2026-08

# A two-stage u-net based hierarchical pipeline for kidney and renal tumor segmentation

Dhara, Showndorjo

Independent University, Bangladesh (IUB)


# A Two-Stage U-Net Based Hierarchical Pipeline for Kidney and Renal Tumor Segmentation

August 2026

Prepared by:

Showndorjo Dhara
ID: 2110001

Aeinun Nahar Anny
ID: 2010328

Department of Computer Science and Engineering
Independent University, Bangladesh

Supervised by:

Md Rashedur Rahman, D. Eng
Assistant Professor

Department of Computer Science and Engineering
Independent University, Bangladesh

A Final Year Design Project Report submitted for the Degree of Bachelor's of Computer Science & Engineering

# Attestation

I am aware of the fact that plagiarism is strictly prohibited and it conflicts with my university's rules and regulations. I guarantee the authenticity of my work. I have used some copyrighted material and models of others in my project which has been properly cited following the international standards proper guidelines.

**Author Name:**

. . . . . .

Signature:

. . . . . .

**Author Name:**

. . . . . .

Signature:

. . . . . .

# Acknowledgement

We express our sincere gratitude to our supervisor, Md Rashedur Rahman (D.Eng), for his invaluable guidance throughout this project. We also thank Dr. Saadia Binte Alam for her thoughtful feedback and guidance, which significantly improved the quality and rigor of our work.

We are grateful to the Center for Computational and Data Sciences (CCDS) and Independent University, Bangladesh, for providing computational resources that were essential to our research.

We would also like to acknowledge the technical support provided by Mr. Siam Tahsin Bhuiyan, Mr. Sefatul Wasi, Mr. Fatin Israq, Mr. Riyadul Islam, and Ms. Halima Khatun, whose contributions helped carry this project forward.

Finally, we used Large Language Models (LLM), specifically Claude Sonnet 5 from Anthropic and ChatGPT from OpenAI, to improve the clarity and language of our writing. These tools were used only for editorial purposes, and we confirm that the method, hypothesis, data analysis, result generation, and figures were produced entirely by the authors.

# List of Publications

**A. N. Anny**, **S. Dhara**, S. Wasi, R. Rahman, S. T. Bhuiyan, F. Israq, A. Islam, and S. B. Alam, "A two-stage U-Net based hierarchical pipeline for kidney and renal tumor segmentation," in 2026 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECTTHCON), Dhaka, Bangladesh, Sep. 4–5, 2026.

# Abstract

Kidney cancer diagnosis and treatment planning depend heavily on accurate segmentation of the kidney and any renal tumors from CT scans. Manual segmentation by radiologists is reliable, but it takes time and different radiologists often draw slightly different boundaries. Deep learning offers a way to automate this process. However, many existing solutions rely on three-dimensional models that need heavy computational resources, which are not always available in clinical settings.

This report takes a different approach. It builds a two-stage, two-dimensional U-Net pipeline for kidney and tumor segmentation using the KiTS23 dataset. In the first stage, the model segments the kidney from a full CT slice. In the second stage, it segments the tumor, but only within a cropped region around the kidney. This region is defined using ground-truth kidney masks, so that the two stages can be evaluated independently. Splitting the task this way cuts down background noise and helps with the class imbalance problem that makes tumor segmentation so difficult.

For the experiments, 100 patient cases were selected from KiTS23, following the dataset's official split. This produced 21,140 slices for kidney segmentation and 12,408 kidney-cropped slices for tumor segmentation. Five encoder backbones were tested in the same U-Net setup, under the same training conditions: ResNet-18, ResNet-50, DenseNet-121, EfficientNet-B3, and MobileNetV2. All five were initialized with ImageNet pre-trained weights.

The results show a clear pattern. In kidney segmentation, every encoder performed well, with F1 scores above 0.94. EfficientNet-B3 came out on top, reaching an F1 score of 0.9541 and an IoU of 0.9122. Its compound scaling seems to help it capture the kidney's fairly consistent shape across patients. Tumor segmentation told a different story. Scores were lower overall and varied more between encoders, ranging from 0.6012 to 0.7195 in F1. Here, ResNet-18 performed best, with an F1 of 0.7195 and an IoU of 0.5619, beating out both ResNet-50 and DenseNet-121 by a wide margin.

This reversal is one of the main findings of the report. No single encoder worked best for both tasks. Larger, deeper networks helped with kidney segmentation, where understanding the overall organ shape matters most. But for tumors, which are small and vary a lot in appearance, a lighter and shallower network generalized better. Deeper models seemed to overfit more easily on the smaller, more irregular tumor regions.

Overall, this work shows that a well-designed 2D hierarchical pipeline can achieve strong segmentation results without the computational cost of full 3D approaches. It also shows that encoder choice should not be treated as one-size-fits-all. Matching the encoder to the specific demands of each stage matters, and this report provides a reproducible baseline for future work to build on.

# Contents

# List of Figures

# List of Tables

# Chapter-1: Introduction

Kidney cancer represents a significant and increasing global health burden. The GLOBOCAN database of the World Health Organization indicates that around 434,419 new cases of kidney cancer were reported globally in 2022. During the same year, the disease caused an estimated 155,702 deaths [2]. Over the past three decades, kidney cancer has become more common, especially in many high-income nations. This pattern has been influenced by a number of causes. One significant factor is thought to be the rising incidence of obesity and hypertension. Furthermore, more tiny, asymptomatic kidney cancers that could have gone undetected have been accidentally discovered as a result of the widespread use of cross-sectional imaging [3].

Renal tumors are the primary cause of kidney cancer. Renal cell carcinomas, which originate from the renal tubules' epithelial lining, account for the majority of malignant malignancies. Oncocytomas and angiomyolipomas are among the many benign kidney tumors. Accurate diagnosis is difficult since these benign lesions frequently resemble malignant tumors on medical imaging. Treatment choices are directly impacted by the ability to distinguish between benign and malignant kidney cancers. The best course of action depends on a number of factors, including the size of the tumor, its proximity to the renal sinus or collecting system, and the degree of penetration into the perinephric fat or renal vein. Patients may have radical nephrectomy, nephron-sparing partial nephrectomy, or active surveillance, depending on these features. The stage at diagnosis matters a lot for how kidney cancer turns out. The American Cancer Society reports a five-year relative survival rate of 93% for localized disease. That number drops to 76% for regional disease, and just 19% once the cancer has spread to distant sites [4]. This significant variation emphasizes how crucial it is to identify kidney masses early and accurately. In clinical practice, accurately describing a kidney tumor is frequently more crucial than merely determining its presence.

Contrast-enhanced computed tomography (CT) is widely regarded as the standard imaging technique for diagnosing, staging, and monitoring kidney tumors. It gives radiologists high-resolution images of the renal parenchyma so they can assess lesions according to their shape, enhancing patterns, and connections to surrounding anatomical structures [5]. These CT scans can be used for a number of computational purposes. The first is detection, which assesses the presence and approximate location of a suspicious lesion. The second is categorization, in which a detected lesion is categorized as either benign or malignant, or as a particular subtype of tumor. The third is segmentation, which pinpoints the precise borders of the kidney and the tumor at the pixel or voxel level. Detection and classification are helpful for clinical decision support and diagnosis, but they don't offer the precise spatial information needed in many medical applications. An accurate three-dimensional contour of the tumor and its relationship to surrounding blood

vessels is necessary for surgical planning. In a similar vein, changes in tumor volume over time are used to evaluate therapy response. Precise boundary delineation, as opposed to a straightforward bounding box or class designation, is also necessary for quantitative study of tumor burden. These factors make segmentation essential to the investigation of kidney tumors. It offers the comprehensive anatomical data required for numerous clinical procedures. Thus, segmentation rather than just detection or classification is the main focus of this report [6].

Kidney and tumor segmentation is often carried out manually by qualified radiologists in current clinical practice. This method is dependable, but it takes a lot of time and work. Inter-observer variability, in which various radiologists may generate somewhat different segmentations, also has an impact. These differences may affect later clinical judgments and result in inconsistent volumetric measurements [7].

Due to these constraints, scientists have worked hard to create automated segmentation techniques that don't need much human participation. Medical image analysis has significantly improved due to the quick development of deep learning, especially convolutional neural networks (CNNs). These models have enabled automated segmentation with promising accuracy at a much larger scale than manual methods allow, although expert review remains essential in clinical practice [8]. U-Net has emerged as one of the most popular deep learning models for medical image segmentation. When learning features at various scales, its encoder-decoder structure and skip connections aid in maintaining spatial information [9]. Consequently, U-Net and its variations have demonstrated excellent performance in a variety of anatomical segmentation tasks and medical imaging modalities. U-Net is one of the most widely used architectures in medical image segmentation research due to its continuous success [10].

Even with these developments, it is still difficult to accurately segment renal malignancies from CT volumes. The extreme class disparity in abdominal CT scans is one of the primary causes. Only a small percentage of the total voxels are occupied by tumor pixels; the majority of the voxels are either part of the surrounding background or healthy tissue [11]. Model training is challenging because of this asymmetry. When a model just predicts the background, it frequently converges to a subpar result. The wide variety in kidney tumor appearance presents another difficulty. The size of tumors varies greatly, from tiny, sub-centimeter lesions to massive masses that take up a sizable amount of the renal parenchyma. They also differ in morphology, ranging from uneven and lobulated masses to clearly defined circular lesions. Additionally, different tumor types have different CT intensities. While some tumors are hypodense or cystic, others seem hyperdense and hypervascular [12]. These variations make accurate segmentation difficult when relying only on image intensity. They also make it challenging for a single-stage model to both identify the kidney and precisely delineate tumor boundaries.

Researchers have suggested hierarchical or coarse-to-fine segmentation techniques to get around these problems. In these methods, a larger and more recognizable anatomical structure is first identified using an initial model. A region of interest (ROI) is defined in this stage. The tumor inside the chosen region is then thoroughly segmented using a second model [13]. This method breaks down a challenging multi-class segmentation problem into a number of easier binary tasks. This reduces background interference and improves the model's ability to concentrate on small and heterogeneous tumor areas. The efficacy of this method for organ and tumor segmentation has been shown in earlier research. Additionally,

some studies have shown that hierarchical models achieve competitive or improved results compared to single-stage models on clinically relevant datasets [14]. Beyond increased accuracy, automated segmentation has further advantages. These techniques can produce more accurate volumetric measurements and cut down on the amount of time radiologists must spend manually annotating data. These advancements can promote more accurate diagnostic evaluations in many clinical contexts and improve the reliability of therapy response assessment.

## 1.1 Background

Kidney tumor segmentation research has advanced thanks to a number of publicly accessible benchmark datasets and contests. The Kidney Tumor Segmentation (KiTS) challenges, such as KiTS19, KiTS21, and KiTS23, have significantly advanced this discipline. These datasets offer expert-annotated segmentation masks for the kidney parenchyma and renal cancers in addition to contrast-enhanced abdomen CT scans [1]. Additionally, a consistent evaluation framework for contrasting various segmentation techniques has been provided by the KiTS challenges. Consequently, the availability of these datasets has significantly advanced research in this field. Many studies have been conducted based on them, leading to increasingly precise and trustworthy models over time. This benchmark is further expanded by the KiTS23 dataset, which incorporates numerous contrast phases and more clinical variability. With these changes, KiTS23 is now among the most complete datasets available for kidney tumor segmentation research and more accurately represents routine clinical imaging. The availability of these benchmark datasets has accelerated the development and evaluation of increasingly sophisticated segmentation algorithms.

Traditional image processing approaches like thresholding, morphological procedures, and deformable contour models were the mainstay of early automated kidney segmentation systems. When the kidney borders were distinct and healthy, these techniques worked fairly effectively. However, when tumors altered the CT images' intensity distribution or warped the organ's shape, their performance deteriorated. Subsequently, machine learning techniques like random forests and support vector machines were presented. Compared to conventional methods, these models offered greater versatility and employed handcrafted image features. However, because the quality of the findings was mostly dependent on manually constructed features, their performance was still constrained. Medical image segmentation has significantly improved since deep learning emerged. End-to-end models that learn hierarchical image features directly from the input data may now be trained thanks to the availability of large annotated datasets and more powerful computational resources [15]. Within this paradigm, some of the most popular designs for volumetric medical image segmentation are U-Net and its three-dimensional variations. Furthermore, self-configuring frameworks like nnU-Net dynamically modify training parameters, network architecture, and preprocessing procedures according to the features of a given dataset [16].

The use of pre-trained encoder backbones for renal tumor segmentation has also been investigated in recent works. Usually, models that were first trained on extensive natural image datasets are used to create these encoders. Beyond architectural improvements like nnU-Net, another line of work has focused on improving encoder initialization through transfer

learning. It has been demonstrated that transfer learning from ImageNet-pre-trained models enhances model generalization and shortens training time. Because annotated datasets are frequently scarce and models trained from scratch are more likely to overfit, it is especially helpful for medical imaging tasks [17]. U-Net-based segmentation models have included a number of encoder architectures, such as ResNet, DenseNet, EfficientNet, and MobileNetV2 [18] [19] [20] [21]. The performance of these backbone networks on various segmentation tasks has been assessed. Previous research indicates that no single encoder works well in all circumstances. The dataset's properties and the segmentation task determine which encoder is best. In certain situations, lightweight designs outperform deeper models, particularly when there is a substantial risk of overfitting and little training data is available.

## 1.2 Motivation

The therapeutic treatment of kidney cancer depends on the precise detection, location, and volumetric evaluation of renal malignancies. Despite significant advancements, there are still a number of issues with automatic segmentation techniques. Strong segmentation performance is attained by numerous three-dimensional deep learning models. However, they frequently need for large computational resources, which might not be available in clinical settings with low resources. Furthermore, they do not always outperform well-designed two-dimensional methods [22]. U-Net based two-dimensional models have a number of useful benefits. They may readily employ encoders that have already been trained on big natural image datasets, consume less memory, and train more effectively. However, a lot of current research uses other experimental techniques or evaluates these models on comparatively little datasets. This makes it challenging to adequately evaluate their efficacy in actual clinical applications and to compare results across research.

The absence of a thorough and consistent comparison of several encoder backbones under identical experimental conditions is another drawback of the current research. Many studies compare a limited number of models or concentrate on a single architecture. Because of this, there is little data demonstrating how a larger variety of contemporary encoders operate under the same training settings on the same dataset. Because of this, it is challenging to determine which architectural options are best for segmenting renal tumors as opposed to medical images in general. Kidney tumor segmentation has also been investigated using hierarchical segmentation. Nevertheless, a lot of current techniques rely on three-dimensional networks, which demand a lot of processing power. This restricts its application in settings with constrained hardware. Furthermore, the design decisions that affect two-dimensional hierarchical pipelines' performance have not been thoroughly and consistently investigated.

This report is organized around a single research question: does encoder-backbone optimality remain consistent across both stages of a hierarchical segmentation pipeline, or is it inherently stage-dependent? The study is not intended to introduce a novel architecture. Instead, it is structured as a controlled experiment. The U-Net decoder, preprocessing pipeline, loss function, and training protocol are held constant across all five encoder backbones and both segmentation stages. As a result, any differences observed in performance can be attributed to the encoder architecture itself, rather than to confounding variations

in the experimental setup.

This report aims to address these limitations by developing, implementing, and evaluating a hierarchical two-stage, two-dimensional U-Net framework for kidney and renal tumor segmentation using the KiTS23 dataset. The main goal is to develop a computationally efficient and reproducible segmentation pipeline that can be used as a reliable baseline. The proposed framework also investigates the performance of different encoder backbones in both stages of the segmentation process. The findings are expected to provide practical guidance for selecting suitable encoder architectures while demonstrating the effectiveness of a two-dimensional hierarchical approach for kidney tumor segmentation.

## 1.3 Objective

This report pursues two complementary research objectives:

- To compare the performance of five pre-trained encoder backbones, ResNet-18, ResNet-50, DenseNet-121, EfficientNet-B3, and MobileNetV2, under identical experimental conditions. Each encoder is integrated into the same U-Net architecture and evaluated independently in both segmentation stages. This is done to determine whether encoder-backbone optimality stays consistent across stages, or whether it is fundamentally stage-dependent, and to identify suitable encoder backbones for each task.
- To establish a reproducible benchmark using the official KiTS23 dataset split and standard evaluation metrics. This benchmark is intended to support consistent comparison with existing and future segmentation methods, assuming similar evaluation conditions are followed. It also evaluates whether a well-designed two-dimensional hierarchical pipeline can provide strong segmentation performance while requiring fewer computational resources than three-dimensional approaches.

# Chapter-2: Literature Review

## 2.1 Prior Studies

### 2.1.1 Kidney Segmentation

Kidney segmentation from CT images has been widely studied for more than two decades. During this time, the field has progressed through several stages. Early studies focused on manual and semi-automatic methods, followed by classical image processing and machine learning techniques. More recently, deep learning has become the dominant approach because of its strong segmentation performance. The following subsections review the major developments in each stage.

#### 2.1.1.1 Classical and Machine Learning Methods

Traditional image processing approaches like thresholding, morphological filtering, active contour modeling, and atlas-based registration were the mainstay of early kidney segmentation techniques. When the kidney borders were healthy and distinct, these methods yielded reasonable results. However, in pathological cases when tumors affected the kidney's shape, increased the parenchymal density, or produced hazy borders with surrounding organs, their performance declined. Later, machine learning techniques were developed to enhance segmentation performance. These methods improved their ability to adjust to patient differences by learning from image data. Nevertheless, their total performance was constrained by their continued reliance on manually created features. Many of these constraints were resolved with the invention of deep learning. Deep learning models learned image features directly from the input data rather than using manually created features. Better segmentation performance resulted from the models' ability to collect both local details and more general spatial information thanks to this end-to-end learning process [15].

#### 2.1.1.2 Deep Learning Methods

Kidney segmentation made significant progress with the advent of fully convolutional deep learning models. U-Net emerged as one of the most popular architectures for biomedical image segmentation among these models. The network can retain spatial information while

learning features at various levels because to its encoder-decoder structure and skip links [9]. This method was then expanded to volumetric CT data via the 3D U-Net architecture. It examines three-dimensional image patches rather than processing individual slices. As a result, the model may utilize data from nearby slices that is unavailable in two-dimensional approaches [23]. Hierarchical three-dimensional fully convolutional networks allowed for further advancements. From volumetric CT images, these models are able to segment several organs at once. Additionally, they better capture the anatomical links between adjacent organs, which improves segmentation performance [24].

The advent of self-configuring deep learning frameworks was a significant advancement in medical image segmentation. nnU-Net is one of the most popular instances. Based on the properties of a dataset, this system automatically chooses training parameters, network design, and preprocessing procedures. Consequently, it offers robust baseline performance for a variety of segmentation tasks and minimizes the requirement for manual configuration [16]. Subsequent research assessed nnU-Net's performance and capacity for generalization across various medical image segmentation tasks. Its efficacy as a general-purpose baseline was validated by these investigations. Additionally, they stressed how crucial it is to compare nnU-Net with task-specific models using consistent validation processes [25]. Additionally, transformer-based architectures for kidney and medical image segmentation have been introduced. TransUNet captures both local image features and global contextual information by combining transformer-based self-attention with convolutional neural networks. On a number of organ segmentation tasks, it has outperformed traditional CNN-based models [26]. Swin-Unet is another noteworthy model that employs a transformer-based encoder-decoder architecture with shifted window focus. This solution maintains a respectable level of computing efficiency while capturing long-range image dependence. On various medical image segmentation benchmarks, it has demonstrated competitive performance [27]. Although these architectures improve segmentation accuracy, most rely on computationally expensive three-dimensional processing or increasingly complex attention-based designs, which limits their applicability in resource-constrained environments.

The development of kidney segmentation techniques has been significantly influenced by the KiTS issues. A uniform criterion for assessing kidney and kidney tumor segmentation models was introduced by the KiTS19 challenge. For kidney segmentation, the top-performing techniques obtained Dice values greater than 0.97. However, due to the wide variance in tumor size and appearance [12], tumor segmentation remained far more challenging. The goal of the KiTS23 challenge research was to enhance three-dimensional U-Net models using various training and post-processing techniques. These investigations looked at how segmentation performance was impacted by variables such patch size, batch normalization, and test-time augmentation [28]. The significance of meticulous preprocessing and data augmentation was further emphasized by other techniques created for the KiTS23 challenge. On the highly varied and clinically representative KiTS23 dataset [29], these methods enhanced model performance.

### 2.1.2 Tumor Segmentation

For a number of reasons, segmenting renal tumors is more difficult than segmenting the entire kidney. The size, form, and appearance of tumors vary widely. Additionally, their

limits are more challenging to distinguish because they are frequently situated near the renal sinus and collecting system. Furthermore, only a tiny percentage of the CT volume is occupied by tumor tissue, which causes a significant class imbalance during model training [11]. Researchers have created specific segmentation techniques that go beyond the typical U-Net architecture in order to overcome these issues. These methods are intended to enhance the precise identification of kidney cancers in intricate clinical settings.

Multi-scale feature learning has become an effective approach for renal tumor segmentation, especially because tumors can vary greatly in size. Multiscale supervised three-dimensional U-Net models apply deep supervision at different resolution levels. This allows the network to learn useful features from both coarse and fine image scales. As a result, these models improve the segmentation of small tumors that are often difficult to detect at lower resolutions [30]. Researchers have also combined contextual deformable attention with edge-enhanced U-Net architectures. These models capture image features at multiple spatial scales while improving the delineation of tumor boundaries. They make better use of the spatial relationship between tumors and the surrounding kidney tissue, leading to more accurate segmentation results [31].

Attention mechanisms have been widely used in encoder-decoder architectures for tumor segmentation. They help the model reduce the influence of irrelevant background regions and focus more on the small areas that contain tumor tissue. This improves the model's ability to identify and segment challenging lesions. The FR2PAttU-Net model introduced feature recalibration and pyramid pooling attention modules into the U-Net architecture for kidney tumor segmentation. These attention mechanisms enhance multi-scale feature learning and improve the detection of small and heterogeneous tumors [32]. EfficientNet-based encoders have also been incorporated into U-Net models for kidney tumor segmentation using CT images. Studies have shown that the compound scaling strategy of EfficientNet provides effective multi-scale feature representations. This makes the architecture well suited for segmenting renal tumors with large variations in size [33].

Class imbalance is one of the major challenges in renal tumor segmentation because tumor regions occupy only a small portion of the CT image compared with the background. To address this issue, several specialized loss functions have been proposed. One example is Unified Focal Loss, which combines Dice loss and cross-entropy loss within a single framework. It also applies focal weighting to reduce the influence of easily classified background pixels and place greater emphasis on difficult tumor regions. This approach has shown consistent improvements on several medical image segmentation tasks with severe class imbalance [34]. Researchers have also investigated how inference strategies affect segmentation performance. A UNet-PWP model combined with inference distribution optimization and explainability analysis showed that factors such as patch sampling and prediction aggregation can improve tumor segmentation accuracy. The study also highlighted their importance for interpreting model predictions in clinical applications [35].

Renal tumor analysis has been expanded beyond binary segmentation in recent research. Certain methods segment tumors and classify different subtypes of clear cell renal cell carcinoma. By examining their enhancement patterns across several contrast phases, convolutional neural networks trained on multiphase CT images have been used to segment renal masses and differentiate between benign and malignant lesions [36]. Instead than just focusing on the tumor, other research has segmented several renal structures. For

instance, RenalSegNet automatically segments renal malignancies in contrast-enhanced CT images together with the renal veins and arteries. This offers more anatomical data that can help with clinical applications and surgical planning [37].

### 2.1.3 Hierarchical and Multi-Stage Segmentation

Hierarchical segmentation is a widely used approach for organ and tumor segmentation. In this strategy, an initial stage first identifies the organ and defines a region of interest (ROI). A second stage then performs detailed segmentation within that region. This approach has been shown to reduce the effects of class imbalance and background interference in medical image segmentation tasks. The main idea behind this strategy is that tumor segmentation becomes easier when the model focuses only on the organ that contains the tumor. Restricting the search area removes most of the background pixels and allows the model to concentrate on the relevant anatomical region. This simplifies the segmentation task and helps the model learn more accurate tumor boundaries.

A broad standard for assessing medical image segmentation techniques across many organs and illnesses was developed by the Medical Segmentation Decathlon. The findings demonstrated that performance is frequently more affected by elements like data augmentation, normalization, and suitable loss functions than by modifications to network design alone [38]. Kidney and tumor segmentation have also been done using the two-stage segmentation method. The initial network in Yao et al.'s automated system locates the kidney and creates areas of interest (ROIs). Tumor segmentation within these ROIs is then carried out by the second network. Compared to a single-stage segmentation approach, their results demonstrated superior performance [13]. Similar dual-stage frameworks identify the tumor within the cropped kidney region after first segmenting the kidney. This study showed that the performance of tumor segmentation in the second stage is directly impacted by the accuracy of kidney segmentation in the first stage [14].

Analysis of kidney abnormalities in CT images has also been done using the segmentation-guided approach. For a subsequent classification or segmentation task, this architecture uses an initial segmentation stage to provide spatial information. Research has demonstrated that integrating lesion characterisation and localization can enhance performance on clinically relevant diagnostic tasks [39]. Cross-dimensional transfer learning has also been investigated for medical image segmentation. These techniques use the advantages of both approaches by transferring learnt features between two-dimensional and three-dimensional models. The objective is to integrate the greater spatial information found in volumetric models with the efficiency of slice-based processing [40].

### 2.1.4 Encoder Backbone Architectures

An essential component of a U-Net-based segmentation model is the encoder backbone selection. It has an impact on both the computational cost of training and inference as well as the quality of the derived image features. One of the most used encoder architectures is ResNet. In order to solve the vanishing gradient issue in deep neural networks, it provides residual shortcut connections. Both ResNet-18 and ResNet-50 have demonstrated

high performance on many tasks and have been effectively applied to medical image segmentation [18]. Another popular encoder architecture is DenseNet. Each layer receives feature maps from every layer before it thanks to tight connections. During training, this design enhances gradient flow and promotes feature reuse. When training data is few, these features are particularly helpful [19].

EfficientNet is designed to improve performance by scaling the network depth, width, and input resolution in a balanced way. This compound scaling approach allows the model to achieve high accuracy while using fewer parameters than many conventional architectures [20]. When used as the encoder in U-Net models for kidney tumor segmentation, EfficientNet has shown good performance by learning image features at multiple scales. This makes it suitable for segmenting renal tumors that vary considerably in size [33]. MobileNetV2 is another widely used encoder architecture. It is based on inverted residual blocks and depthwise separable convolutions, which reduce computational cost while maintaining effective feature extraction. Because of its lightweight design, MobileNetV2 is well suited for applications where computational resources are limited or fast inference is required [21]. Taken together, these encoders reflect a trade-off between representational power and computational cost. ResNet and DenseNet tend to offer strong feature extraction and stable training, making them suitable when sufficient computational resources are available, whereas EfficientNet achieves comparable or better accuracy with substantially fewer parameters through its balanced scaling strategy. MobileNetV2, in contrast, prioritizes low latency and reduced computational overhead, often at some cost to representational capacity, which makes it a practical choice for resource-constrained settings such as the one considered in this study.

In medical image analysis, initializing encoder backbones with weights pre-trained on large image datasets like ImageNet has become standard procedure. During training, transfer learning speeds up the convergence of models and frequently enhances their capacity for generalization. These advantages are particularly significant when there is a scarcity of labeled medical imaging data [17]. The degree to which pre-trained features transfer to medical imaging tasks has also been the subject of numerous investigations. According to their findings, the degree of feature transfer during training and the similarity between the source and target datasets determine how successful transfer learning is. When choosing encoder backbones and determining how to optimize them for a particular segmentation task, these observations offer helpful guidance.

### 2.1.5 Data Augmentation and Training Strategies

An essential component of training medical image segmentation models is data augmentation. It is particularly helpful when there is a limited quantity of annotated data. Scaling, flipping, and random rotation are common geometric augmentation methods. Gaussian noise addition, contrast alteration, brightness correction, and other intensity-based augmentations are also frequently employed. These methods decrease overfitting and broaden the variety of training data [41]. Additionally, more sophisticated augmentation techniques based on oversampling have been presented by researchers. These methods produce more training examples for the dataset's underrepresented areas. It has been demonstrated that these tactics increase model resilience, especially in medical image segmentation problems with significant class imbalance [42].

The quick development of deep learning in medical image processing has been summed up in recent review articles. Numerous applications, such as image segmentation, detection, classification, and registration, are covered in these reviews. Additionally, they address a number of persistent issues, including algorithmic bias, insufficient validation using multi-center datasets, and the disconnect between research findings and practical clinical use [43]. According to these studies, the science has advanced to a point where variables other than model architecture are becoming more significant. It is presently thought that developing trustworthy techniques that may be used in clinical practice requires careful experimental design, strict evaluation procedures, and open reporting of outcomes.

### 2.1.6 Summary and Research Gap

This review shows that segmentation performance has improved through architectures like U-Net, nnU-Net, TransUNet, and Swin-Unet. However, most high-performing models rely on complex or three-dimensional designs, which limits their use in resource-constrained settings. Looking more closely at the literature, three related gaps stand out, and this report addresses all three.

First, prior hierarchical or cascaded pipelines typically report only a single winning encoder backbone per stage. This gives a point estimate for that design choice, not a systematic comparison across candidate architectures. Second, these studies routinely assign different network families to each stage by design. A heavier network is often used for organ localization, and a lighter one for lesion detection. However, the counterfactual assignments are rarely tested symmetrically. Because of this, it remains unclear whether the reported encoder choice is actually optimal, or simply convenient. Third, no existing study reports a same-architecture, same-hyperparameter, both-stage ablation across a common pool of encoders. This kind of comparison is necessary to expose whether architectural sensitivity is truly stage-dependent, or just an artifact of how each study happened to set things up.

In short, the gap identified here is not a missing model. It is a missing controlled comparison. This report addresses these gaps by developing a reproducible, two-dimensional hierarchical segmentation pipeline. It is evaluated on the official KiTS23 dataset split. The aim is to achieve strong performance with lower computational cost than three-dimensional approaches.

## 2.2 Challenges

Kidney tumor segmentation presents a number of well-known technical difficulties. Severe class disparity is one of the most important. Only a small percentage of the CT volume is made up of tumor voxels; the majority of voxels are either background or normal tissue. This imbalance might lead to a model predicting merely the background rather than accurately identifying tumor locations, which complicates model training. The wide variety in kidney tumor appearance presents another difficulty. The size of tumors varies; they might be huge masses or tiny, sub-centimeter lesions. Their forms might range from lobular and irregular formations to clearly delineated borders. Additionally, they exhibit distinct

patterns of intensity on CT scans, such as hypodense cystic lesions and hypervascular hyperdense tumors [12]. Standard loss functions like binary cross-entropy might not offer the optimum segmentation performance due to these difficulties. As a result, several studies employ mixed loss functions or dice-based loss functions that minimize the impact of the dominating background class while giving the foreground tumor region more weight [34].

Although technique is computationally efficient, two-dimensional slice-based segmentation has certain drawbacks. The model is unable to fully utilize the spatial information provided across adjacent CT slices because each slice is analyzed independently. This could result in discrepancies between neighboring slices and lower the final three-dimensional reconstruction's accuracy [15]. Segmenting very tiny tumors is also challenging. They are only shown by a few pixels in some CT slices. Because of this, even little segmentation mistakes can have a discernible impact on the outcome. Performance evaluation may also be less trustworthy because to the scarcity of tiny tumor samples. The inconsistency of manual annotations presents another difficulty. For the same image, various radiologists may create somewhat different segmentation masks. This variance introduces label noise that can impact both model training and evaluation, even though it is often smaller for kidney segmentation than for tumor segmentation [7].

## 2.3 Scope of Study

Using a hierarchical two-dimensional U-Net framework on the KiTS23 dataset, this study focuses on automated kidney and renal tumor segmentation. Three primary design decisions determine the study's scope. First, a fully three-dimensional pipeline is replaced with a two-dimensional slice-based method. This decision offers a segmentation framework that is simpler to replicate and execute, lowers computational needs, and permits the use of ImageNet pre-trained encoder backbones. Second, there are two consecutive binary steps in the segmentation task. The kidney is segmented in the first stage, and the tumor inside the renal region is segmented in the second step. Reducing background interference and improving the detection of small tumors are two benefits of restricting the search region to the kidney. Third, using the identical experimental conditions, the study contrasts five encoder backbone topologies in both segmentation stages. This helps find appropriate encoder backbones for kidney and tumor segmentation and offers a fair evaluation of their performance. This work provides a reproducible evaluation framework for hierarchical two-dimensional kidney tumor segmentation by integrating these design decisions. Additionally, it offers an alternative to completely automated frameworks like nnU-Net, where many architectural choices are made automatically [25], while addressing the dearth of thorough encoder comparisons in the literature.

# Chapter-3: Dataset

## 3.1 Dataset Description

This study uses the KiTS23 dataset as the primary source of CT images for training, validation, and evaluation of the proposed hierarchical segmentation framework. KiTS23 is the latest version of the Kidney Tumor Segmentation (KiTS) challenge dataset and extends the earlier KiTS19 and KiTS21 datasets [1]. The dataset contains standardized contrast-enhanced abdominal CT scans collected from patients who underwent partial or radical nephrectomy for suspected renal malignancy, along with expert-annotated segmentation masks for the kidney parenchyma and renal tumors. Compared with the previous KiTS datasets, KiTS23 includes greater clinical diversity, with CT scans acquired using different contrast phases, scanner protocols, and a larger patient population. These characteristics make the dataset more challenging and provide a better representation of clinical imaging data for kidney and renal tumor segmentation research [29]. This diversity is important for developing models that generalise well beyond a single clinical setting, since real-world deployment requires robustness to variation in scanner hardware and acquisition protocol.

The KiTS23 dataset contains more than 300 patient cases stored in NIfTI format [44]. Each case includes a three-dimensional abdominal CT volume and a corresponding segmentation mask, with each voxel labeled as background, kidney parenchyma, or tumor. The segmentation masks were created by trained clinical annotators and are used as the ground truth for evaluating segmentation models. In this study, a subset of 100 cases was selected from the full dataset due to the computational cost of training and comparing multiple encoder backbones across two segmentation stages. The selected cases were chosen to include a wide range of tumor sizes, shapes, and contrast enhancement patterns, ensuring that the subset remains representative of the full dataset's variability. This provides a diverse and manageable sample for evaluating the proposed segmentation pipeline. Selecting a representative subset also allows multiple encoder architectures to be trained and compared within a reasonable computational budget, without sacrificing the clinical variability needed for a meaningful evaluation.

The CT scans in the KiTS23 dataset include a wide range of Hounsfield Unit (HU) values because they contain different tissue types and were acquired using different contrast protocols. In portal venous phase CT images, the kidney parenchyma typically has HU values between 100 and 200. Tumor intensity varies considerably depending on factors such as vascularity and the timing of contrast enhancement. Because of these variations, segmentation methods that rely only on image intensity are often not sufficient. Deep

Table 3.1: Summary of KiTS23 Dataset.

| **Attribute** | **Details** |
|---|---|
| Modality | CT (Contrast-enhanced, multiple phases) |
| Total Cases (Full Dataset) | 300 |
| Cases Used in This Study | 100 |
| Annotation Classes | Kidney, Tumor, Cyst |
| Classes Used in This Study | Kidney (Stage 1), Tumor (Stage 2) |
| Original Format | NIfTI |
| Annotation | Clinical annotators, reviewed by radiologists |

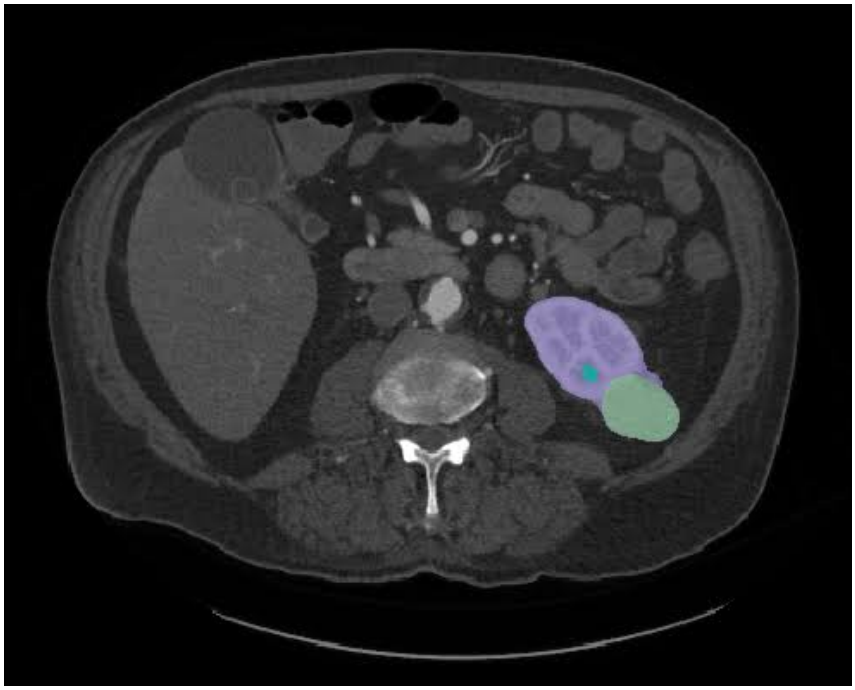

Figure 3.1: Example annotated axial CT slice from the KiTS23 dataset showing kidney (purple), tumor (green), and cyst (blue) labels [1].

learning models can learn more complex spatial and contextual features, making them better suited for this task. The KiTS23 dataset is designed to support the development and evaluation of fully automated segmentation methods without requiring manual interaction during inference [12]. This makes it a suitable benchmark for evaluating hierarchical, encoder-based segmentation pipelines such as the one proposed in this study.

## 3.2 Dataset Preprocessing

A preprocessing pipeline was created to transform the volumetric data into standardized two-dimensional image-mask pairs for training because the proposed model has a two-dimensional architecture while the KiTS23 dataset is supplied as three-dimensional CT volumes. There are two distinct branches to the preprocessing procedure. While one branch prepares the data for tumor segmentation, the other prepares the data for kidney segmentation. Every branch is created in accordance with the specifications of the

segmentation stage that it corresponds to. Although the two branches share several common steps, such as orientation alignment and HU normalisation, they differ in output resolution and mask construction to match the requirements of each stage. This separation ensures that each segmentation stage receives inputs that are appropriately scaled and formatted for its specific task.

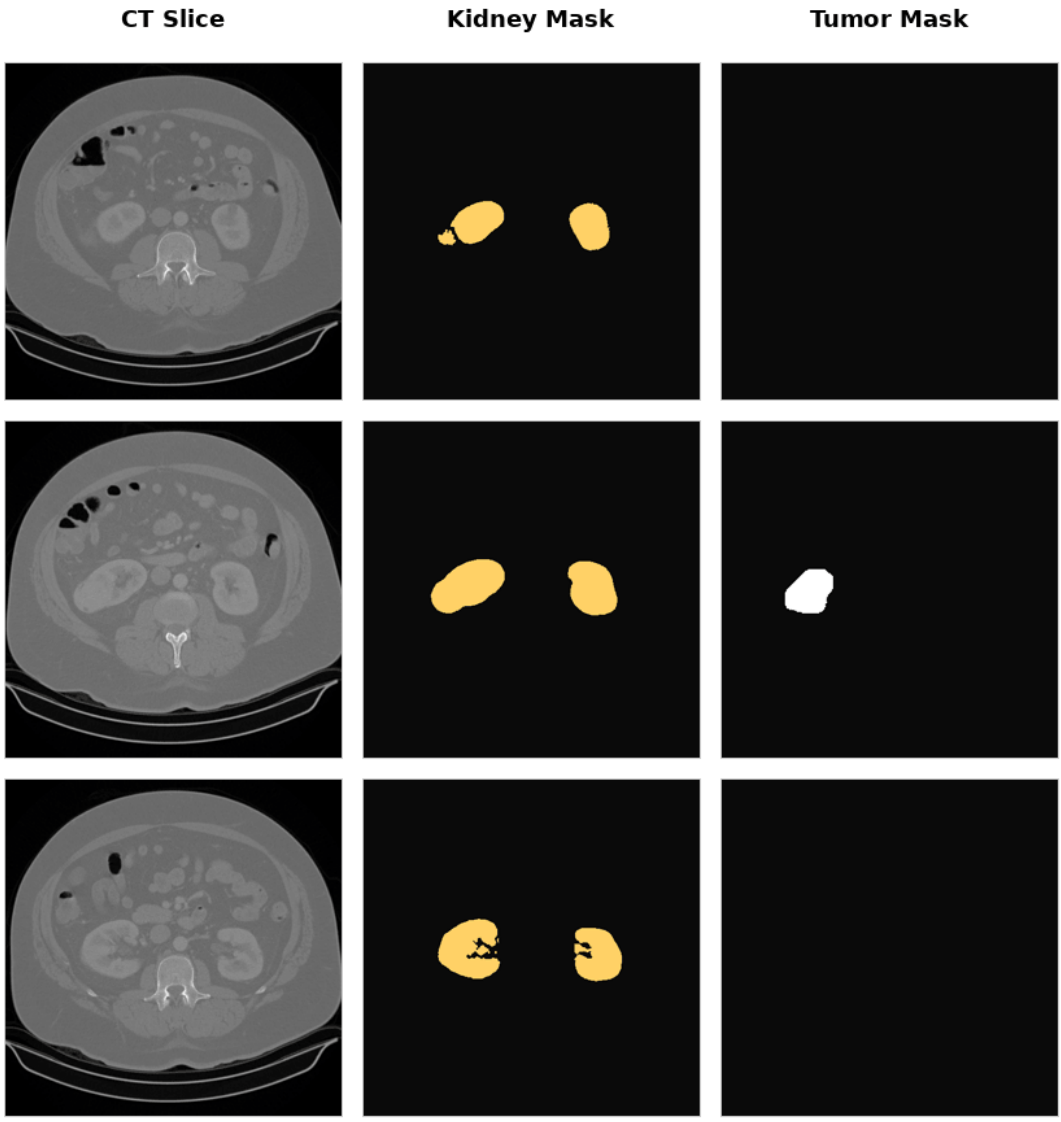


Figure 3.2: Representative axial CT slices with corresponding kidney and tumor segmentation masks. Each row shows a single slice across three views: the original CT image (left), the ground-truth kidney mask (middle, shown in yellow), and the ground-truth tumor mask (right, shown in white).

## 3.3 Dataset Preparation

The 100 selected KiTS23 cases were divided into training, validation, and test sets for each stage of the segmentation pipeline. The official KiTS23 data split was followed to ensure that the experiments are reproducible and that the results can be compared fairly with other methods evaluated on the same benchmark. The dataset was split at the subject level rather than the slice level. This means that all CT slices from a single

patient were assigned to only one subset, either training, validation, or testing. No patient appeared in more than one subset. Subject-level splitting is important in medical image analysis because a single CT volume can produce hundreds of axial slices. If slices from the same patient were included in both the training and test sets, the model could learn patient-specific anatomical features during training. This would introduce data leakage and lead to overly optimistic performance estimates. Splitting the data by subject provides a more reliable evaluation because it measures how well the model performs on previously unseen patients, which better reflects real clinical use [38].

Table 3.2: Dataset Splits for the Two-Stage Pipeline.

| **Task** | **Training Slices** | **Validation Slices** | **Test Slices** |
|---|---|---|---|
| Kidney Segmentation | 13,485 | 2,969 | 4,686 |
| Tumor Segmentation | 7,954 | 1,827 | 2,627 |

For the kidney segmentation stage (Stage 1), the 100 selected cases produce a total of 21,140 two-dimensional axial slices. This is because every axial slice across the full CT volume is included, even those without visible kidney tissue. These slices are divided into 13,485 training slices, 2,969 validation slices, and 4,686 test slices. For the tumor segmentation stage (Stage 2), the kidney-guided ROI extraction process generates 12,408 slices. Only slices containing kidney tissue are retained for this stage. The dataset consists of 7,954 training slices, 1,827 validation slices, and 2,627 test slices. Although ROI extraction reduces the overall number of slices, the left and right kidneys are cropped separately. As a result, a single patient can contribute two kidney-centered image crops from the same axial slice.

The difference in the number of slices also reflects the preprocessing strategy used in each stage. In Stage 1, the entire axial slice is resized to $224 \times 224$ pixels and used for kidney segmentation. Slices without kidney tissue are retained as negative examples so that the model learns to identify both kidney and background regions correctly. In Stage 2, only kidney-centered regions of interest are processed. Slices containing kidney tissue are cropped, and separate image crops are generated for the left and right kidneys. This increases the number of training samples while allowing the model to focus on the relevant anatomical region. The approach follows the hierarchical segmentation strategy and helps reduce the effect of severe class imbalance in renal tumor segmentation [11].

# Chapter-4: Methodology

The proposed two-stage hierarchical segmentation paradigm employed in this investigation is explained in this chapter. The U-Net design, the five pre-trained encoder backbones, the training configuration, the evaluation technique, and the transfer learning approach are all presented. The methodology is intended to offer a consistent and equitable comparison of various encoder backbones. Two segmentation tasks with varying degrees of complexity are used to assess all models under identical experimental settings.

## 4.1 Proposed Method

Figure 4.1 shows the overall pipeline of the proposed framework. The pipeline has four main stages. First, raw CT volumes are preprocessed into 2D slice-mask pairs. Second, a U-Net model segments the kidney from each slice. Third, ground truth kidney masks are used to extract kidney-centered ROIs. Fourth, a second U-Net model segments the tumor within each ROI. The same five pre-trained encoder backbones are used in both stages: ResNet-18, ResNet-50, DenseNet-121, EfficientNet-B3, and MobileNetV2. Each encoder is trained and evaluated independently in both stages. This allows encoder performance to be compared without errors from one stage affecting the other, and helps identify which architecture suits coarse organ localization versus fine-grained lesion detection.

### 4.1.1 Preprocessing

Each raw 3D NIfTI CT volume is first reoriented to a standard anatomical coordinate system. This is done using `nib.as_closest_canonical`. Reorientation ensures that all volumes follow a consistent axis convention regardless of the original scanner protocol. The volume is then split into 2D axial slices. Each slice is paired with its corresponding ground truth mask. This slice-based approach allows the use of 2D encoder backbones pre-trained on ImageNet [17]. It also reduces the computational cost compared to training a full 3D segmentation network.

Every slice is clipped to a soft-tissue Hounsfield Unit (HU) window. This step removes irrelevant intensity ranges such as air, bone, and metal artifacts. The clipped intensities are then min-max normalized to the range [0, 255]. Normalization improves training stability and helps the network converge faster. The normalized slice is converted into an 8-bit grayscale image. A fixed 90-degree rotation and a horizontal flip are applied to every

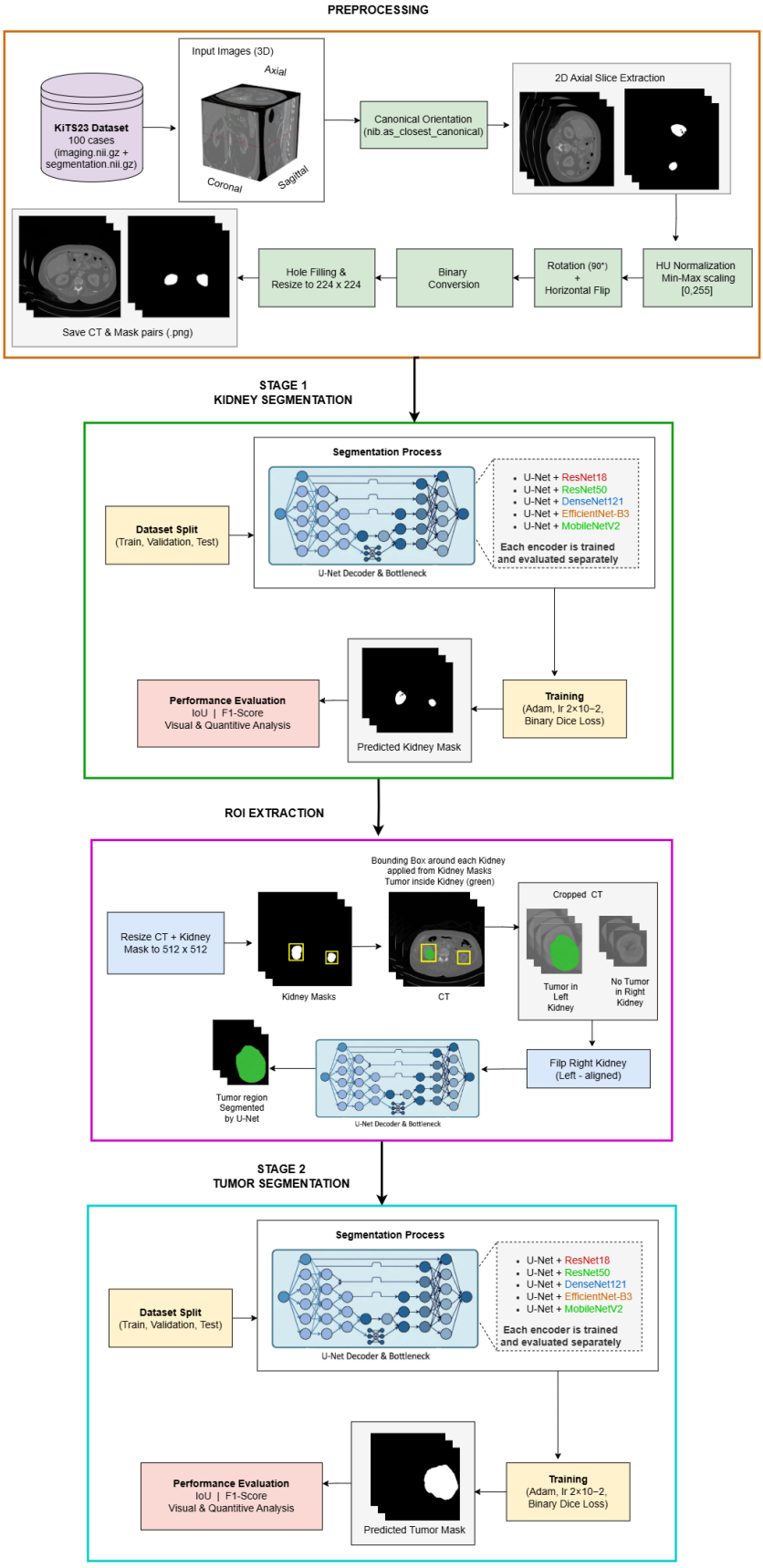


Figure 4.1: Overview of the proposed two-stage hierarchical segmentation pipeline. Stage 1 produces binary kidney masks from full axial CT slices. Stage 2 uses kidney-guided ROI crops to produce binary tumor masks.

slice. This ensures a consistent anatomical orientation across all cases, since raw NIfTI volumes can vary in axis ordering between scanners.

The original multi-class ground truth mask contains separate labels for kidney, tumor, and cyst. These labels are merged into a single binary foreground class for Stage 1. This simplification is necessary because Stage 1 is only concerned with organ-level localization, not with distinguishing internal structures. A morphological hole-filling operation is then applied using OpenCV [45]. This makes the binary masks smoother and more continuous, and removes small annotation gaps that could otherwise mislead the network during training. Each image-mask pair is resized to $224 \times 224$ pixels. This resolution balances segmentation detail with computational efficiency. The final CT slices and masks are saved as PNG file pairs for training. Slices with no foreground content are optionally filtered out to reduce class imbalance during training.

### 4.1.2 Stage 1: Kidney Segmentation

In Stage 1, a U-Net model is trained to produce a binary kidney mask from each preprocessed axial slice. The dataset is split into training, validation, and test sets. This split ensures that model performance is evaluated on unseen cases rather than on data used during training. One of five pre-trained backbones replaces the standard U-Net encoder in each experiment. A bottleneck layer and skip connections link the encoder to the decoder. Skip connections preserve spatial detail that would otherwise be lost during downsampling, which is important for accurate boundary prediction. The decoder is initialized from scratch for every encoder configuration. This ensures that any performance difference comes from the encoder architecture alone, rather than from differences in decoder design.

All encoder backbones start from ImageNet pre-trained weights. These weights are fine-tuned during training. Transfer learning from ImageNet allows the network to leverage general low-level visual features, such as edges and textures, even though the source domain differs from medical imaging. The model is trained using the Adam optimizer with a learning rate of $2 \times 10^{-4}$. A binary Dice loss function is used to handle the foreground-background imbalance, since kidney regions typically occupy only a small portion of the full axial slice. The trained model outputs a predicted kidney mask for each input slice.

Model performance is evaluated using Intersection over Union (IoU) and F1-score. Both metrics are standard choices for segmentation tasks and provide complementary insight into overlap accuracy and boundary precision. Both visual and quantitative analyses are performed on the test set. Visual inspection helps identify systematic failure cases, such as under-segmentation near kidney boundaries, that quantitative metrics alone may not reveal. This evaluation is repeated independently for each of the five encoder backbones, allowing a direct ranking of encoder performance for the kidney segmentation task.

### 4.1.3 ROI Extraction

Kidney-guided regions of interest (ROIs) are built directly from the raw dataset rather than from the Stage 1 predictions. Ground truth kidney masks are used for this step, both

during training and during evaluation of Stage 2. This design choice allows Stage 1 and Stage 2 to be evaluated independently. It isolates tumor segmentation performance from any errors introduced by the kidney segmentation stage, making it possible to determine whether encoder suitability is stage-dependent. This separation is particularly important because a poorly performing kidney segmentation model could otherwise mask the true tumor segmentation capability of a given encoder.

Before cropping, each CT slice and its ground truth kidney mask are resized to $512 \times 512$ pixels. This preserves spatial detail during the crop, which is critical since tumors can be small relative to the full kidney region. The same HU normalization used in Stage 1 is applied again at this stage. Connected component analysis is then applied to the kidney mask. The centroid position of each component is used to separate the left and right kidney. This step is necessary because a single axial slice may contain both kidneys, and each must be processed as an independent ROI.

A bounding box is generated for each kidney region. Five pixels of padding are added to each side of the box. The padding ensures that tumor tissue near the kidney boundary is not accidentally excluded from the crop. The same bounding box is applied to the CT slice and the corresponding tumor mask. This produces a kidney-centered crop for each kidney. All right-kidney crops are horizontally flipped to match the left-kidney orientation. This gives a consistent anatomical layout across all ROI crops, which simplifies learning for the Stage 2 network since it no longer needs to account for left-right anatomical variation.

### 4.1.4 Stage 2: Tumor Segmentation

In Stage 2, a separate U-Net model is trained to predict a binary tumor mask from each kidney ROI crop. The same five pre-trained encoder backbones from Stage 1 are used here. Each encoder is again trained and evaluated independently. The dataset split, encoder initialization, and decoder setup follow the same protocol as Stage 1. Keeping these conditions identical ensures that any performance difference between the two stages can be attributed to the nature of the task rather than to inconsistencies in the experimental setup.

Training uses the Adam optimizer with a learning rate of $2 \times 10^{-4}$. Binary Dice loss is again used as the objective function. This keeps the training configuration identical between the two stages. The model outputs a predicted tumor mask for each ROI crop. Because tumors vary widely in size, shape, and contrast enhancement, the network must learn more fine-grained features than in Stage 1, where the kidney boundary is comparatively more consistent across cases.

The predicted tumor masks are compared against the expert-annotated ground truth. IoU and F1-score are used as the evaluation metrics. Visual and quantitative analyses are performed for each encoder backbone. This allows a direct comparison of encoder performance across both the kidney and tumor segmentation tasks, and shows whether the best-performing encoder differs between the two stages. The results of this comparison are discussed in detail in the following section, along with an analysis of failure cases for each encoder architecture.

## 4.2 Segmentation Architecture: U-Net

### 4.2.1 U-Net Overview and Design Philosophy

The U-Net architecture, a convolutional neural network created for biomedical image segmentation [9], forms the basis of both phases in the proposed workflow. U-Net was created to address a frequent problem in medical image analysis: reliable pixel-level segmentation with comparatively little annotated training data. An encoder-decoder structure with skip links between matching encoder and decoder layers makes up the architecture. Fine spatial information that may otherwise be lost during the encoder's downsampling operation is preserved by these skip connections. Consequently, the decoder may produce segmentation masks with clear object boundaries that are more precise.

For this study, U-Net was chosen for a number of reasons. First, it has demonstrated excellent performance across many organs and imaging modalities and has been frequently utilized for medical image segmentation. It is therefore a trustworthy starting point for assessing segmentation models. Second, it can combine several pre-trained encoder backbones without altering the decoder thanks to its encoder-decoder structure. Because of this, U-Net is a good choice for a fair comparison of various encoder architectures. Furthermore, as is frequently the case in medical image analysis, U-Net works effectively even in situations where the quantity of training data is quite little. Ultimately, the network's output is a pixel-wise probability map, which makes it appropriate for the binary kidney and tumor segmentation tasks taken into consideration in this study.

### 4.2.2 U-Net Architecture Details

The U-Net architecture consists of three main components: an encoder (contracting path), a bottleneck, and a decoder (expanding path), as illustrated in Figure 4.2.

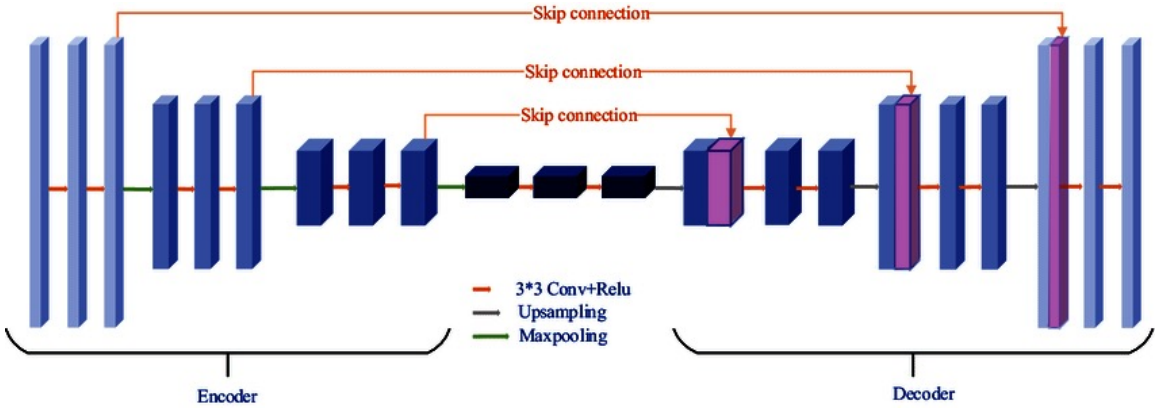


Figure 4.2: U-Net architecture used in this study. The encoder (left) progressively downsamples the input; the decoder (right) restores spatial resolution via transposed convolutions. Skip connections (horizontal arrows) fuse encoder features into each decoder level.

The **encoder path** gradually reduces the spatial resolution of the input image while increasing the number of feature channels. At each level, two 3×3 convolutional layers

with ReLU activation extract image features. A 2×2 max-pooling layer then reduces the spatial dimensions by half. As the network becomes deeper, it learns features at different levels, ranging from edges and textures in the early layers to more complex anatomical patterns in the deeper layers.

The **bottleneck** is the deepest part of the network. At this stage, the feature maps have the smallest spatial resolution and the largest number of channels. The bottleneck captures high-level contextual information from the entire input image, which helps the decoder reconstruct accurate segmentation masks.

The **decoder path** gradually restores the original spatial resolution using transposed convolution layers. At each level, the upsampled feature maps are concatenated with the corresponding feature maps from the encoder through skip connections. These skip connections combine high-level contextual information with fine spatial details, improving the accuracy of the reconstructed segmentation mask. After concatenation, two 3×3 convolutional layers with ReLU activation further refine the extracted features.

The **output layer** uses a 1×1 convolution to map the final feature representation to a single output channel for binary segmentation. The output is a pixel-wise probability map that represents the likelihood of each pixel belonging to the foreground class, which is the kidney in Stage 1 and the tumor in Stage 2.

Each of the five pre-trained backbone architectures listed in the next section is used in this investigation in place of the conventional U-Net encoder. To ensure that reported performance discrepancies are attributed to encoder design, the decoder architecture is maintained consistent across all encoder variants.

## 4.3 Pre-Trained Encoder Backbones

This study evaluates five encoder backbones. Hierarchical feature extraction is approached differently by each design. All five encoders are integrated into the same U-Net framework and trained under the same experimental conditions to guarantee a fair comparison. Weights pre-trained on the ImageNet dataset [18] are used to initialize all encoder backbones. Prior to fine-tuning on CT images, these pre-trained weights give the models general image properties including edges, textures, and basic forms. Compared to training the models from random initialization, this initialization frequently results in superior performance and increases training efficiency. This is especially beneficial given the relatively limited size of the medical imaging subset used in this study, where training from scratch could easily lead to overfitting. Each backbone also differs in depth, parameter count, and computational cost, which allows the study to examine the trade-off between model complexity and segmentation accuracy.

### 4.3.1 ResNet-18: Residual Networks

ResNet-18 is the shallowest member of the Residual Network family introduced by He et al. [18]. The defining contribution of the ResNet architecture is the residual connection,

which allows the network to learn residual functions relative to the layer input rather than unreferenced mappings. Formally, if the desired mapping for a block is $\mathcal{H}(\mathbf{x})$, the block is reformulated to learn $\mathcal{F}(\mathbf{x}) = \mathcal{H}(\mathbf{x}) - \mathbf{x}$, so that the full output becomes:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}) + \mathbf{x} \tag{4.1}$$

where $\mathbf{x}$ is the block input and $\mathcal{F}(\mathbf{x})$ is the learned residual function. This reformulation addresses the vanishing gradient problem by providing a direct gradient path through the shortcut connection.

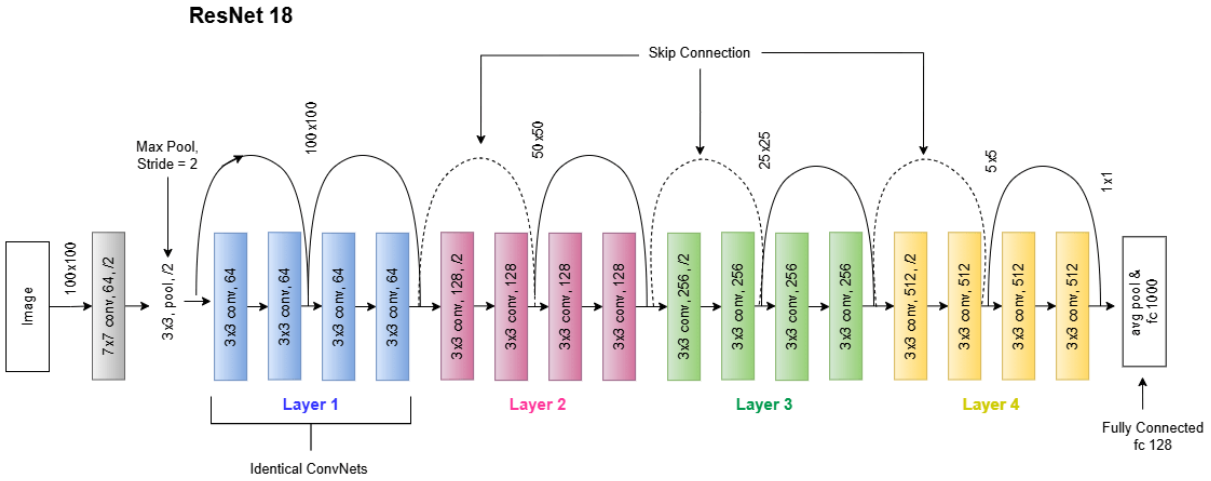


Figure 4.3: ResNet-18 Architecture.

ResNet-18 consists of 18 layers arranged into four stages of residual blocks. As the network becomes deeper, the spatial resolution of the feature maps decreases while the number of feature channels increases. Each residual block contains two 3×3 convolutional layers connected by a shortcut connection. ResNet-18 was selected to evaluate whether a relatively shallow residual network can learn features that are sufficient for both kidney and tumor segmentation. It also allows us to examine whether a lightweight architecture performs better on small region-of-interest (ROI) tasks, such as renal tumor segmentation.

### 4.3.2 ResNet-50: Deeper Residual Learning

ResNet-50 extends the residual learning framework with a deeper architecture using bottleneck residual blocks. Unlike the basic block in ResNet-18, which applies two 3×3 convolutions, the bottleneck block applies a 1×1 convolution to reduce channel dimensionality, a 3×3 convolution on the reduced-dimension representation, and a 1×1 convolution to restore channel dimensionality. This design enables greater effective depth while managing computational cost.

ResNet-50 was included to examine whether a deeper residual network can improve segmentation performance compared with ResNet-18. A deeper architecture may capture more complex and high-level image features, which could be beneficial for kidney segmentation, where understanding the overall organ shape and anatomical context is important. Its performance is also evaluated on the tumor segmentation task to determine whether the additional network depth provides similar benefits for localized regions of interest.

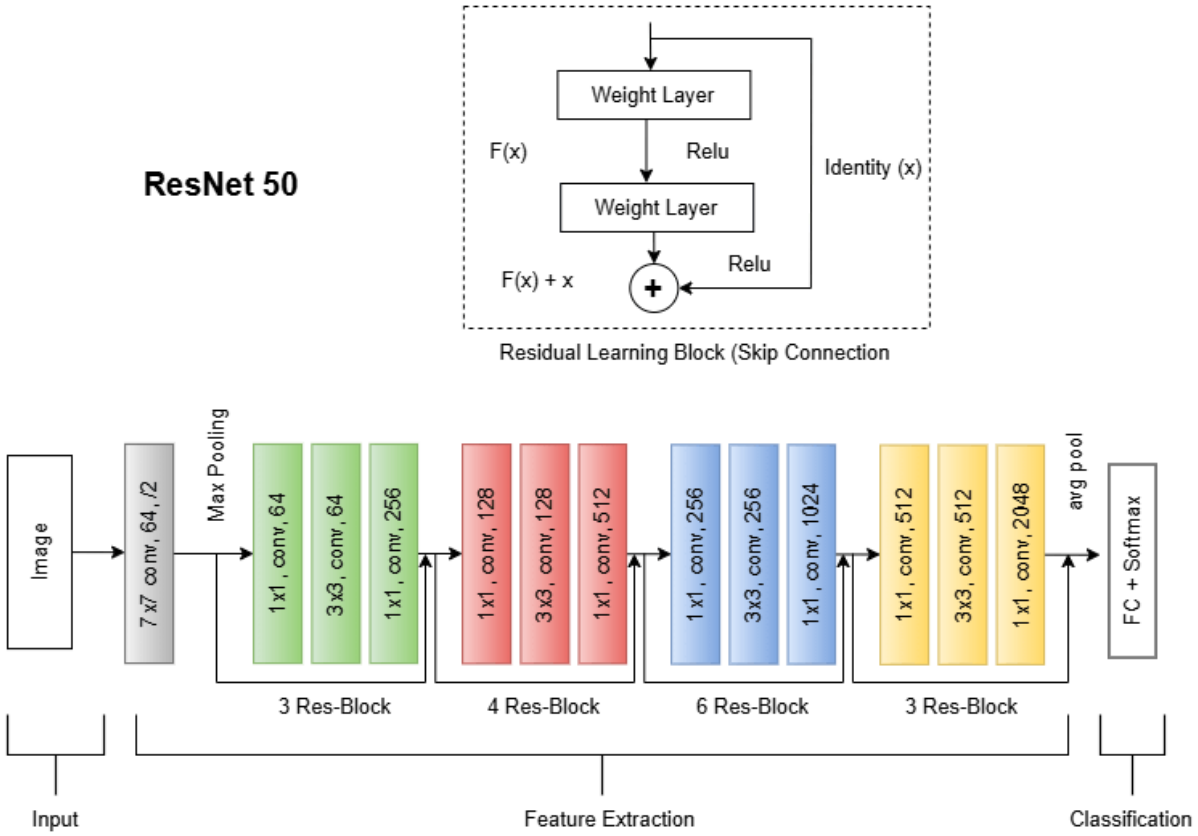


Figure 4.4: ResNet-50 Architecture.

### 4.3.3 DenseNet-121: Dense Connectivity

DenseNet-121, introduced by Huang et al. [46], replaces sequential layer stacking with a dense connection pattern in which every layer within a dense block receives input from all preceding layers. Formally, the output of layer $l$ is:

$$\mathbf{x}_l = \mathcal{H}_l([\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{l-1}]) \tag{4.2}$$

where $[\mathbf{x}_0, \ldots, \mathbf{x}_{l-1}]$ denotes the concatenation of feature maps from all previous layers and $\mathcal{H}_l$ is the composite function (batch normalization, ReLU, and 3×3 convolution) applied at layer $l$.

This dense connectivity improves the flow of information and gradients throughout the network while encouraging feature reuse. As a result, DenseNet-121 can achieve strong performance with fewer parameters than many comparable ResNet models. DenseNet-121 consists of 121 layers organized into four dense blocks separated by transition layers. Each layer adds 32 new feature channels, known as the growth rate, allowing the network to continuously build richer feature representations as it becomes deeper.

### 4.3.4 EfficientNet-B3: Compound Scaling

EfficientNet, introduced by Tan and Le [20], addresses how to scale a neural network architecture most effectively. Rather than scaling network depth, width, or input resolution

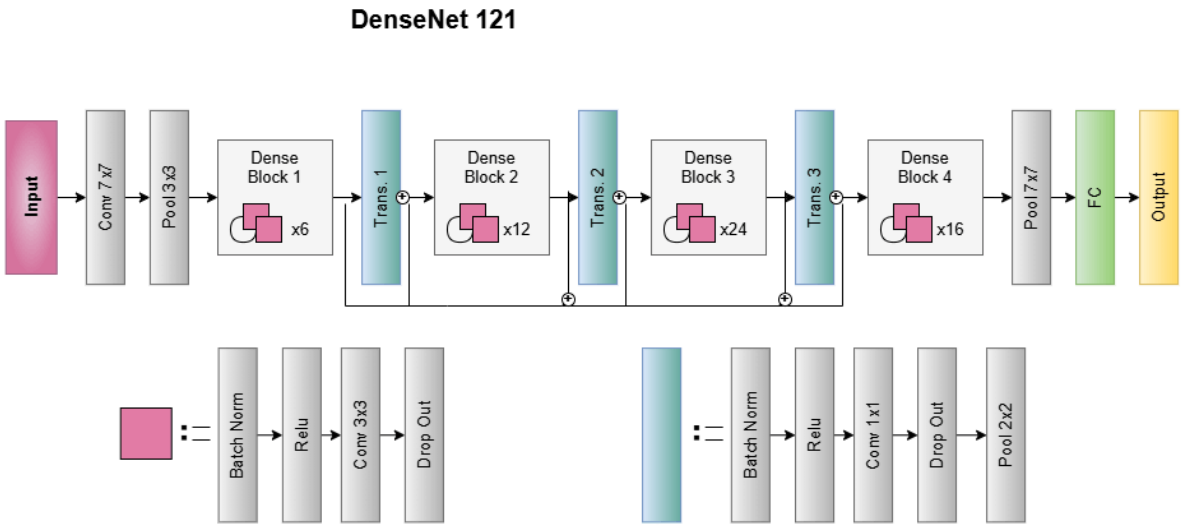


Figure 4.5: DenseNet-121 Architecture.

independently, EfficientNet introduces compound scaling, which scales all three dimensions simultaneously according to a fixed ratio governed by a scaling coefficient $\varphi$:

$$\text{depth} = \alpha^{\varphi}, \quad \text{width} = \beta^{\varphi}, \quad \text{resolution} = \gamma^{\varphi} \tag{4.3}$$

where $\alpha$, $\beta$, and $\gamma$ are constants determined through neural architecture search on the base EfficientNet-B0 model, and $\varphi$ is a user-specified coefficient.

EfficientNet-B3 is the third variant in the EfficientNet family ($\varphi = 3$).

EfficientNet-B3 is built using mobile inverted bottleneck (MBConv) blocks, depthwise separable convolutions, and squeeze-and-excitation modules. These components improve feature extraction while keeping the number of model parameters relatively low. The squeeze-and-excitation modules also help the network emphasize more informative feature channels during training. EfficientNet-B3 was included in this study to examine whether its compound scaling strategy improves the learning of multi-scale features required for kidney and tumor segmentation. Its performance is compared with the other encoder backbones under the same experimental conditions.

### 4.3.5 MobileNetV2: Lightweight Architecture

MobileNetV2, introduced by Sandler et al. [21], was designed specifically for deployment in resource-constrained environments. The primary architectural innovation is the inverted residual block, which first expands the input channel dimension by a factor of six using a 1×1 pointwise convolution, applies a depthwise 3×3 convolution on the expanded representation, and then projects back to a smaller output channel dimension using a linear 1×1 pointwise convolution. The use of a linear projection in the final step prevents the ReLU non-linearity from destroying information in the low-dimensional representation.

MobileNetV2 is designed for computational efficiency. It requires approximately 300 million multiply-accumulate (MAC) operations, making it much lighter than many conventional convolutional networks. It also has significantly fewer parameters and offers faster inference,

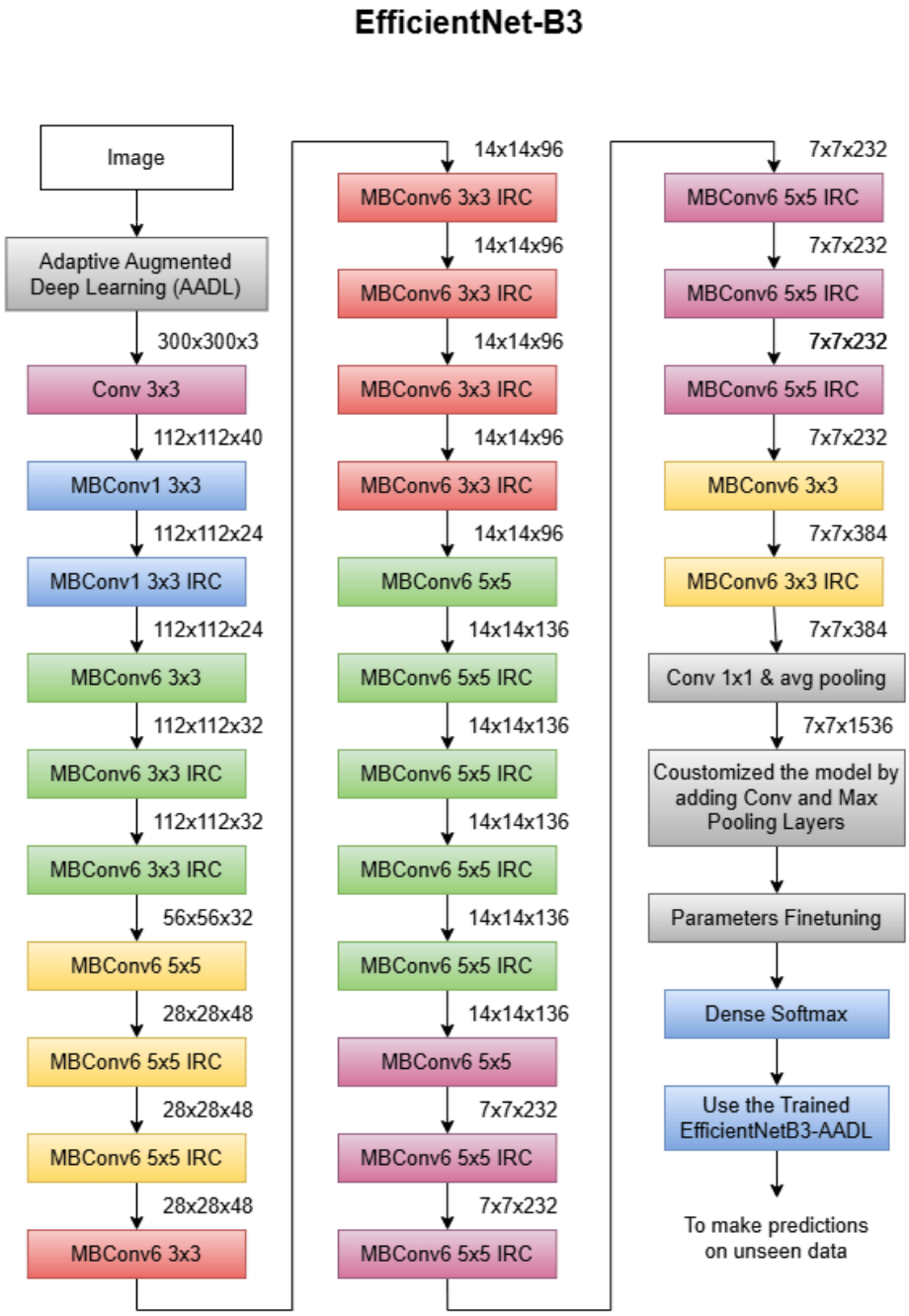


Figure 4.6: EfficientNet-B3 Architecture.

making it suitable for applications with limited computational resources. MobileNetV2 was included in this study to investigate whether a lightweight encoder can achieve segmentation performance comparable to larger architectures. This is particularly important for clinical settings, where fast inference and efficient resource utilization are practical considerations.

## 4.4 Training Protocol

All five encoder backbones are initialized with weights pre-trained on the ImageNet dataset before being fine-tuned on the KiTS23 segmentation tasks. ImageNet pre-training provides

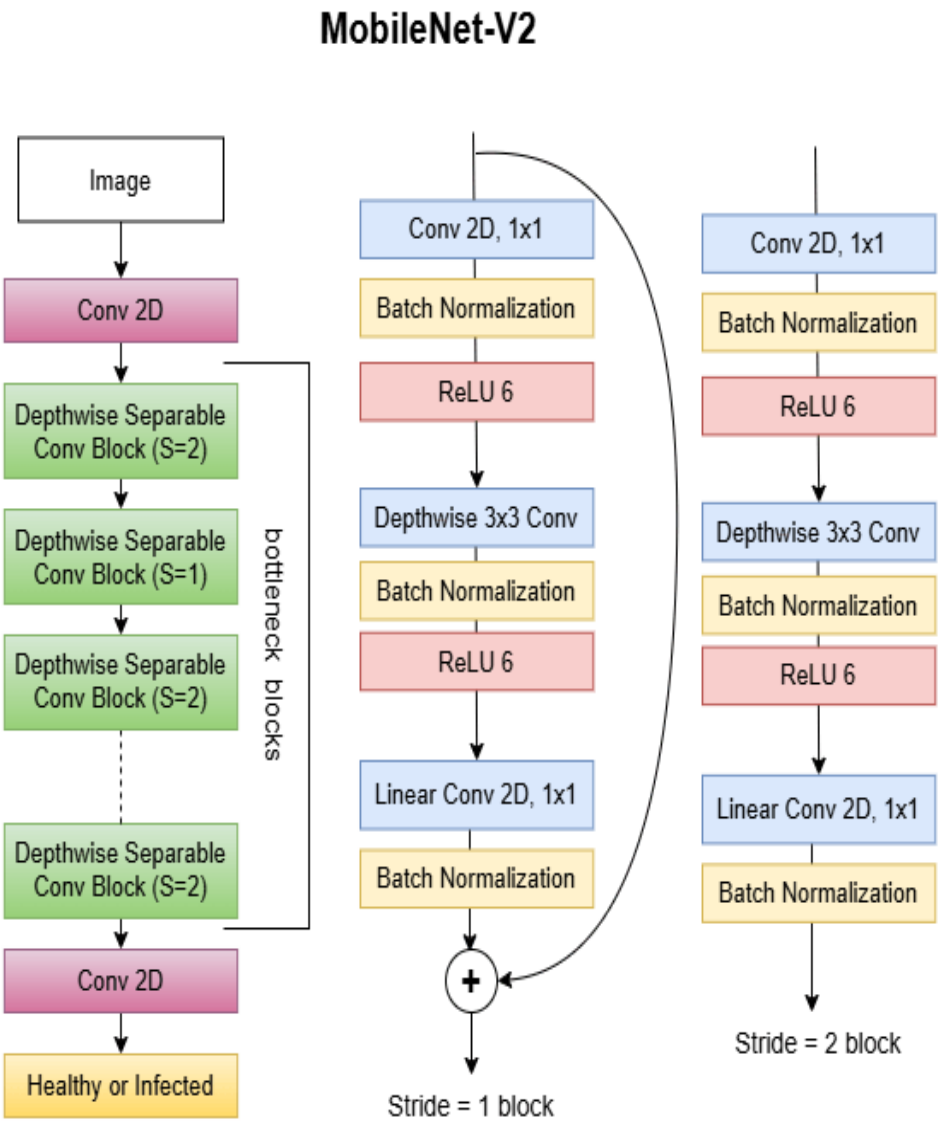


Figure 4.7: MobileNet-V2 Architecture.

general visual features, such as edges, textures, and simple shape patterns, which serve as a useful starting point even though natural images and CT images are different. Using pre-trained weights helps the models converge more quickly during training and often achieves better performance than training from random initialization, especially when the available medical imaging data are limited.

In contrast, the U-Net decoder is initialized from scratch because it is designed specifically for the segmentation task and does not have a suitable pre-trained counterpart. During training, both the encoder and decoder are optimized together. This allows the encoder to adapt its learned features to CT images while benefiting from the initialization provided by ImageNet pre-training.

All models in both stages of the pipeline are trained using identical hyperparameters, ensuring that observed performance differences are attributable to architectural rather than training configuration differences.

Training is performed using the Adam optimizer [47] with an initial learning rate of $2 \times 10^{-4}$

Table 4.1: Summary of Hyperparameters.

| **Parameter** | **Details** |
|---|---|
| Loss Function | Binary Dice Loss (both stages) |
| Optimizer | Adam, initial learning rate = $2 \times 10^{-4}$ |
| LR Scheduler | CosineAnnealingLR |
| Epochs / Batch Size | 100 epochs, batch size 4 |
| Hardware | NVIDIA RTX 4070 Super, 12 GB VRAM |
| Framework | PyTorch and PyTorch Lightning |

and a cosine annealing learning rate scheduler, which gradually reduces the learning rate over the course of training according to:

$$\mathrm{lr}(t) = \mathrm{lr}_{\min} + \tfrac{1}{2}(\mathrm{lr}_{\max} - \mathrm{lr}_{\min})\left(1 + \cos\left(\frac{\pi t}{T}\right)\right) \tag{4.4}$$

where $t$ is the current epoch and $T$ is the total number of epochs. Here, $\mathrm{lr}(t)$ represents the learning rate at epoch $t$, while $\mathrm{lr}_{\min}$ and $\mathrm{lr}_{\max}$ denote the minimum and maximum learning rate values, respectively.

This schedule allows the model to explore the loss landscape with a higher learning rate in early training and to fine-tune weights more precisely in later epochs.

All models are trained for 100 epochs with a batch size of 4. The training objective for both tasks is Binary Dice Loss, denoted as $\mathcal{L}_{\mathrm{Dice}}$.

Binary Dice Loss is used as the training objective for both tasks:

$$\mathcal{L}_{\mathrm{Dice}} = 1 - \frac{2\sum_i y_i \hat{y}_i}{\sum_i y_i + \sum_i \hat{y}_i} \tag{4.5}$$

where $y_i$ is the ground-truth binary label and $\hat{y}_i$ is the predicted probability at pixel $i$, summed over all pixels in the batch.

Dice Loss was chosen because it directly optimizes the overlap between predicted and ground truth regions and is more robust to class imbalance than cross-entropy loss.

Online data augmentation is applied during training, including random rotations, random scaling, random horizontal and vertical flips, intensity variations (brightness and contrast adjustments), and Gaussian blur.

These augmentations expose the model to a broader range of appearances and reduce the risk of overfitting to the specific acquisition characteristics of the training cases.

Input images are resized to $224 \times 224$ pixels for Stage 1 kidney segmentation.

Stage 2 tumor segmentation operates on kidney-guided ROI crops at variable sizes determined by the kidney bounding box.

All models are trained on an NVIDIA RTX 4070 Super GPU (12 GB VRAM) using PyTorch and PyTorch Lightning.

## 4.5 Evaluation Framework

Segmentation performance is evaluated using two standard metrics: Intersection over Union (IoU) and F1 Score (equivalent to the Dice Similarity Coefficient). Both metrics are computed at the pixel level by comparing predicted binary masks to expert ground truth masks.

### 4.5.1 Confusion Matrix

The evaluation is based on the following confusion matrix quantities:

- **True Positives (TP):** Pixels correctly predicted as foreground (kidney or tumor).
- **False Positives (FP):** Background pixels incorrectly predicted as foreground (over-segmentation).
- **False Negatives (FN):** Foreground pixels missed by the model (under-segmentation).
- **True Negatives (TN):** Background pixels correctly predicted as background.

True negatives (TN) are included in the confusion matrix, however they are not taken into account when calculating the F1 score or IoU. As noted in Chapter 1 and Section 2.2, this is particularly crucial for the kidney and tumor segmentation tasks taken into consideration in this study, since the foreground only takes up a small percentage of each CT slice, leading to a significant class imbalance. Because the majority of pixels are in the background class, metrics that incorporate TN, like pixel-wise accuracy, may yield values that are unduly optimistic. Even if a model is unable to accurately identify the kidney or tumor, it may still achieve high accuracy if it predicts nearly all pixels as background.

In contrast, IoU and the F1 score are calculated using only true positives (TP), false positives (FP), and false negatives (FN). By focusing on the overlap between the predicted and ground truth foreground regions, these metrics provide a more reliable measure of segmentation quality and are better suited for evaluating models under severe class imbalance.

### 4.5.2 IoU

Intersection over Union (IoU), also known as the Jaccard Index, measures the ratio of the intersection between the predicted and ground truth masks to their union:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{4.6}$$

IoU ranges from 0 (no overlap) to 1 (perfect overlap) and penalizes both false positives and false negatives equally. It is insensitive to class imbalance between foreground and background pixels. This makes it well suited for evaluating small structures such as

tumors and cysts. A higher IoU indicates that the predicted mask closely matches the true boundary of the target region. Even small boundary misalignments can noticeably lower the IoU score. For this reason, IoU is considered a strict measure of segmentation quality.

### 4.5.3 F1 Score

The F1 Score measures the harmonic mean of precision and recall, and is mathematically equivalent to the Dice Similarity Coefficient (DSC):

$$\mathrm{F1} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{4.7}$$

The F1 Score weights the intersection term twice in the numerator. This makes it slightly more tolerant of small boundary errors than IoU. As a result, F1 values are typically higher than IoU values for the same prediction. The F1 Score is one of the most widely used metrics in medical image segmentation. It balances the trade-off between over-segmentation and under-segmentation. A high F1 Score indicates that the model reliably captures the shape and extent of the target region.

Both metrics are widely reported in the medical image segmentation literature and are computed and reported here to enable direct comparison with prior work.

# Chapter-5: Results and Analysis

This chapter presents the quantitative evaluation outcomes of the five encoder-based U-Net models across both stages of the hierarchical segmentation pipeline.

## 5.1 Kidney Segmentation

All five encoder variants were trained and evaluated on the binary kidney segmentation task based on the KiTS23 training, validation and test splits as described in Chapter 3. Quantitative results on the test set are shown in Table 5.1.

Table 5.1: Performance Metrics for Kidney Segmentation (Stage 1).

| **Encoder** | **IoU** | **F1 Score** |
|---|---|---|
| ResNet-18 | 0.8946 | 0.9444 |
| ResNet-50 | 0.9054 | 0.9503 |
| MobileNetV2 | 0.8914 | 0.9426 |
| EfficientNet-B3 | **0.9122** | **0.9541** |
| DenseNet-121 | 0.9008 | 0.9478 |

All five encoders achieved robust kidney segmentation results with F1 scores above 0.94 in all cases. The improved baseline performance on all architectures points to morphological consistency of the kidney as a segmentation target and validates the efficacy of the preprocessing pipeline in generating standardized and informative training inputs.

EfficientNet-B3 obtained better results on both metrics, F1 score of 0.9541 and IoU of 0.9122. The second rank was obtained by ResNet-50 with F1 score of 0.9503 and IoU of 0.9054. DenseNet-121 achieved the third rank with F1 score of 0.9478 and IoU of 0.9008. The ResNet-18 and MobileNetV2 were ranked 4th and 5th, respectively, with a small performance gap between the encoders. The F1 score range for all five encoders is only 0.0115 (from 0.9426 to 0.9541).

The small performance variance shows that all five architectures are capable of effectively learning the essential feature to accurately locate the kidney.

The better performance of ResNet-50 (second best) compared to ResNet-18 (fourth best) suggests that depth in the network is beneficial for kidney segmentation, possibly because it is more capable of representing global shape context. The fact that MobileNetV2 performs

competitively with a much smaller number of parameters demonstrates that computational efficiency and segmentation accuracy are not mutually exclusive for this task.

Comparative analysis of representative ground truth and predictions for kidney segmentation in Figure 5.1.

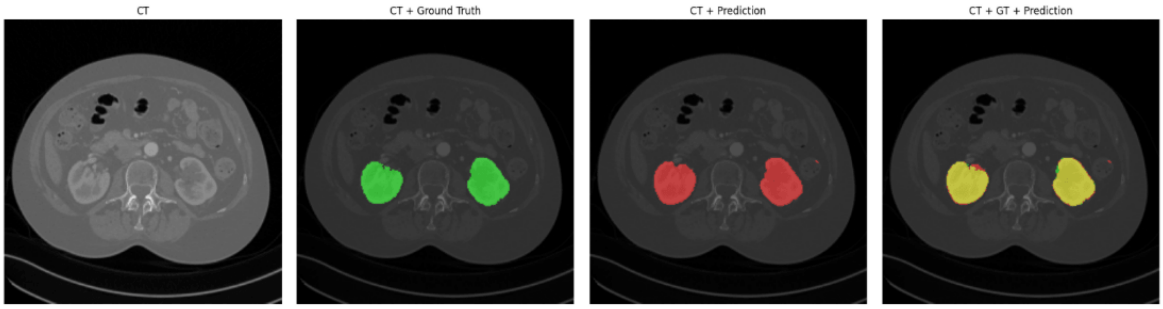


Figure 5.1: Ground truth and predicted kidney segmentation masks for representative test cases. From left - input CT slice, expert ground truth mask, ground truth overlaid on CT (green), model prediction(red), and prediction overlaid on CT (yellow).

## 5.2 Tumor Segmentation

Absolute performance metrics are worse for all encoders indicating that tumor segmentation is significantly more challenging than kidney segmentation. Quantitative results are presented in Table 5.2.

Table 5.2: Performance Metrics for Tumor Segmentation (Stage 2).

| **Encoder** | **IoU** | **F1 Score** |
|---|---|---|
| ResNet-18 | **0.5619** | **0.7195** |
| ResNet-50 | 0.4394 | 0.6106 |
| MobileNetV2 | 0.4612 | 0.6313 |
| EfficientNet-B3 | 0.4835 | 0.6518 |
| DenseNet-121 | 0.4298 | 0.6012 |

ResNet-18 achieved the superior tumor segmentation performance with an F1 score of 0.7195 and IoU of 0.5619, significantly outperforming the second-best with a difference of 0.0677 in F1 score. EfficientNet-B3 obtained the second highest F1 score with 0.6518, followed by MobileNetV2 with 0.6313, ResNet-50 with 0.6106 and DenseNet-121 with 0.6012.

Stage 2 is quite different in the order of performance compared to Stage 1. Regarding tumor segmentation, EfficientNet-B3, best performer in kidney segmentation, is second. Interestingly, ResNet-18, which is ranked fourth for kidney segmentation, is the clear winner for tumor segmentation. Although ResNet-50 performs better than ResNet-18 in Stage 1, it has a much lower F1 score (0.6106) compared to ResNet-18 (0.7195) in Stage 2, indicating that the depth and parameters of ResNet-50 are detrimental to the tumor detection task.

The markedly lower absolute performance metrics in Stage 2 compared to Stage 1 – best F1 score dropping from 0.9541 to 0.7195 – underscore the intrinsic difficulties of renal tumor segmentation even within the context of kidney-guided regions of interest.

A major challenge for learning algorithms is the intra-class variability of tumors in size, shape and CT attenuation.

Some example comparisons of ground truth vs. prediction for tumor segmentation are shown in Figure 5.2.

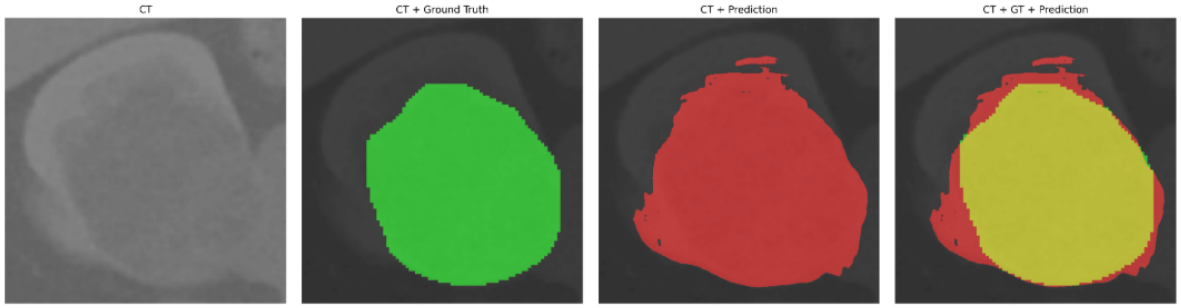


Figure 5.2: Ground truth and predicted tumor segmentation masks for representative test cases. From left - kidney-guided ROI crop, expert ground truth tumor mask, ground truth overlaid on CT (green), model prediction (red), and prediction overlaid on CT (yellow).

## 5.3 Training Curves

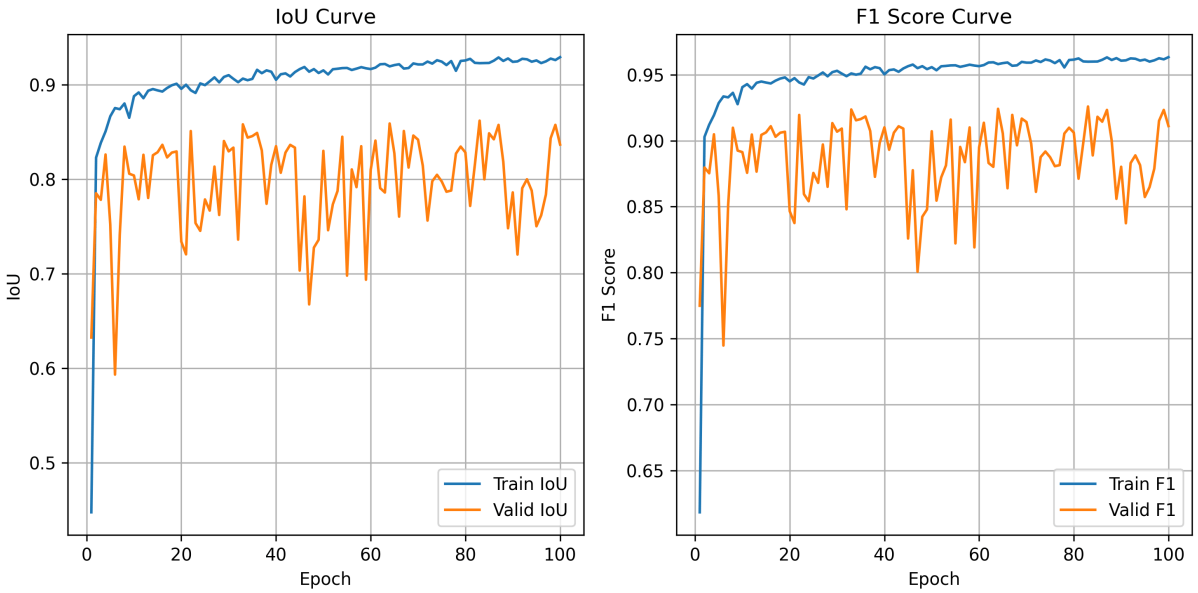


Figure 5.3: Training and validation IoU and F1 score curves for kidney segmentation using EfficientNet-B3 over 100 epochs.

Figure 5.3 shows the training and validation curves for kidney segmentation using EfficientNet-B3. Both IoU and F1 score rise steadily over the 100 epochs. Training IoU reaches around 0.93 and training F1 reaches around 0.96 by the last epoch. The validation curves follow the same upward pattern but fluctuate more between epochs. Most of the

later validation F1 values fall between 0.85 and 0.92. The gap between training and validation stays small throughout, which shows the model is not overfitting badly.

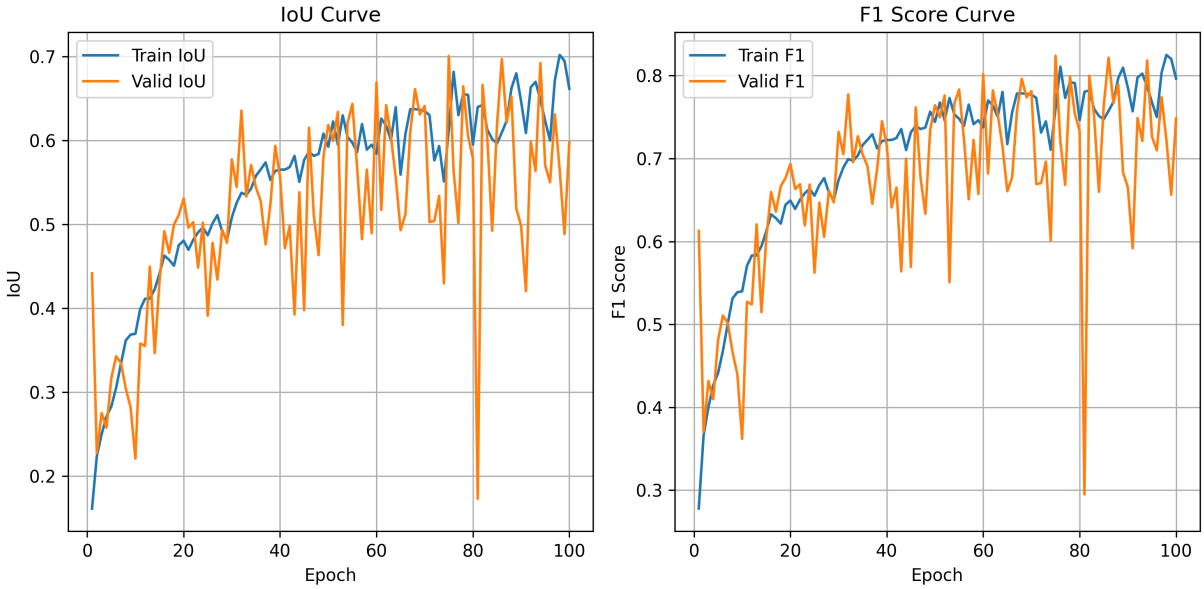


Figure 5.4: Training and validation IoU and F1 score curves for tumor segmentation using ResNet-18 over 100 epochs.

Figure 5.4 shows the training and validation curves for tumor segmentation using ResNet-18. Both metrics climb more slowly here, and the curves are noisier than in kidney segmentation. Training IoU reaches around 0.70 and training F1 reaches around 0.82 by epoch 100. The validation curves swing widely from epoch to epoch. There is a sharp drop near epoch 81, where validation IoU and F1 fall to their lowest points in the whole run before recovering. This instability shows how sensitive tumor segmentation is to a handful of hard cases in the validation set. Even with this noise, the overall trend across epochs still moves upward.

## 5.4 Performance Summary and Overall Trends

Table 5.3 presents the F1 score results for both stages side by side to facilitate direct comparison of encoder behaviour across tasks.

Table 5.3: F1 Score Performance Summary Across Both Stages.

| **Encoder** | **Stage 1 (Kidney) F1** | **Stage 2 (Tumor) F1** |
|---|---|---|
| ResNet-18 | 0.9444 | **0.7195** |
| ResNet-50 | 0.9503 | 0.6106 |
| MobileNetV2 | 0.9426 | 0.6313 |
| EfficientNet-B3 | **0.9541** | 0.6518 |
| DenseNet-121 | 0.9478 | 0.6012 |

The most interesting pattern in this summary is the total reversal of the performance pecking order between the two stages. EfficientNet-B3 is the best encoder for kidney

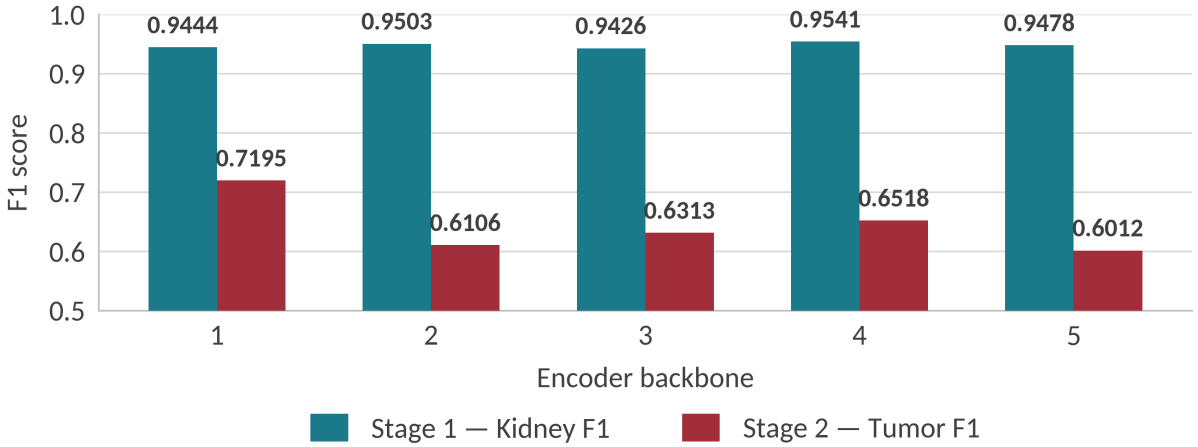


Figure 5.5: Performance Summary Across Both Stages (F1 Score by encoder and stage).

segmentation and the second best for tumor segmentation. ResNet-18 is the worst encoder for kidney segmentation but the best for tumor segmentation. Within the evaluated encoder backbones and the KiTS23 dataset, no single encoder achieved the best performance in both stages. This suggests that the hierarchical pipeline may benefit from selecting task-specific encoders rather than relying on one architecture for both stages.

Furthermore, the performance variation across encoders is much larger for Stage 2 (F1 range: 0.6012–0.7195, range size: 0.1183) than for Stage 1 (F1 range: 0.9426–0.9541, range size: 0.0115). The difference in encoder performance is an order of magnitude indicating that the tumor segmentation task is especially sensitive to architectural choices in the encoder. When the task is relatively simple and all encoders are close to optimal performance, architectural differences are reduced. In very hard problems, like tumor segmentation, architectural differences appear to play a more prominent role in performance, although other factors such as training data and hyperparameters may also contribute.

# Chapter-6: Discussion

## 6.1 Task-Dependent Architectural Performance

These results directly answer the core research question posed in Chapter 1: encoder-backbone optimality is not consistent across both stages of the proposed hierarchical segmentation pipeline. EfficientNet-B3 was the best-performing encoder for kidney segmentation, with an F1 score of 0.9541, yet it placed only second in tumor segmentation. ResNet-18 tells the opposite story: the weakest encoder for kidney segmentation, but the best for tumor segmentation, with an F1 score of 0.7195, beating the next-best encoder by a margin of 0.0677. No single encoder came out on top in both stages, confirming that encoder-backbone optimality is fundamentally stage-dependent rather than architecture-dependent.

This task dependency arises from structural differences between the two segmentation targets. The kidney parenchyma is a morphologically consistent, anatomically robust structure, well suited to architectures capable of capturing multi-scale global shape features. Renal tumors, by contrast, are small, heterogeneous, and localised, requiring precise detection within a narrow region of interest, a setting where less complex, parameter-light architectures generalise better.

This has a direct practical consequence: a hierarchical pipeline should assign a different, task-specific encoder to each stage rather than reusing one architecture throughout. Using EfficientNet-B3, the best kidney encoder, for tumor segmentation would perform considerably worse than using ResNet-18. A single architecture optimised for both tasks would compromise performance on each. The comparative results in Chapter 5 provide clear, evidence-based guidance for this configuration decision.

## 6.2 Compound Scaling for Morphologically Consistent Structures

That the EfficientNet-B3 was superior for kidney segmentation is consistent with the theoretical properties of its compound scaling architecture. Compound scaling achieves a balanced architecture by jointly scaling network depth, width and input resolution according to a pre-defined formula, which enables the model to learn features of different levels of granularity. The ability to extract multi-scale features is directly beneficial for

kidney segmentation, as the model needs to learn to represent a structure that varies in size from patient to patient, appears at different locations within the image field of view and has a consistent but not identical morphology.

The compound scaling principle ensures that the network does not over-specialise in any feature dimension. However, depth scaling only in the network can achieve increasingly abstract semantic features at the price of losing the fine-grained spatial details. A width-only network could increase the feature diversity at a certain resolution, but it might not have the depth for high-level semantic understanding. A resolution-scaled network can learn fine details at a higher computational cost. The balanced scaling of EfficientNet-B3 does not suffer from these problems and results in representations that are spatially detailed, semantically rich and multi-scale, properties that fit well with the requirements of kidney parenchyma segmentation.

The inferior performance of EfficientNet-B3 in tumor segmentation, despite its multi-scale feature extraction capabilities, might imply another limitation: the morphological variability and small size of tumors might lead the compound-scaled architecture to excessively distribute its representational capacity among scales, instead of focusing it on the particular spatial scales pertinent to small tumors.

## 6.3 Residual Connections for Localized Tumor Detection

The excellent results achieved by ResNet-18 in tumor segmentation, in addition to its advantage over the deeper ResNet-50, provides useful insight into the requirements for small lesion detection in limited regions of interest. The shallow architecture of ResNet-18 (18 layers vs. 50 in ResNet-50) and its residual connections seem to provide a useful inductive bias for the tumor segmentation task.

This advantage can be accounted for by several mechanisms. Initially, ResNet-18 has a smaller number of parameters compared to ResNet-50, DenseNet-121 and EfficientNet-B3, so it is less likely to overfit on the tumor segmentation training dataset. Tumor annotations are inherently noisy at the boundary level, and a model with fewer parameters can learn to capture the overall tumor region without fitting the specific boundary irregularities that characterise individual training examples.

ResNet-18's more superficial receptive field may be more accurately calibrated to the spatial scale of renal tumors in the region of interest of the kidney. Deeper networks have larger effective receptive fields, which make them more suitable for understanding global shapes, but potentially harmful when the target structure is small and localised.

The simplicity of the residual learning process, which combines a learned residual function with the identity mapping, may allow successful gradient propagation without the representational complexity that might cause more complex architectures to learn spurious correlations in the training data.

The poor performance of ResNet-50 compared to ResNet-18 in Stage 2 (F1: 0.6106 vs.

0.7195) is a clear evidence towards the hypothesis that more depth is not beneficial and could be harmful for tumor detection in small ROIs. This result is important and counterintuitive: more depth and more parameters, which are usually good for classification benchmarks, can be bad for small-lesion segmentation tasks, where the main failure modes are overfitting and receptive field mismatch.

## 6.4 Dense Connectivity and Feature Reuse

The lowest tumor segmentation F1 score (0.6012) was provided by DenseNet-121 among the five encoders but a competitive kidney segmentation performance (0.9478 F1). This pattern suggests that the dense connectivity architecture, while beneficial for gradient flow and feature reuse in large, morphologically consistent structures, may have some limitations for detecting small lesions.

One possible explanation is that dense connectivity creates larger and larger concatenated feature representations that increase linearly with network depth by connecting each layer to all subsequent layers. The feature dimension in any layer of a compact block equals the sum of the output channels of all previous layers plus the input channels. This can lead to very high dimensional feature representations in deep dense blocks which may hinder effective routing through the U-Net decoder for the accurate localisation needed in small-lesion segmentation. The decoder has to learn to extract relevant tumor-discriminating features from an increasing concatenated feature space, which may be detrimental when the training set is moderate.

Moreover, the dense connectivity, where each layer is allowed to access the raw input features and all the intermediate representations, may promote the network to leverage low-level intensity features that are specific to the acquisition characteristics of the training data. This might lead to less robust representations compared to architectures that progressively abstract from low-level features, which is crucial in tumor segmentation where generalisation ability across different tumor appearances is essential.

This difference indicates that feature reuse alone does not necessarily improve segmentation performance. In the case of small and heterogeneous tumors, retaining a large number of features throughout the network may make it more difficult to identify the features that are most relevant to the lesion. The results therefore suggest that the suitability of an encoder depends on the segmentation target, and that an architecture that performs well for the kidney does not necessarily provide the most effective feature representation for localized tumor segmentation.

## 6.5 Lightweight Architectures and Computational Efficiency

The results of MobileNetV2 in both stages show that the performance of lightweight models can be similar to larger models with substantially lower computational cost.

This competitive performance profile is of particular importance from a deployment point of view. MobileNetV2 introduces inverted residual blocks with depthwise separable convolutions to provide about 8 to 9 times parameter efficiency over regular convolutional networks of similar depth. This computational efficiency is important for practical deployment viability in clinical setting when inference is performed on standard workstations without dedicated high performance GPU.

The observation that MobileNetV2 outperforms both ResNet-50 and DenseNet-121 in tumor segmentation provides further support for the hypothesis that smaller, more efficient architectures may generalise better to the more challenging problem of small-lesion detection, where the risk of overfitting is higher and the benefits of larger model capacities are reduced.

## 6.6 Two-Stage Hierarchical Design Benefits

The pipeline addresses the class imbalance problem in the tumor segmentation training dataset by restricting the tumor segmentation model to kidney-guided regions of interest derived from ground-truth kidney masks. The ratio of tumor pixels to background pixels is much higher in a kidney ROI than in an entire CT slice of the abdomen, which allows the Dice Loss to more effectively guide the network to learn features that differentiate tumors. The ROI extraction removes a lot of the anatomically irrelevant background tissue which could otherwise lead to false positive tumor predictions and thus improves the specificity of the tumor segmentation model.

These benefits are reflected in the results in Stage 2 where all encoders achieve F1 scores above 0.60 even for the challenging tumor segmentation task. We expect that incorporating 3D context, larger training datasets, and loss functions designed specifically for infrequent foreground regions in future work will further improve these results.

# Chapter-7: Conclusion

## 7.1 Summary

This report presents a two-level hierarchical segmentation framework for automated segmentation of kidneys and renal tumors using the KiTS23 dataset. The pipeline decomposes multi-class segmentation into two sequential binary sub-tasks. Stage 1 performs kidney localisation. Stage 2 performs tumor segmentation within kidney-centered regions of interest. A uniform preprocessing pipeline was used across both stages. This included Hounsfield Unit normalisation, canonical orientation alignment, morphological hole filling, and kidney-guided ROI extraction via connected component analysis.

Five pre-trained encoder backbones were evaluated within the U-Net architecture. These were ResNet-18, ResNet-50, DenseNet-121, EfficientNet-B3, and MobileNetV2. Stage 1 kidney segmentation performed very well across all encoders, with F1 scores above 0.94. This strong performance is due to the homogeneous structure of the renal parenchyma and the standardised preprocessing steps. EfficientNet-B3 achieved the best kidney segmentation result, with an F1 score of 0.9541 and an IoU of 0.9122. This is attributed to its compound scaling of depth, width, and resolution.

Tumor segmentation was the main challenge in this pipeline. Renal tumors vary substantially in size, shape, and CT attenuation compared to the surrounding parenchyma. This caused Stage 2 performance to be noticeably lower than Stage 1 across all encoders. The lighter ResNet-18 encoder achieved the best tumor segmentation result, with an F1 score of 0.7195 and an IoU of 0.5619. This suggests that lightweight residual architectures are better suited to extracting localised tumor features. The two-stage design helped reduce class imbalance and false positives from surrounding structures, improving detection of small tumor lesions.

Preprocessing had a significant impact on performance at both stages. HU normalisation and morphological hole-filling improved mask consistency during training. The kidney-guided ROI extraction in Stage 2 reduced background interference and kept tumor segmentation focused on anatomically relevant regions. Overall, these steps improved the repeatability and clinical relevance of the pipeline.

These results show that a well-designed 2D hierarchical pipeline can produce clinically relevant segmentation outcomes. Task-specific preprocessing and appropriate encoder selection at each stage played a key role in this outcome. The framework offers a reproducible and computationally efficient baseline for future comparison against 3D and more

complex architectures.

## 7.2 Limitations

There are several important limitations to this study that limit the generalisability of the findings and point to future research.

The study uses only 100 cases from the KiTS23 dataset, which is sufficient for a proof-of-concept evaluation, but is a small sample of the full dataset. The limited sample size limits the diversity of CT acquisition protocols, patient demographics and tumor morphologies in the training data which may impede generalisability of the reported performance to broader clinical populations. Furthermore, the dataset split used in this study was performed using the official KiTS23 split, but the fact that model selection may be sensitive to validation performance on this very split is a methodological aspect to be considered when interpreting the results.

Second, the pipeline operates on CT volumes as individual 2D axial slices without taking into account the spatial context between slices along the Z-axis. The incremental approach may result in inconsistencies between consecutive predictions and reduce the accuracy of 3D reconstruction from the predicted masks. In comparison to fully three-dimensional methods, 2D methods have many practical advantages such as lower memory consumption, faster training and easier deployment, but the main drawback is the lack of volumetric context.

Even in the kidney-centered regions of interest used in Stage 2, the imbalance of classes between tumor and background pixels is still an important issue. While the two-stage strategy mitigates the global imbalance of full CT volumes, the local imbalance of cropped kidney regions persists, impacting training stability and detection to small, low-contrast tumors. This is particularly relevant in poorly differentiated tumors from the surrounding renal tissue.

Finally, all encoders in this work were pre-trained with weights from ImageNet, a natural image dataset that has very different statistical properties than CT medical images. Although transfer learning from ImageNet provides useful low-level feature representations, the domain gap between natural images and CT images may impair the discriminative power of the learned features for detailed medical structures. This baseline analysis does not take into account the study of encoders pre-trained on datasets of medical imaging or domain adapted by further pre-training.

## 7.3 Social Impact Analysis

Bringing this hierarchical segmentation pipeline into clinical use would have effects that go beyond raw technical performance. Automated kidney and tumor segmentation can take over much of the manual annotation work currently done by radiologists. This frees up time for diagnostic reasoning and patient care, which matters most in high-volume

hospitals. Because the framework produces standardized volumetric measurements, it may also reduce the variability seen between different clinicians reading the same scan. Over time, this could lead to more consistent treatment planning and follow-up across institutions [7].

The two-dimensional, lightweight design of this pipeline also matters for healthcare equity. It does not require the heavy computational resources that three-dimensional models typically need. This makes it easier to deploy in resource-limited settings, such as regional or rural hospitals without dedicated GPU infrastructure. In practice, this could help close some of the gap in diagnostic capability between large tertiary centers and smaller, less-resourced facilities.

That said, automated tools like this one bring their own social risks. Clinicians may start relying on the model's output too heavily, even in cases where the prediction is wrong. This kind of automation bias is a real concern, and so is the possibility that trainees lose some of their manual segmentation and interpretation skills over time. Tumor segmentation performance in this study was noticeably weaker than kidney segmentation, with F1 scores ranging from 0.60 to 0.72. Given this gap, it is important that tools like this be treated as decision-support aids rather than replacements for expert radiological judgment.

## 7.4 Environmental Impact Analysis and Sustainability

Training and running deep learning models is not free of environmental cost. GPU computation consumes a lot of electricity, and this translates into real carbon emissions [48]. This is especially worth thinking about in medical imaging research, where three-dimensional architectures keep growing in size and hyperparameter searches often involve running the same experiment dozens of times.

The design of this report pushes back against that trend to some extent. Instead of a fully three-dimensional approach, a two-dimensional slice-based hierarchical pipeline was used. This cuts down the computational and energy demands of both training and inference by a meaningful margin. Every experiment here ran on a single consumer-grade NVIDIA RTX 4070 Super GPU with 12 GB of VRAM. Compare this to the multi-GPU clusters that volumetric models like 3D U-Net or nnU-Net usually need, and the difference in energy footprint per experiment is significant.

The encoder comparison in this report also has something to say about sustainability. MobileNetV2, the lightest of the five encoders tested, held its own in both segmentation stages. It did this with far fewer parameters and multiply-accumulate operations than ResNet-50, DenseNet-121, or EfficientNet-B3. This lines up with the broader "Green AI" idea that efficiency deserves as much attention as accuracy when choosing a model for deployment [49]. Lightweight architectures like this one lower both the energy cost of running inference at scale and the hardware barrier to clinical adoption. Going forward, work that extends this pipeline to larger datasets or full 3D architectures should report computational cost and estimated energy use alongside accuracy. That way, sustainability can be weighed against performance gains directly, rather than as an afterthought.

## 7.5 Ethical Considerations

A few ethical dimensions are worth addressing here. First, the KiTS23 dataset used in this study is made up of de-identified, publicly released imaging data. It was collected under institutional ethical approval and patient consent procedures set up by the original challenge organizers [1]. No additional patient-identifiable information was accessed or processed at any point in this report.

Second, this study used a limited sample of 100 cases out of the more than 300 available in the full dataset. This raises concerns about representativeness and algorithmic bias. The training data may not fully capture the diversity of patient demographics, scanner vendors, contrast protocols, and tumor subtypes found in real clinical populations. Because of this, the reported performance may not generalize equally well across all patient groups. If a model like this were deployed without broader validation, it could risk providing unequal diagnostic benefit across populations. Groups underrepresented in the training data would likely be affected the most [50].

Third, deep learning models are often criticized for being difficult to interpret. In a clinical setting, this matters a lot. A segmentation error, such as a missed tumor or an under-segmented one, could have real consequences for how a patient is managed. This is exactly why a human-in-the-loop workflow is not optional but necessary. Model outputs should be treated as decision support, something a radiologist reviews and confirms, not as a final diagnostic conclusion on its own. This aligns with current guidance on trustworthy AI in healthcare [51]. Finally, before this pipeline could ever be used in real clinical practice, it would need to go through formal regulatory evaluation, such as medical device certification, along with prospective validation across multiple centers. Only then could it ethically be used to inform patient care.

## 7.6 Future Work

This basis opens itself up to a number of interesting avenues of future research. An immediate next step is to extend the training to the entire KiTS23 dataset, including the full spectrum of clinical cases, contrast phases and acquisition protocols. By expanding the training data, more empirical evidence on the generalisability of the two-stage pipeline will be provided and a more reliable comparison with published three-dimensional methods on the same benchmark will be possible.

Another important focus is the investigation of 2.5D and 3D segmentation approaches. Providing the multi-channel input with neighbouring axial slices within a 2.5D framework would provide spatial context across slices, and would relax the full memory cost of volumetric convolution. Fully 3D architectures, e.g. 3D U-Net, would improve volumetric coherence along the Z-axis and may improve the consistency of kidney boundary delineation and tumor detection across slices.

The class imbalance problem in tumor segmentation needs dedicated research. Specialised loss functions like focal loss or compound loss, Dice-cross-entropy formulations and tar-

geted data augmentation techniques that over-sample infrequent tumor morphologies may improve model sensitivity to small and low-contrast lesions. Methods of curriculum learning that progressively increase the difficulty of the training samples could help to improve the stability of the convergence.

Pre-training the encoder on medical imaging data is a potentially useful direction. The foundation models that are trained on large CT datasets, or the encoders that are fine-tuned on abdominal CT volumes, may provide more discriminative features for renal structures than those initialised with ImageNet weights. Self-supervised pre-training with unlabelled CT data could be a scalable strategy to improve the encoder features with no requirement of large-scale manual annotation.

Finally, rigorous multi-center validation across varying CT acquisition protocols, scanner manufacturers, and patient demographics is paramount to advance the pipeline toward clinical utility. The final step for practical implementation as a radiologist-assistance system, would be integration into a prototype clinical decision-support tool, with sufficient explainability mechanisms and compliance with regulatory standards.

Despite these limitations, the main goal of this report is successfully achieved: to provide a rigorous, reproducible baseline analysis outlining the strengths and weaknesses of structured 2D methods. The findings here lay a strong foundation for systematic development of more complex architectures, training methods and clinical validation efforts.

# Bibliography


[1] N. Heller, N. Sathianathen, A. Kalapara, E. Walczak, K. Moore, H. Kaluzniak, J. Rosenberg, P. Blake, Z. Rengel, M. Oestreich *et al.*, "The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes," *arXiv preprint arXiv:1904.00445*, 2019.

[2] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[3] U. Capitanio and F. Montorsi, "Renal cancer," *The Lancet*, vol. 387, no. 10021, pp. 894–906, 2016.

[4] American Cancer Society, "Survival rates for kidney cancer," https://www.cancer.org/cancer/types/kidney-cancer/detection-diagnosis-staging/survival-rates.html, 2026, accessed: Aug. 4, 2026.

[5] M. S. Davenport, E. M. Hu, A. D. Smith, H. Chandarana, K. Hafez, G. S. Palapattu, J. Stuart Wolf Jr, S. G. Silverman, and S. of Abdominal Radiology Disease Focused Panel on Renal Cell Carcinoma Davenport Matthew Chandarana Hersh Silverman Stuart Israel Gary Leyendecker John Vikram Raghu Raman Steve Remer Erick Smith Andrew Wang Jane, "Reporting standards for the imaging-based diagnosis of renal masses on ct and mri: a national survey of academic abdominal radiologists and urologists," *Abdominal Radiology*, vol. 42, no. 4, pp. 1229–1240, 2017.

[6] N. J. Sathianathen, N. Heller, R. Tejpaul, B. Stai, A. Kalapara, J. Rickman, J. Dean, M. Oestreich, P. Blake, H. Kaluzniak *et al.*, "Automatic segmentation of kidneys and kidney tumors: the kits19 international challenge," *Frontiers in Digital Health*, vol. 3, p. 797607, 2022.

[7] J.-G. Lee, S. Jun, Y.-W. Cho, H. Lee, G. B. Kim, J. B. Seo, and N. Kim, "Deep learning in medical imaging: general overview," *Korean journal of radiology*, vol. 18, no. 4, pp. 570–584, 2017.

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

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[10] N. Siddique, P. Sidike, C. Elkin, and V. Devabhaktuni, "U-net and its variants for medical image segmentation: theory and applications," *arXiv preprint arXiv:2011.01118*, 2020.

[11] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical image analysis*, vol. 63, p. 101693, 2020.

[12] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han *et al.*, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge," *Medical image analysis*, vol. 67, p. 101821, 2021.

[13] N. Yao, H. Hu, C. Han, J. Nan, Y. Li, and F. Zhu, "An automated two-stage approach to kidney and tumor segmentation in ct imaging," *Technology and Health Care*, vol. 32, no. 5, pp. 3279–3292, 2024.

[14] N. Karunanayake, L. Lu, H. Yang, P. Geng, O. Akin, H. Furberg, L. H. Schwartz, and B. Zhao, "Dual-stage ai model for enhanced ct imaging: precision segmentation of kidney and tumors," *Tomography*, vol. 11, no. 1, p. 3, 2025.

[15] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of digital imaging*, vol. 32, no. 4, pp. 582–596, 2019.

[16] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[17] G. Hidy, B. Bakos, and A. Lukács, "Enhancing pretraining efficiency for medical image segmentation via transferability metrics," *arXiv preprint arXiv:2410.18677*, 2024.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[20] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[21] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[22] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[23] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[24] H. R. Roth, H. Oda, Y. Hayashi, M. Oda, N. Shimizu, M. Fujiwara, K. Misawa, and K. Mori, "Hierarchical 3d fully convolutional networks for multi-organ segmentation," *arXiv preprint arXiv:1704.06382*, 2017.

[25] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, and P. F. Jaeger, "nnu-net revisited: A call for rigorous validation in 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 488–498.

[26] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang *et al.*, "Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, vol. 97, p. 103280, 2024.

[27] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *European conference on computer vision*. Springer, 2022, pp. 205–218.

[28] K.-H. Uhm, H. Cho, Z. Xu, S. Lim, S.-W. Jung, S.-H. Hong, and S.-J. Ko, "Exploring 3d u-net training configurations and post-processing strategies for the miccai 2023 kidney and tumor segmentation challenge," in *International Challenge on Kidney and Kidney Tumor Segmentation*. Springer, 2023, pp. 8–13.

[29] A. Myronenko, D. Yang, Y. He, and D. Xu, "Automated 3d segmentation of kidneys and tumors in miccai kits 2023 challenge," in *International Challenge on Kidney and Kidney Tumor Segmentation*. Springer, 2023, pp. 1–7.

[30] W. Zhao, D. Jiang, J. P. Queralta, and T. Westerlund, "Multi-scale supervised 3d u-net for kidneys and kidney tumor segmentation," *arXiv preprint arXiv:2004.08108*, 2020.

[31] R. SSRM *et al.*, "Multi-scale and spatial information extraction for kidney tumor segmentation: A contextual deformable attention and edge-enhanced u-net." *Journal of Imaging Informatics in Medicine*, vol. 37, no. 1, pp. 151–166, 2024.

[32] P. Sun, Z. Mo, F. Hu, F. Liu, T. Mo, Y. Zhang, and Z. Chen, "Kidney tumor segmentation based on fr2pattu-net model," *Frontiers in oncology*, vol. 12, p. 853281, 2022.

[33] A. Abdelrahman and S. Viriri, "Efficientnet family u-net models for deep learning semantic segmentation of kidney tumors on ct images," *Frontiers in computer science*, vol. 5, p. 1235622, 2023.

[34] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Computerized Medical Imaging and Graphics*, vol. 95, p. 102026, 2022.

[35] P. K. Rao, S. Chatterjee, M. Janardhan, K. Nagaraju, S. B. Khan, A. Almusharraf, and A. I. Alharbe, "Optimizing inference distribution for efficient kidney tumor segmentation using a unet-pwp deep-learning model with xai on ct scan images," *Diagnostics*, vol. 13, no. 20, p. 3244, 2023.

[36] V.-O. Bolocan, M. Secareanu, E. Sava, C. Medar, L. S. C. Manolescu, A.-Cătălin Rașcu, M. G. Costache, G. D. Radavoi, R.-A. Dobran, and V. Jinga, "Convolutional neural network model for segmentation and classification of clear cell renal cell carcinoma based on multiphase ct images," *Journal of Imaging*, vol. 9, no. 12, p. 280, 2023.

[37] R. Khan, C. Chen, A. Zaman, J. Wu, H. Mai, L. Su, Y. Kang, and B. Huang, "Renalsegnet: automated segmentation of renal tumor, veins, and arteries in contrast-enhanced ct scans," *Complex & Intelligent Systems*, vol. 11, no. 2, p. 131, 2025.

[38] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers *et al.*, "The medical segmentation decathlon," *Nature communications*, vol. 13, no. 1, p. 4128, 2022.

[39] J. Faruk, S. B. Alam, S. S. T. Elma, S. Wasi, R. Rahman, and S. Kobashi, "Kidney abnormality detection using segmentation-guided classification on computed tomography images," in *2024 International Conference on Machine Learning and Cybernetics (ICMLC)*. IEEE, 2024, pp. 414–419.

[40] H. Messaoudi, A. Belaid, D. B. Salem, and P.-H. Conze, "Cross-dimensional transfer learning in medical image segmentation with deep learning," *Medical image analysis*, vol. 88, p. 102868, 2023.

[41] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET image processing*, vol. 16, no. 5, pp. 1243–1267, 2022.

[42] L. Wu, J. Zhuang, W. Chen, Y. Tang, C. Hou, C. Li, Z. Zhong, and S. Luo, "Data augmentation based on multiple oversampling fusion for medical image segmentation," *PloS one*, vol. 17, no. 10, p. e0274522, 2022.

[43] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Medical image analysis*, vol. 79, p. 102444, 2022.

[44] National Institute of Mental Health, "Neuroimaging informatics technology initiative (NIfTI)," https://nifti.nimh.nih.gov/, 2026, accessed: 2026.

[45] OpenCV, "Open source computer vision library," https://opencv.org/, last accessed: 2026-01-15.

[46] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," pp. 4700–4708, 2017.

[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[48] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 3645–3650.

[49] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[50] D. S. Char, N. H. Shah, and D. Magnus, "Implementing machine learning in health care - addressing ethical challenges," *New England Journal of Medicine*, vol. 378, no. 11, pp. 981–983, 2018.

[51] High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy ai," European Commission, Tech. Rep., 2019.