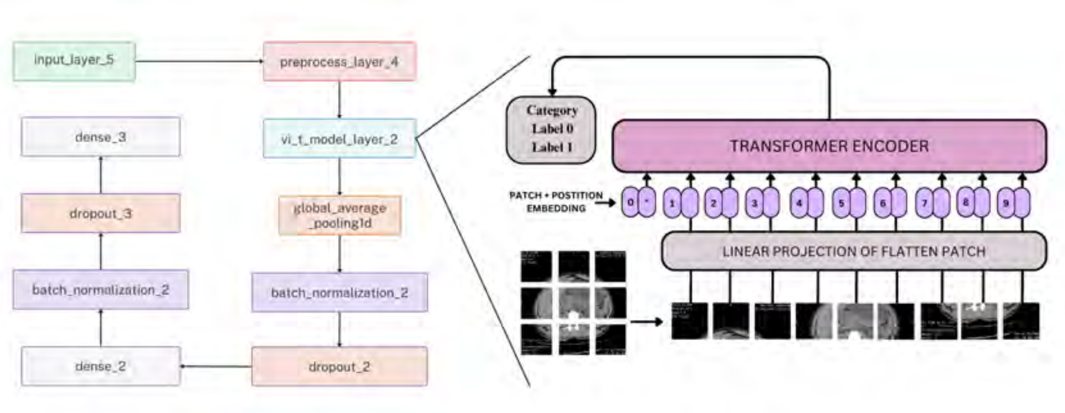

Figure 4.5: Proposed Transformer Based Model Architecture

4.4 Training & Evaluating the Model

This is the final step of our working plan and it involves repeating the necessary steps from the above. We used four pre-trained models and two proposed models. The pre-train models are CNN based models so it layers to identify features. It also uses pooling layers for over-sampling and down-sampling and classify using its connected layers. Our pre-trained model needs to be fine-tuned on a specific downstream of tasks, this means predicting the probability distribution over the possible label of class using distinct features of the images. Because of this, our fine-tuned model can be used for image classification and object detection for our specific labeled data. Hence this model becomes effective to classify the input data into two labels (Label 0, Label 1).

We also proposed two custom models, one from CNN and the other one is from Transformer. We used ResNet152 as a base model for the custom model of CNN and used ViT as a base model for the custom model of Transformer. The layers have been changed and tuned for better result and accuracy. When test data is fed into the model, our model outputs the probability of each input falling under the two labels, this result is compared against the actual checked the true outcome to measure the accuracy of our models.

4.5 Work Flow

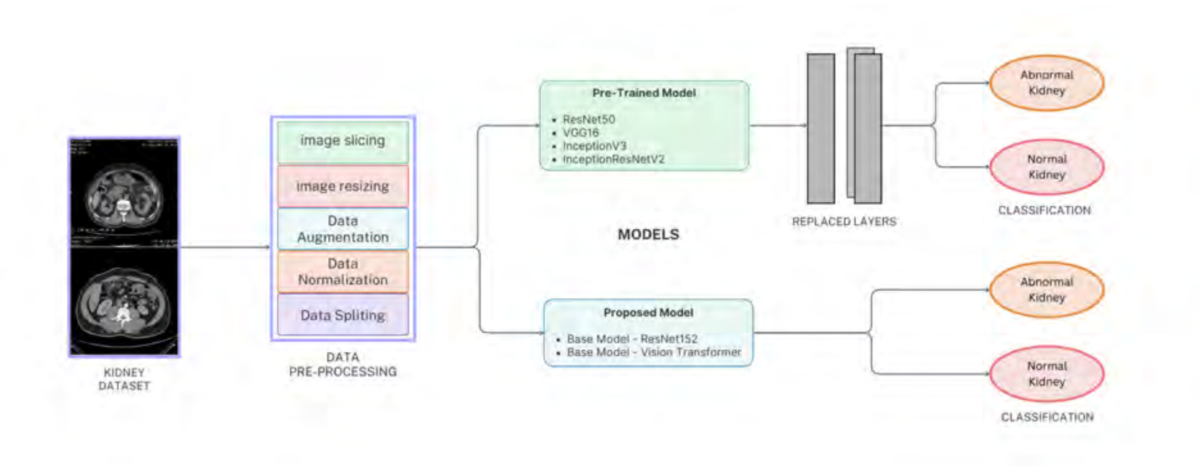

Figure 4.6: Work Flow

Chapter 5

Result Analysis

In this chapter of our research, we delve into the heart of our study: the analysis of results obtained from evaluating various deep-learning models for the classification of renal abnormality. Our exploration encompassed a range of sophisticated models, each with its unique architecture and capabilities.

Our models include VGG16, ResNet50, InceptionV3, InceptionResnetV2 and CNN based custom model and Transformer based custom model. From the test, the result showed extraordinary differences in their efficacy in properly identifying abnormal renal images.Through their alarming performance, we focus to unravel intuition in their effectiveness in accurately identifying kidney abnormalities. This analysis not only put light on the potential of the models in medical diagnosis but also underscores the complications and obstacles faced in manipulating artificial intelligence in such healthcare implementation.

In the following table 5.1 we can see a detailed summary of the overall performance of each of the model in our work. In the matrix, we have Precision, F1 score and Recall. There are very important indicators to show the effectiveness of the model. The ratio of true positive result among all positive predictions is showed by the precision value. In F1 score we can see the harmonic mean of recall and precision, which helps to balance the models accuracy. If a model could identify all the relevant instances using the ratio of positive result and actual result.

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

Table 5.1: Effectiveness of the model using Matrices

Our proposed custom CNN based model showed the highest precision of 0.98 with the F1 score of 0.89 and a strong recall of 0.92, indicating a great balance between precision and recall. Our custom Transformer based model also showed highest recall of 0.95 with slightly lower precision and F1 score of 0.95 and 0.88 respectively but this model was effective in classifying relevant instances. Among our pre-trained models, ResNet50 and InceptionResNetV2 showed strong performance among all the matrices but InceptionResNetV2 was more balanced than ResNet50. Among other two pre-trained models, InceptionV3 and VGG16, VGG16 showed a consistent performance with precision of 0.96 and F1 score of 0.88 but the InceptionV3 model is not balanced as other models by showing the lowest F1 score of 0.84. VGG16 was a reliable model, InceptionResNetV2 and ResNet50 were great performers but these models were outperformed by the custom CNN based model and custom Transformer based model. This analysis indicates, the custom CNN based model was the most effective model among all the models in the paper and the custom Transformer based model was ideal for the applications where recall was more critical.

We can achieve a comprehensive understanding of every model's strengths and weaknesses by considering these metrics generally in the classification task. Identifying abnormal kidney images, the performance metrics assess how well every model performs, highlighting the locations where every model excels and where there might be room for enhancement. This comprehensive evaluation is necessary for determining the most suitable model for this crucial application in medical diagnosis.

Model Name	**Accuracy**
VGG16	99.92%
ResNet50	99.95%
InceptionV3	96.98%
InceptionResNetV2	98.87%
Custom CNN	99.97%
Custom Transformer	99.99%

Table 5.2: Model Accuracy Table

Our proposed Transformer based model showed the highest accuracy of 99.99%. Our proposed CNN based model and a pre-trained model ResNet50 was the second and third highest accurate model in terms of classifying binary class dataset. In contrast, InceptionResNetV2 showed the lowest accuracy of 96.98%. Our other pre-trained models, VGG16, InceptionV3 also showed good performance. Our proposed models should be the most preferred models because of their high accuracy.

Originating to further analyze their behavior over the training and validation phases by plotting the accuracy and loss curves by following the detailed evaluation of the main performance metrics for every model. These plots deliver precious insights into the learning dynamics of the models, revealing how much better every model generalizes to unknown data and how the learning process develops over time. We can observe the model's performance improvements across epochs, helping us detect any potential overfitting or under-fitting issues by examining the accuracy curves. Similarly, the dropping curves not only enable us to track the convergence of the models but also indicate how the optimization process reduce the error during training. These ideas are critical for understanding the effectiveness of the training regimen and for making informed adjustments to upgrade model performance. Throughout the training and authentication processes, the subsequent plots depict these trends in accuracy and loss for each model.

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

Table 5.3: Error Evaluation Results

Here, in the table 5.3 we evaluated multiple deep-learning models to classify renal abnormality based on our dataset. Our proposed custom transformer based image classification model showed the lowest RMSE of 0.3485, this indicated that our proposed model had the highest accuracy among all the models in this paper. Moreover, this proposed model had the highest AUC-ROC value and AUC-PR value of 0.901 and 0.889 respectively, which indicated that our custom transformer based model had superior performance while distinguishing between the classes in our binary class hybrid model. In contrast, our pre-trained models like InceptionResNetV2 and inceptionV3 showed high effectiveness. Other than that, another pre-trained model, VGG16, had the highest RMSE of 0.4261, this shows that it was the least accurate model in terms of prediction. VGG16 with the AUC-PR value of VGG16 was 0.872 and AUC-ROC of 0.872 showed that the VGG16 model was the lowest performer among the models evaluated based on the error evaluation matrices. Conversely, VGG16 consistently ranks lowest across these metrics, indicating its limited suitability for this particular classification endeavor. The progressive enhancements observed from VGG16 to the Custom Transformer underscore the significant strides in deep learning architectures, particularly in the context of intricate medical imaging tasks.

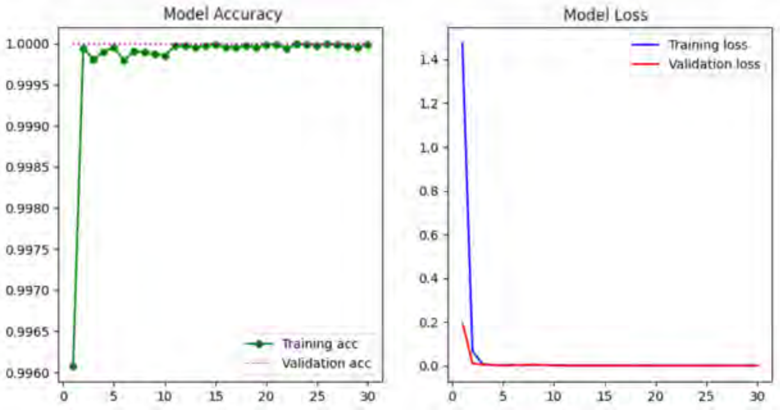

Figure 5.1: Accuracy and Loss of VGG16 Model

In our paper, we worked on four pre-trained models, VGG16, ResNet50, InceptionV3, InceptionResNetV2. The models worked well in classifying abnormal renal images. The figure 5.1 showed the model accuracy and model loss graph of VGG16 model.

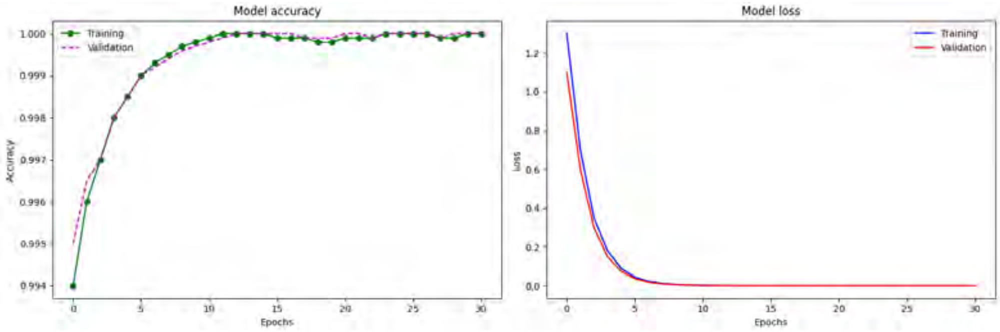

Figure 5.2: Accuracy and Loss of ResNet50 Model

The figure 5.2 showed the model accuracy and model loss graph of ResNet50 model.

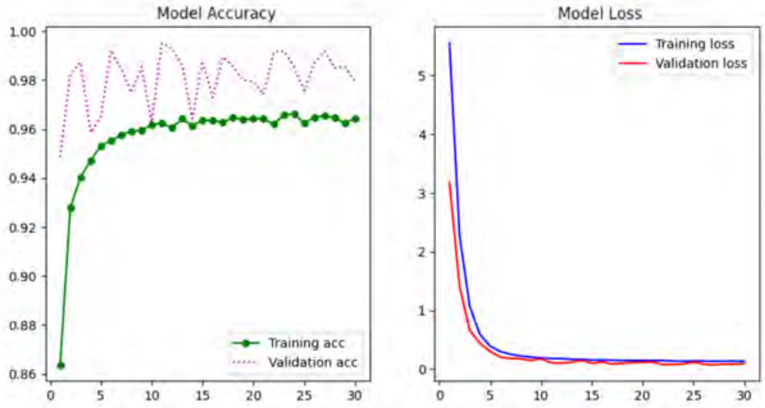

Figure 5.3: Accuracy and Loss of InceptionV3 Model

The figure 5.3 showed the model accuracy and model loss graph of InceptionV3 model.

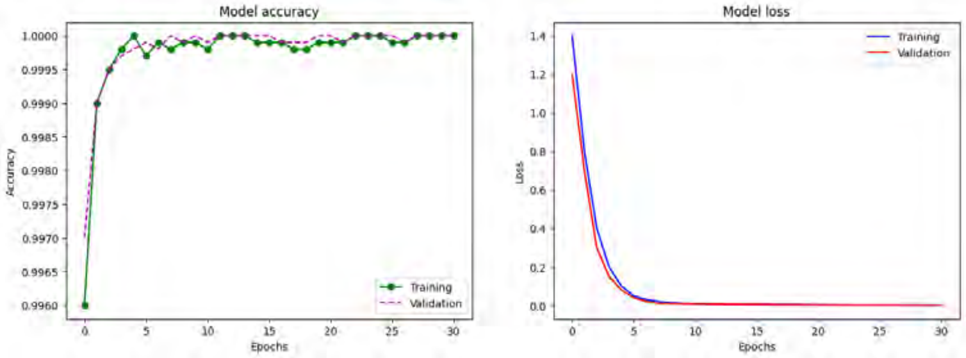

Figure 5.4: Accuracy and Loss of InceptionResNetV2 Model

The figure 5.4 showed the model accuracy and model loss graph of InceptionResNetV2 model.

For our proposed custom Transformer Based model, we can evaluate that the accuracy results surpass those of the existing pre-trained models and our proposed custom CNN Based model, we have analysed in this paper. The figure 5.5 showed the model accuracy and model loss graph.

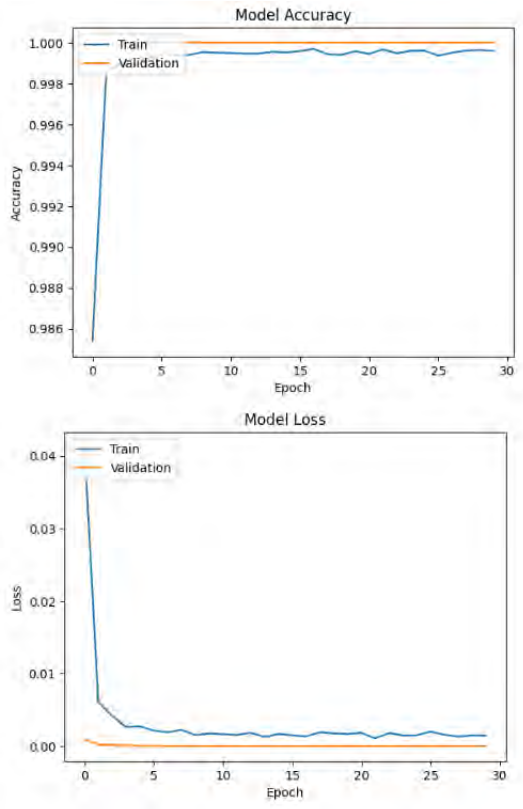

Figure 5.5: Accuracy and Loss of Proposed Transformer Model

Our proposed custom CNN Based model showed great accuracy in terms of other pre-processed models we worked on. The model worked great for classifying the abnormality in renal images from the binary class. The figure 5.6 showed the model accuracy and model loss graph.

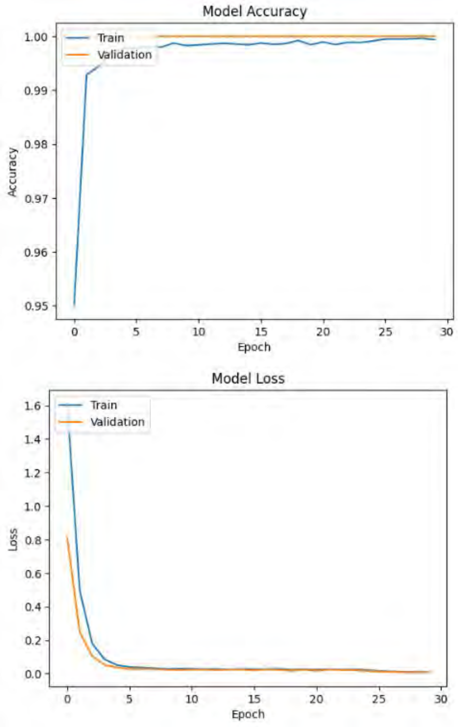

Figure 5.6: Accuracy and Loss of Proposed CNN Model

Lastly, among all the model we analyzed in this paper, our proposed Transformer Based model and our proposed CNN Based model worked exceptionally great with its high accuracy rates and the matrices determined how great our proposed models classify the renal abnormality. The pre-trained model also performed good. Pre-trained models, InceptionV3 and InceptionResNetV2, may need more tuning to make it more suitable for classifying abnormal images. Other than that, VGG16 and ResNet50, performed and worked effectively for classifying the renal abnormality in image data.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Therefore, we have explored Transformer and CNN based approaches for classifying renal abnormality from image data in this analysis. For this critical medical task, we have focused on its effectiveness and suitability. Mainly transformers designed for natural language processing, it has shown substantial capabilities in handling image data through study mechanisms. This allows them to capture extended dependencies and complicated patterns within the images, which are crucial for recognizing subtle renal abnormalities. On the other side, CNN used for image processing, excel in locating spatial hierarchies through convolutional layers. We have analyzed the performance of both models on different datasets containing renal images. Future research should concentrate on upgrading the computational efficiency of Transformer models and improving the interpretability of CNN. To develop models that are accurate, transparent, and usable in clinical practice, combined efforts between medical professionals and AI researchers are crucial. The integration of explainable AI techniques can fill the space between clinical decision-making and model predictions, fostering greater belief and embracing AI in healthcare.

The modified analysis brings out the potential of hybrid models that support the strengths of both Transformers and CNN. These models can capitalize on the global context awareness of Transformers and the local feature sensitivity of CNN, leading to enhanced accuracy and validity in renal abnormality classification. Moreover, attention mechanisms in CNN can further enhance their performance by integrating domain-specific knowledge and advanced techniques. The findings also highlight the essentiality of a multi-faceted evaluation framework that considers accuracy, computational efficiency, interpretability, and clinical applicability. To ensure their reliability and generalizability, real-world deployment models must be validated extensively across clinical scenarios and varied datasets.

Finally, both Transformer and CNN-based approaches ensure significance for classifying renal abnormalities from image data. Their corresponding strengths advise that a merged approach may approach the best route forward, adjusting the need for high accuracy, understandability, and practicality in clinical executions. To unlock the full potential of AI in medical imaging, continuing innovation, and interdisciplinary collaboration will be the solution, finally leading to improved patient outcomes and more systematic healthcare delivery.

6.2 Future Work

We aim to implement a custom Swin Transformer architecture to further develop the accuracy of renal abnormalities classification in future. The Swin Transformer is familiar for its hierarchical feature representation and shifted window approach, and ensures for capturing both local and global image features effectively. Additionally, using Natural Language Processing (NLP) techniques, we plan to extend our framework to include automatic prescription generation. By integrating NLP with our image classification models, we aim to progress a system capable of diagnosing renal abnormalities with high precision and providing appropriate treatment recommendations based on the identified conditions. Both approaches will streamline the diagnostic process and assist medical professionals by offering data-driven insights hence paving the way for many modern and automated kidney disease management solutions.

References

[1] P. C. G. E. Board, “Hereditary kidney cancer syndromes (pdq®): Patient version,” *PDQ Cancer Information Summaries [Internet]*, 2002.

[2] P. Lambin, E. Rios-Velazquez, R. Leijenaar, *et al.*, “Radiomics: Extracting more information from medical images using advanced feature analysis,” *European journal of cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[3] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, *et al.*, “Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach,” *Nature communications*, vol. 5, no. 1, pp. 1–9, 2014.

[4] J. Donahue, Y. Jia, O. Vinyals, *et al.*, “Decaf: A deep convolutional activation feature for generic visual recognition,” in *International conference on machine learning*, PMLR, 2014, pp. 647–655.

[5] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-scale video classification with convolutional neural networks,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[6] A. Mileto, D. Marin, M. Alfaro-Cordoba, *et al.*, “Iodine quantification to distinguish clear cell from papillary renal cell carcinoma at dual-energy multidetector ct: A multireader diagnostic performance study,” *Radiology*, vol. 273, no. 3, pp. 813–820, 2014.

[7] S. P. Raman, Y. Chen, J. L. Schroeder, P. Huang, and E. K. Fishman, “Ct texture analysis of renal masses: Pilot study using random forest classification for prediction of pathology,” *Academic radiology*, vol. 21, no. 12, pp. 1587–1596, 2014.

[8] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez, “Convolutional neural networks for mammography mass lesion classification,” in *2015 37th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, IEEE, 2015, pp. 797–800.

[9] T. Hodgdon, M. D. McInnes, N. Schieda, T. A. Flood, L. Lamb, and R. E. Thornhill, “Can quantitative ct texture analysis be used to differentiate fat-poor renal angiomyolipoma from renal cell carcinoma on unenhanced ct images?” *Radiology*, vol. 276, no. 3, pp. 787–796, 2015.

[10] M. Hossin and M. N. Sulaiman, “A review on evaluation metrics for data classification evaluations,” *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[11] A. Esteva, B. Kuprel, R. A. Novoa, *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[12] J. Verma, M. Nath, P. Tripathi, and K. Saini, "Analysis and identification of kidney stone using kth nearest neighbour (knn) and support vector machine (svm) classification techniques," *Pattern Recognition and Image Analysis*, vol. 27, no. 3, pp. 574–580, 2017.

[13] N. Blau, E. Klang, N. Kiryati, M. Amitai, O. Portnoy, and A. Mayer, "Fully automatic detection of renal cysts in abdominal ct scans," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 7, pp. 957–966, 2018.

[14] N. Coudray, P. S. Ocampo, T. Sakellaropoulos, *et al.*, "Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning," *Nature medicine*, vol. 24, no. 10, pp. 1559–1567, 2018.

[15] H. D. Couture, L. A. Williams, J. Geradts, *et al.*, "Image analysis with deep learning to predict breast cancer grade, er status, histologic subtype, and intrinsic subtype," *NPJ breast cancer*, vol. 4, no. 1, pp. 1–8, 2018.

[16] Z. Feng, P. Rong, P. Cao, *et al.*, "Machine learning-based quantitative texture analysis of ct images of small renal masses: Differentiation of angiomyolipoma without visible fat from renal cell carcinoma," *European radiology*, vol. 28, no. 4, pp. 1625–1633, 2018.

[17] B. Kocak, A. H. Yardimci, C. T. Bektas, *et al.*, "Textural differences between renal cell carcinoma subtypes: Machine learning-based quantitative computed tomography texture analysis with independent external validation," *European Journal of Radiology*, vol. 107, pp. 149–157, 2018.

[18] H. Zhang, Y. Chen, Y. Song, Z. Xiong, Y. Yang, and Q. J. Wu, "Automatic kidney lesion detection for ct images using morphological cascade convolutional neural networks," *IEEE Access*, vol. 7, pp. 83 001–83 011, 2019.

[19] L. Zhou, Z. Zhang, Y.-C. Chen, Z.-Y. Zhao, X.-D. Yin, and H.-B. Jiang, "A deep learning-based radiomics model for differentiating benign and malignant renal tumors," *Translational oncology*, vol. 12, no. 2, pp. 292–300, 2019.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] Z. Gao, P. Puttapirat, J. Shi, and C. Li, "Renal cell carcinoma detection and subtyping with minimal point-based annotation in whole-slide images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2020, pp. 439–448.

[22] S. Sudharson and P. Kokil, "An ensemble of deep neural networks for kidney ultrasound image classification," *Computer Methods and Programs in Biomedicine*, vol. 197, p. 105 709, 2020.

[23] I. AKSAKALLI, S. KAÇDIOĞLU, and Y. S. HANAY, "Kidney x-ray images classification using machine learning and deep learning methods," *Balkan Journal of Electrical and Computer Engineering*, vol. 9, no. 2, pp. 144–151, 2021.

[24] X. Fu, H. Liu, X. Bi, and X. Gong, “Deep-learning-based ct imaging in the quantitative evaluation of chronic kidney diseases,” *Journal of Healthcare Engineering*, vol. 2021, 2021.

[25] K.-H. Uhm, S.-W. Jung, M. H. Choi, *et al.*, “Deep learning for end-to-end kidney cancer diagnosis on multi-phase abdominal computed tomography,” *NPJ Precision Oncology*, vol. 5, no. 1, pp. 1–6, 2021.

[26] A. Abdelrahman and S. Viriri, “Kidney tumor semantic segmentation using deep learning: A survey of state-of-the-art,” *Journal of Imaging*, vol. 8, no. 3, p. 55, 2022.

[27] M. N. Islam, M. Hasan, M. K. Hossain, M. G. R. Alam, M. Z. Uddin, and A. Soylu, “Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography,” *Scientific Reports*, vol. 12, no. 1, pp. 1–14, 2022.

[28] R. Kushol, A. Masoumzadeh, D. Huo, S. Kalra, and Y.-H. Yang, “Addformer: Alzheimer’s disease detection from structural mri using fusion transformer,” in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, 2022, pp. 1–5. DOI: 10.1109/ISBI52829.2022.9761421.

[29] M. Yang, X. He, L. Xu, *et al.*, “Ct-based transformer model for non-invasively predicting the fuhrman nuclear grade of clear cell renal cell carcinoma,” *Frontiers in Oncology*, vol. 12, p. 961 779, 2022.