Computer Aided Medical Procedures
Prof. Dr. Nassir Navab

Dissertation

Machine Learning Methods for Segmentation in Autosomal Dominant Polycystic Kidney Disease

Kanishka Sharma

Fakultät für Informatik
Technische Universität München

Technische Universität München
Fakultät für Informatik
Lehrstuhl für Informatikanwendungen in der Medizin

Machine Learning Methods for Segmentation in Autosomal Dominant Polycystic Kidney Disease

Kanishka Sharma

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Hans Michael Gerndt

Prüfer der Dissertation: 1. Prof. Dr. Nassir Navab

2. Steven Sourbron, Ph.D.
University of Leeds, United Kingdom

Die Dissertation wurde am 13.06.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 20.10.2017 angenommen.

Abstract

Autosomal Dominant Polycystic Kidney Disease (ADPKD) is the most common inherited cystic kidney disease. It is characterized by the development of fluid-filled cysts and progressive enlargement of the kidneys. So far, there are no existing proven treatments for ADPKD, therefore, an effective disease-modifying drug would have important implications for patients. The increase in kidney volume has been associated with renal function decline and total kidney volume (TKV) is now acknowledged as a prognostic imaging biomarker for use in clinical trials on ADPKD. Therefore, developing efficient computational means for a reliable quantification of TKV is important for assessment of disease progression and evaluation of the efficacy of novel therapies in ADPKD. Currently employed methods for TKV quantification in ADPKD studies include stereology and manual segmentation. Both methods tend to be time consuming, especially in case of high-resolution CT or MR images. For improving applicability in clinical trials, TKV estimation has to be fast, accurate, and reproducible. So far, automatic segmentation in ADPKD has proved to be challenging due to widespread anatomical modifications in the kidneys and adjacent organs caused by development and expansion of irregularly shaped fluid-filled cysts during disease progression. Thus, segmentation of kidneys for TKV quantification in ADPKD is not only important from a clinical point of view, but it is also an interesting and challenging computer vision problem itself. Recently, Random Forests and Deep Learning approaches have gained considerable attention in the field of medical image segmentation. This PhD thesis analyzes the applicability and performance of these machine-learning methods for segmentation of polycystic kidneys to facilitate TKV quantification. In the first approach, a random forest based classifier was developed which requires minimal user interaction. The main novelty of the proposed approach is the use of geodesic distance volumes that contain intensity-weighted distances to a manual outline of the respective kidney in its middle slice (for each kidney) of the CT volume. The method was evaluated qualitatively and quantitatively on CT acquisitions of ADPKD patients using ground truth annotations from clinical experts. Furthermore, a fully automated segmentation method based on deep learning using fully convolutional neural network was developed which does not require hand-crafted features. Both methods were evaluated separately for their respective segmentation performance on complex polycystic kidney images from CT. The method based on deep learning achieves an overall good agreement with manual segmentations from clinical experts and facilitates fast and reproducible measurements of kidney volumes. This thesis demonstrates that machine learning can be successfully used for complex medical image segmentation tasks. Future research on machine learning and its applications in the medical domain might not only lead to improved algorithms for classical computer vision tasks such as image segmentation, but also facilitate holistic physical and biological models integrating heterogeneous clinical data from various sources that foster a thorough understanding of disease development, progression and treatment possibilities.

Zusammenfassung

Die Autosomal-dominante polyzystische Nierenerkrankung (ADPKD) ist eine der verbreitetsten, zystischen Nierenerkrankungen gekennzeichnet durch die Entwicklung mit Flüssigkeit gefüllter Zysten sowie durch eine progressive Vergrößerung der Nieren. Bis dato existiert keine validierte Behandlung und ein effektives krankheitsmodifizierendes Medikament wäre für die betroffenen Patienten von großer Bedeutung. Die Vergrößerung der Niere wurde mit einer Verschlechterung der Nierenfunktion in Verbindung gebracht und das Nierenvolumen (total kidney volume: TKV) gilt als alternativer Biomarker für den Krankheitsverlauf. Demzufolge ist die Entwicklung effizienter, computergestützter Algorithmen zur Überwachung der Nierenvergrößerung mittels TKV-Messungen von enormer Bedeutung für die Bewertung des Krankheitsverlaufs sowie die Analyse der Wirksamkeit neuartiger Therapien. Die für die Bestimmung des TKV etablierten Methoden sind Stereologie und manuelle Segmentierung. Beide Methoden sind, insbesondere im Fall hochauflösender CT- oder MR-Bilder, sehr zeitintensiv. Zur besseren Etablierung von TKV-Messungen in klinischen Studien, müssen diese schnell, präzise und reproduzierbar sein. Bisher hat sich die automatische Segmentierung bei ADPKD als anspruchsvoll erwiesen, nicht zuletzt aufgrund weitreichender anatomischer Veränderungen in den Nieren und den angrenzenden Organen, vor allem bedingt durch die Entstehung und Vergrößerung unterschiedlich geformter und mit Flüssigkeit gefüllter Zysten. Deshalb ist die Segmentierung der Nieren im Falle von ADPKD nicht nur von klinischer Bedeutung, sondern stellt auch für sich betrachtet ein interessantes und herausforderndes Problem des Bildverstehens dar. In der letzten Zeit haben Random-Forest-basierte und Deep-Learning-basierte Methoden großes Aufsehen im Bereich medizinischer Bildsegmentierung erlangt. Diese Doktorarbeit untersucht die Anwendbarkeit und Leistungsfähigkeit solcher Methoden für die Segmentierung polyzystischer Nieren und die TKV-Quantifizierung. Zunächst wurde ein Random-Forest-basierter Klassifikator entwickelt, welcher mit nur wenigen Benutzereingaben auskommt. Das entscheidende Novum dieser Methode ist die Nutzung geodätischer Distanzvolumen, die die bildintensitätsgewichtete Distanz zum manuellen Umriss der entsprechenden Niere im mittleren Schichtbild (jeder Niere) des betreffenden CT-Volumens verwenden. Die Methode wurde qualitativ und quantitativ auf der Grundlage vorhandener CT-Daten von ADPKD-Patienten und unter Verwendung von Goldstandard-Annotationen klinischer Experten evaluiert. Des Weiteren wurde eine Methode basierend auf Deep Learning und faltungsbasierten neuronalen Netzwerken entwickelt, die ohne die manuelle Definition von Merkmalen auskommt. Beide Methoden wurden getrennt und auch für CT-Datensätze komplexer, polyzystischer Nieren, sowohl im Hinblick auf ihre jeweiligen Segmentierungsergebnisse als auch die Genauigkeit der TKV-Messungen evaluiert. Die Deep-Learning-basierte Methode liefert gute Übereinstimmung mit den manuellen Segmentierungen klinischer Experten und ermöglicht somit eine schnelle, reproduzierbare Messung der Nierenvolumina. Diese Arbeit zeigt, dass maschinelles Lernen erfolgreich für komplexe medizinische Bildsegmentierungsaufgaben eingesetzt werden kann. Die weitere Erforschung von Methoden des maschinellen Lernens und deren medizinische Anwendung wird somit möglicherweise nicht nur zu verbesserten Algorithmen für klassische Probleme des Bildverstehens, beispielsweise Bildsegmentierung, führen, sondern auch zu holistischen, physikalischen und biologischen Modellen, welche verschiedene, klinische Informationsquellen einbeziehen und ein tiefgreifendes Verständnis von Krankheitsentstehung, Krankheitsverlauf und Behandlungsmöglichkeiten begünstigen.

Acknowledgements

First and foremost, I would like to thank my advisor Professor Nassir Navab for not only the opportunity to pursue this PhD but also for being a great mentor. I would also like to thank Professor Andrea Remuzzi, Professor Giuseppe Remuzzi, and Dr. Norberto Perico for the opportunity to work on the TranCYST project (Marie Curie Initial Training Networks, EU-FP7/2007–2013 grant: 317246) and for their guidance during my appointment at the Mario Negri Institute (MNI), Italy. I am really grateful to the senior members Dr. Maximilian Baust from CAMP group, Dr. Anna Caroli from MNI, and Dr. Luca Antiga from Orobix, for their guidance, invaluable research contribution and constant encouragement during this PhD. My sincere thanks to the PhD colleagues Christian Rupprecht and Loïc Peter for their help, support and discussions on machine learning that resulted in a highly effective collaboration. I would like to acknowledge the contribution of the TranCYST initiative in providing successful research and collaboration opportunities. For this, I also thank all the members of this project, especially Professor Dorien Peters (coordinator of the TranCYST ITN), and Babs Teng (project manager) for their support, as well as my PhD colleagues: Chiara, Laura V, Aylin, Arianna, Laura R, Zoraide, Martin, Alkaly, and Tareq for sharing this great experience. I am thankful to my fantastic colleagues: Michela, Davide, Sergio, Flavio, Bogdan, and Le Van at MNI, as well as the amazing colleagues at CAMP for their uplifting support during this PhD. I extend deep gratitude to my family and friends for their unwavering patience, support and love. I would like to thank my parents, Urmil and Dinesh Sharma, for being the most significant blessing in my life and enriching it with opportunities and freedom of choice. I am also very grateful to Parul, Pushkar, Vaishali, and Aarav for always being available with love and bringing joy to my life. I thank all my friends around the globe, especially: Eka, Michela & Andrea, Carolina & Matias, the Bozzetto family (Paolo & Gloria, Francesca, Giorgio), and Fiorenza for supporting me in this wonderful journey.

Contents

Part I

Imaging and Analysis in ADPKD

> *Nothing in this life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.*
>
> **— Marie Skłodowska-Curie**
> "quoted in Our Precious Habitat (1973) by Melvin A. Benarde, p. v."

1 Introduction

1.1 Background and Motivation

Autosomal dominant polycystic kidney disease (ADPKD) is the most common hereditary renal disorder, which initiates in utero and is characterized by sustained development and expansion of bilateral renal cysts. It is the fourth leading cause of chronic kidney disease (CKD) worldwide with majority of the patients progressing to end-stage renal disease (ESRD)[44, 45] and, currently no effective drug treatments are known to cure ADPKD. The *glomerular filtration rate* (GFR), an indicator of renal function remains normal for several decades in most of the ADPKD patients, thereby limiting diagnosis and evaluation of disease progression in addition to being unsuitable for studying effective therapies that would mostly have long-term benefits during early stages of ADPKD. In order to identify potential drug treatments for slowing down or even halting disease progression, it is vital to recognise effective biomarkers and their response to new therapies. In this respect, *total kidney volume* (TKV) has been identified as an important *imaging biomarker* of disease progression, allowing early and accurate measurement of cystic burden and likely growth rate in ADPKD. The increase in TKV usually precedes development of renal insufficiency by more than four decades and previous investigations have provided evidence that monitoring TKV is essential for assessment of disease severity, as well as, for predicting disease progression [25]. Currently employed methods for TKV quantification in ADPKD studies include stereology and manual kidney segmentation. Both methods tend to be time consuming and for improving applicability in clinical trials, TKV estimation has to be fast, accurate, and reproducible. However, automatic segmentation in ADPKD is very challenging due to widespread anatomical modifications in the kidneys and adjacent organs, caused by development and expansion of the irregularly shaped cysts during disease progression. Thus, segmentation of polycystic kidneys is not only important from a clinical perspective, but it is also an interesting and challenging problem in the field of computer vision. Since the last decade, pattern recognition algorithms have become widely popular in improving machine intelligence for several tasks including medical image segmentation. Machine-learning models based on efficient feature engineering and representation learning are capable of identifying complex patterns within the data, thereby providing reliable outcomes with good accuracy and generalization. This thesis analyzes the applicability and performance of two separate machine-learning approaches based on *Random Forests* and *Convolutional Neural Networks*, respectively, for segmentation of polycystic kidneys from CT dataset of ADPKD patients to aid TKV computation. We demonstrate that machine learning can be successfully used for complex medical image segmentation tasks.

1.2 Renal Anatomy and Physiology

The kidneys are paired retroperitoneal organs located on either side of the vertebral column between the peritoneum and the posterior muscular wall of the abdominal cavity. The left kidney is located slightly superior than the right kidney due to large size of the liver located on the upper right portion of the abdominal cavity [81]. As shown in figure 1.1, the parenchyma of kidney has two main regions, namely, the outer renal cortex and the inner renal medulla. The medulla has conical subdivisions known as the renal pyramids with bands of renal columns separating adjacent pyramids. The broad base of the renal pyramids faces the renal cortex while the apex, also known as the papilla, points towards the renal pelvis. The concave side of the kidney consists of an indentation known as the renal hilum which provides an entry space for the renal artery, renal vein, and the ureter. The funnel shaped enlarged upper end of the ureter is known as the renal pelvis which allows flow of urine from the kidney to the urinary bladder and is also the point of convergence where a system of ducts named calyces transport urine for excretion.

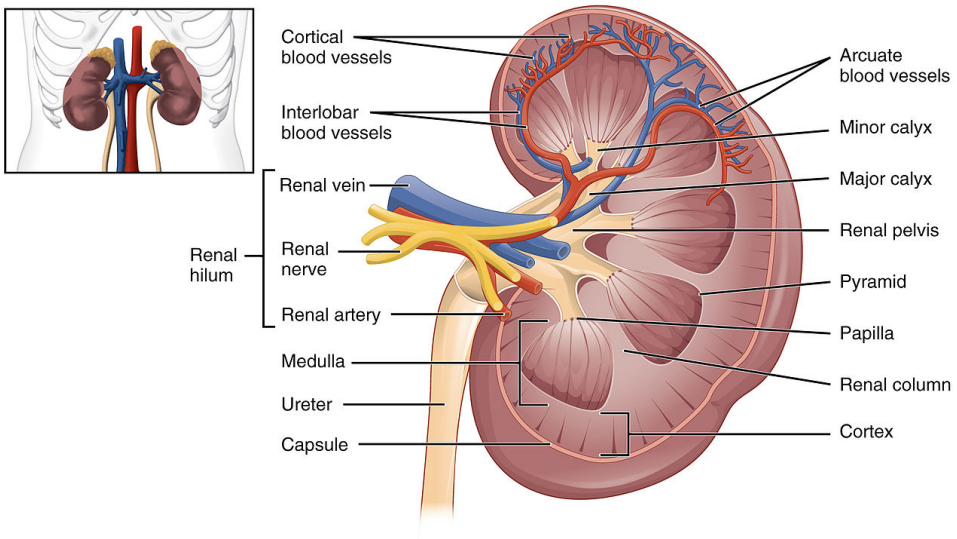


Fig. 1.1. **Normal Kidney Anatomy.** Cross section of a normal kidney showing the outer renal cortex and the inner renal medulla consisting of conical subdivisions known as the renal pyramids. The concave side of the kidney consists of the renal hilum which provides an entry space for the renal artery, renal vein, and the ureter. The funnel shaped enlarged upper end of the ureter is the renal pelvis. (*Image courtesy: cnx.org/content/col11496/1.6/*)

Nephrons, the basic structural and functional unit of the kidney span the cortex and medulla, as shown in figure 1.2. The outer renal cortex contains the glomeruli and convoluted portion of the proximal and distal tubules, while the inner renal medulla is composed of the straight portion of the proximal tubule, the henle's loop and the collecting duct. The normal single kidney volume in a healthy human adult has been estimated to be approximately 202 ± 36 ml (for men) and 154 ± 33 ml (for women) as measured on MRI [29] (mean $\pm$ SD). Additionally, the average size of each kidney is about 10 to 13 cm long, approximately 5 to 7.5 cm wide and 2 to 2.5 cm thick, correlated with the age and height of the subject and corresponding to a kidney weight that varies between 125 and 170 gm.

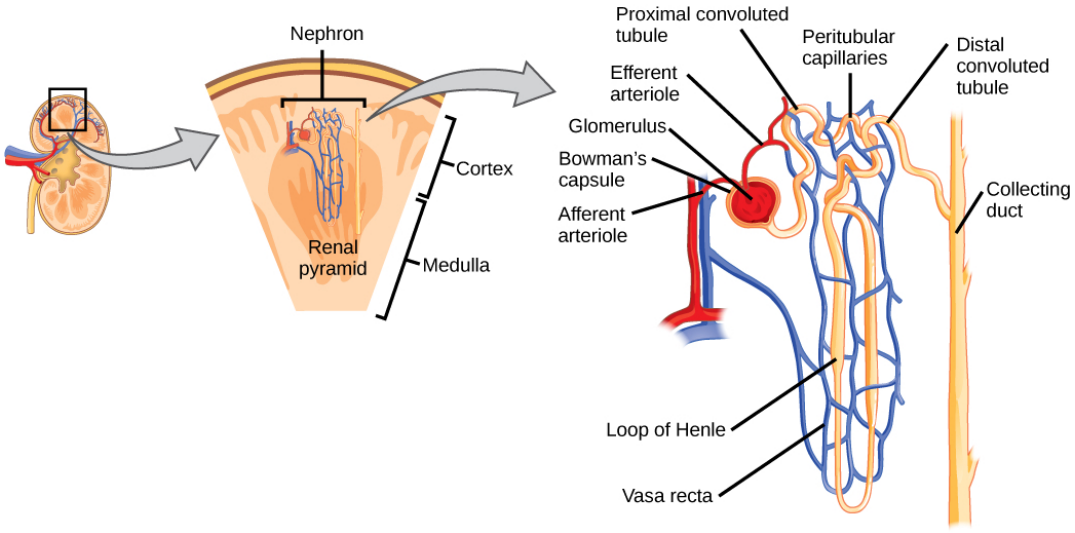


Fig. 1.2. **Normal Kidney Nephron.** Nephrons are the basic structural and functional unit of the kidney. The outer renal cortex contains the glomeruli and convoluted portion of the proximal and distal tubules, while the inner renal medulla is composed of the straight portion of the proximal tubule, the henle's loop and the collecting duct. (*Image courtesy: cnx.org*)

The kidneys have many vital functions such as maintaining whole body homeostasis, blood pressure regulation, purifying blood from toxic metabolic waste products, producing urine, hormones, and absorbing minerals. In the event of a kidney disease, the homeostatic functions of the kidneys are highly compromised leading to serious alteration of volumes and composition of body fluids which is usually accompanied with decreased quality of life for the patient suffering from a kidney disorder. While some disorders such as Acute Kidney Injury (AKI) can be reversed using renal replacement therapy such as hemodialysis, or other specific therapies like administration of intravenous fluids. Complications in an existing kidney problem or other prevailing health problems such as diabetes and high blood pressure can lead to gradual and progressive loss in renal function over a period of months or years, a condition also known as *Chronic Kidney Disease* (CKD). The last stage of CKD is a pathological condition identified as *End Stage Renal Disease* (ESRD), more generally known as kidney failure which necessarily requires the patient to undergo dialysis or kidney transplantation to survive. Patients afflicted with ADPKD are subject to CKD and majority of them reach ESRD.

1.3 Pathogenesis of ADPKD

The cyst formation in ADPKD is known to derive from mutations in *PKD1* and *PKD2* genes encoding the proteins polycystin-1 and polycystin-2, respectively. The PKD1 gene mutation involves approximately 85% cases of ADPKD and these individuals usually show more severe disease with early cyst development and are more likely to progress to ESRD. The possession of two identical forms of PKD1 gene, one inherited from each parent (i.e. PKD1 homozygosity) is known to be lethal in utero [91]. The PKD2 gene mutation is known to affect the remaining 15% cases. However, studies on some families with ADPKD have found neither PKD1 nor PKD2 mutations, postulating that an additional genetic loci may be associated with the disease [4, 15, 36, 85, 139]. The latter two categories of patients (PKD2 gene or postulated additional

genetic loci) are known to present with milder form of the disease but there is also evidence on families with severe clinical courses [4, 36]. In case of heterozygous mutations involving both PKD1 and PKD2 gene, the severity of disease is worse than mutation of a single gene [93]. The course of disease severity has also been linked to inheritence from each parent. It has been suggested that patients inheriting ADPKD from their father experience less severe disease compared to inheritence from the mother [10].

Previous studies have indicated the initiation of renal cyst development in renal tubules and in rare cases the Bowman's capsule [46]. At first, the cysts appear as tiny growths in the renal tubule and eventually expand relentlessly. For several decades, however, the cell proliferation in ADPKD is relatively low but this allows individual cysts to remarkably increase in size (even $>$ 10 cm in diameter) and the combined effect of increased cell proliferation and fluid secretion promotes progressive cyst enlargement [45]. The rate of cyst growth is not significantly different between PKD1 and PKD2 mutations, however, the median age for onset of ESRD is approximately 53 years in patients with PKD1 mutation while, it is estimated to be around 69 years in patients with PKD2 mutation [50]. The prevalence of all cystic manifestations in ADPKD increases with age but no specific pattern of cyst growth has been identified so far and investigations have only suggested that increase in the cyst volume is largely individualized, varying from patient to patient. For every individual with ADPKD, each cyst in a polycystic kidney is considered to function independently, but known to have a constant growth rate. The overall growth of all these individual cysts in both kidneys causes an exponential increase in the *total kidney volume* (TKV), with the oldest and largest cysts accounting for greater effect on the TKV change compared to the younger and smaller cysts [45]. The gross pathology of polycystic kidneys is shown in figure 1.3, depicting independent and heterogeneous growth of cysts in individual kidneys. The variation in kidney shape, size, and volume of polycystic kidneys in comparison to normal kidneys, as well as

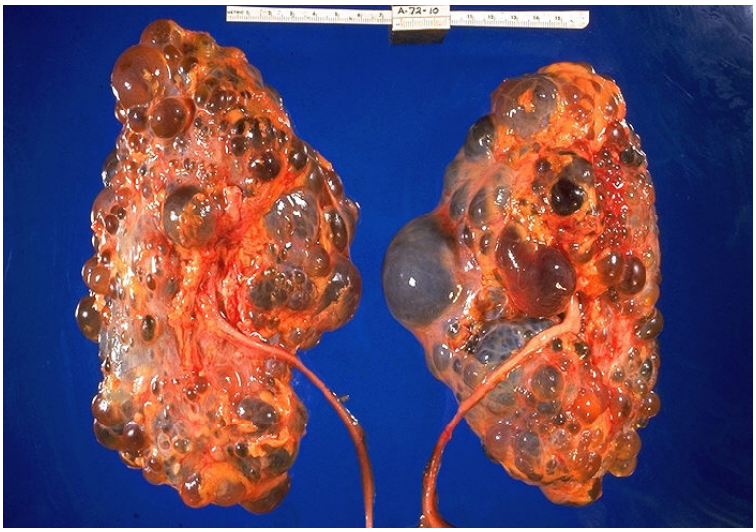


Fig. 1.3. Gross Pathology of Polycystic Kidneys. In ADPKD, increase in the cyst volume is largely individualized, varying from patient to patient. For every individual with ADPKD, each cyst in a polycystic kidney is considered to function independently but known to have a constant growth rate. Eventually, overall growth of all these individual cysts causes an exponential increase in the TKV. (*Image courtesy: phil.cdc.gov/PHIL_Images/02071999/00002/20G0027_lores.jpg*)

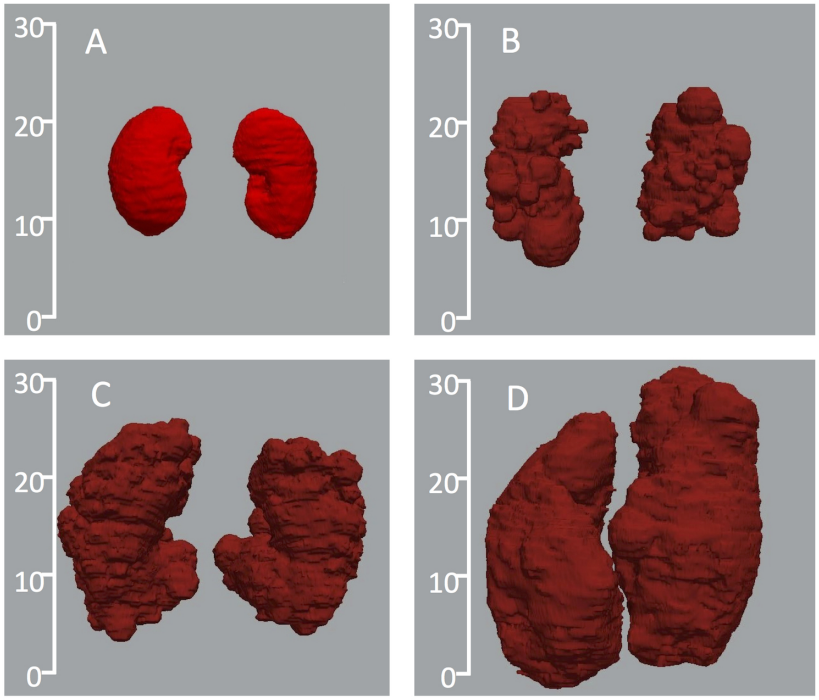


Fig. 1.4. Three-dimensional representation of ADPKD kidneys in comparison with normal kidneys. Scales represent dimension in cm. The kidney shape, size, and volume highly differ between the normal control (panel A: $TKV = 591$ ml) and the patients (panel B: $TKV = 1,327$ ml; panel C: $TKV = 3,026$ ml; panel D: $TKV = 5,836$ ml). TKV is the combined volume of left and right kidneys.

the variability among different ADPKD patients is depicted in figure 1.4. The volume and shape of ADPKD kidneys can vary considerably among different patients. Some polycystic kidneys adopt regular shape but most patients have markedly irregular shaped kidneys with prominent surface irregularities due to the presence of different sized and shaped cysts.

Manifestations of ADPKD also include development of hepatic cysts (70%) and pancreatic cysts (5%), which may spread to the spleen, prostate and seminal vesicles. The number and size of hepatic cysts has shown to correlate with female gender and, severity of the renal disease [16]. Other risks include increased chances of heart valve abnormalities and aneurysms in aorta [133] or in blood vessels at the base of the brain [24]. Moreover, associated clinical symptoms of ADPKD such as hypertension (blood pressure 140/190 mmHg), hematuria, and abdominal pains due to passage of stones and urinary tract infection [136] can lead to renal insufficiency. ADPKD patients progressing to ESRD require hemodialysis, peritoneal dialysis or renal transplantation.

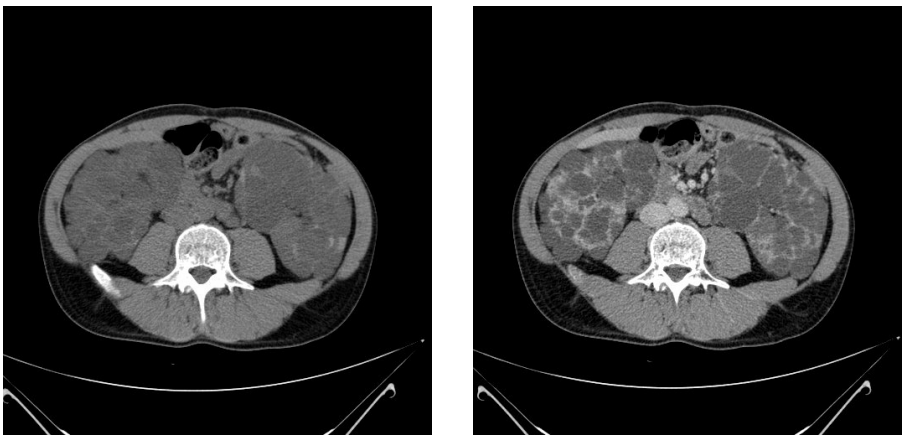
To identify potential drug treatments for slowing down or even halting ADPKD progression, it is vital to recognise effective biomarkers and their response to new therapies. TKV has been identified as an important imaging biomarker for assessment of disease severity and for predicting disease progression in ADPKD. In the next section, we describe different imaging techniques for monitoring morphological changes in the kidneys to aid TKV computation in ADPKD.

1.4 Imaging Techniques in ADPKD

In ADPKD, morphological changes in the kidneys and its compartments can be captured on imaging modalities such as Ultrasound (US), Computed Tomography (CT), or Magnetic Resonance Imaging (MRI). Renal Ultrasonography is currently performed for presymptomatic screening and assessment of ADPKD. With easy accessibility in clinics, US helps to acquire large patient dataset that can be useful in managing ADPKD. However, it suffers from limitations of low spatial resolution, high operator variability, lack of reproducibility and limited accuracy of TKV measurements in comparison with imaging modalities such as CT and MRI. Therefore, it is rather unsuitable for detecting smaller cysts and monitoring short-term morphological changes in ADPKD. Recent work described statistical shape modeling for renal volume measurement on tracked ultrasound using normal kidney shaped phantoms [104] but, further investigations are required to sufficiently express the wide variety of deformations found in polycystic kidneys and to increase the prognostic value of US in ADPKD.

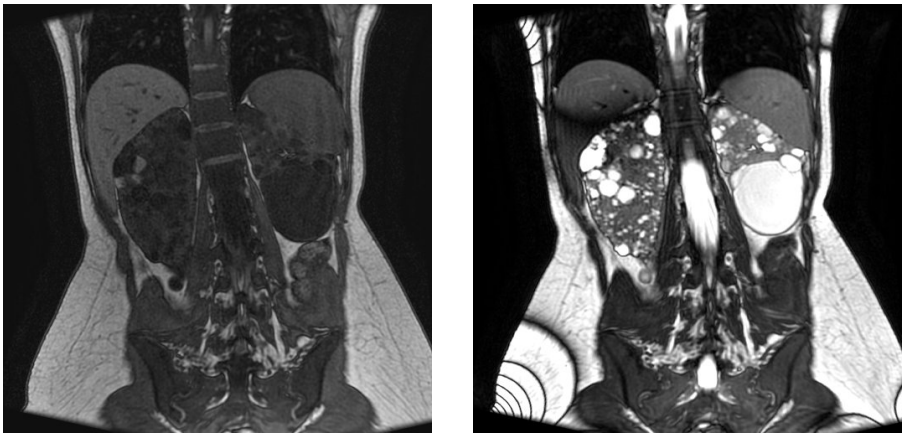
Other imaging modalities such as CT and MRI offer higher spatial resolution, reproducibility, and facilitate detection of smaller cysts ($<$ 1 cm in diameter) that are not captured on US [92, 152]. Several studies have utilized imaging methods based on CT and MRI to reliably and accurately measure TKV in ADPKD patients [5, 6, 26, 27, 47, 48, 70, 71, 110, 122]. The accuracy of TKV measurement using CT and MRI is comparable, however, both modalities have their respective advantages and disadvantages. The first work on TKV computation using CT scans of ADPKD patients was reported by Thomsen et al. [131]. On CT, the abdominal section of a polycystic kidney highlights different pixels based on the tissue radiointensity. The use of contrast agents further enhances the differentiation between cysts, healthy and residual parenchyma as shown in figure 1.5.

While CT acquisitions are relatively faster than MRI, the main disadvantage of CT is the exposure to ionizing radiation and the use of nephrotoxic contrast agents. Despite longer acquisition times of MRI, it is becoming a popular choice of use for imaging studies in



(a) CT image without contrast agent (b) CT image with contrast agent

Fig. 1.5. ADPKD CT Images. (a) Axial section of polycystic kidneys on CT image highlighting different pixels based on the tissue radiointensity. (b) Use of contrast agents further enhances the differentiation between pixels depicting cysts, healthy tissue and residual parenchyma.



(a) MR T1-Weighted Image (b) MR T2-Weighted Image

Fig. 1.6. ADPKD MR Images. (a) T1-weighted acquisition of polycystic kidneys (coronal-view) where parenchyma appears hyperintense while fluid-filled renal cysts appear hypointense. (b) On the contrary, T2-weighted acquisition shows cystic fluid as hyperintense while surrounding parenchyma is hypointense.

ADPKD with the advantages of high signal-to-noise ratio and good contrast between soft tissues. To monitor changes in the kidney morphology, coronal (or axial) T1-weighted acquisitions are generally used where the parenchyma appears hyperintense while the fluid-filled renal cysts appear hypointense as shown in figure 1.6 (a). On the contrary, T2-weighted acquisitions (figure 1.6 (b)) are used mostly for studying the cyst volume, as cystic fluid has high signal intensity relative to surrounding parenchyma thereby appearing hyperintense and distinguished from renal parenchyma which is hypointense [144, 145].

The segmentation of polycystic kidneys for quantifying kidney volumes from CT or MRI is very challenging due to non-uniform renal cyst growth leading to high variability in kidney morphology. As described in the previous section, polycystic kidneys are characterized by their markedly irregular shape and size in comparison to normal kidneys and sometimes surface irregularities are prominent due to the presence of surface cysts of different size. On both CT and MRI, additional clinical complications hindering automated assessment of TKV include the presence of hepatic cysts which appear identical to kidney cysts, as well as, the presence of hemorrhagic renal cysts which appear rather dissimilar to other fluid filled cysts leading to high intensity variability within the kidney. Thus, development of a fully-automated segmentation method for fast and precise TKV estimation remains a challenging problem. In the next chapters, we describe different strategies for segmentation in the domain of medical imaging along with their respective application in ADPKD. Additionally, we compare different methods available for TKV quantification on CT and MR images within clinical studies on ADPKD. The limitations of currently employed methods for TKV computation in APDKD provide good motivation for investigating novel strategies to improve segmentation of polycystic kidneys from acquired imaging (CT or MR) dataset. Therefore, we assess the performance of machine learning based methods for segmentation in ADPKD, details of which are described in later chapters of this dissertation.

Medical Image Segmentation

2

2.1 Introduction

Medical imaging has evolved as a critical component of diagnosis, treatment planning and research outcomes. Imaging modalities such as Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Ultrasound (US) provide non-invasive yet effective ways of revealing normal or diseased anatomy (or physiology) and there is an increasing demand for automatic assessment of data retrieved from these imaging devices. In clinical settings, delineation of anatomical structures acquired on one or more imaging modalities formulates an important task, and is conventionally approached by manually outlining regions of interest on slice-wise sections of the acquired images by clinical experts and trained personnel. Over the recent years, image segmentation algorithms have become increasingly popular for extracting regions of interest, thereby, assisting or even completely automating radiological tasks in several medical applications such as pathology localization, volumetric quantification, computer-assisted surgeries or treatment planning [98]. Image segmentation requires fundamental understanding of the image content and localizing useful properties (or features) within an image facilitates extraction of desired regions of interest. Essentially, a segmentation task aims at partitioning the image into constituent regions that are homogeneous in some respect such as intensity, texture, shape or a combination of representative features. However, automatic segmentation of imaging data particularly acquired on patients in clinical settings proves to be non-trivial due to modality specific as well as anatomy specific limitations. In the acquired images, these challenges appear as undesired intensity variations or texture contrast, imaging artifacts, noise, missing or deformed boundaries between structures, lack of a definitive shape and location owing to morphological deformation. Ongoing research in this field aims to achieve reproducible, accurate and fast segmentation outcomes while addressing the challenges in anatomy specific regions and on various imaging modalities. In the next sections, we describe peculiarities of medical image segmentation and discuss traditional, as well as, recently proposed approaches for segmentation in the medical imaging domain.

2.2 Peculiarities of Medical Image Segmentation

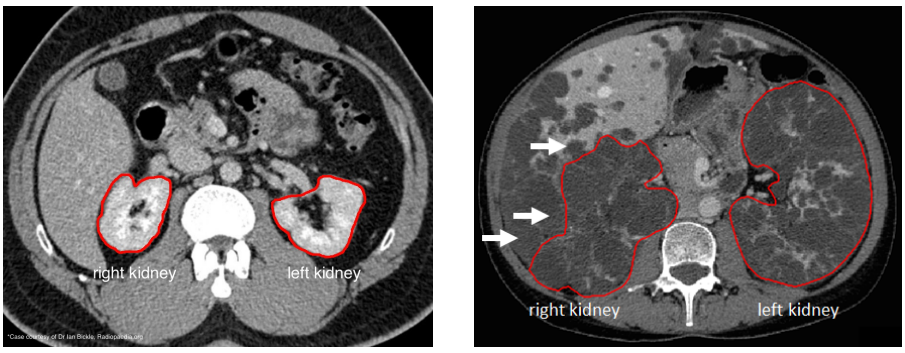
For reproducible, accurate and fast segmentation outcomes, it is important to not only understand the image formation process but also to consider associated anatomy specific and modality specific limitations. In ADPKD, the difficulties in segmentation arise due to one or several issues including: leakage problem, morphological variability, modality specific intensity inhomogeneity, intra and inter-subject intensity differences, partial volume effects, and noise. Additonally, for multi-centric clinical studies, the imaging data frequently suffers from variable quality owing to acquisition on different imaging scanners used at independent acquisition sites. Similarly, the variation in image quality may also be seen within an imaging dataset utilized from separate clinical studies with diverse acquisition protocols. Some peculiarities of medical image segmentation are described below.

Leakage Problem

The leakage problem occurs when the organ to be segmented is surrounded by tissue with similar physical properties. In ADPKD for instance, an extra-renal manifestation includes the presence of hepatic cysts which generally exhibit similar physical properties to the cysts in the kidneys. When visualized on CT or MRI, the cysts in the kidneys and liver are visualized with similar intensity values and as a consequence, the kidney border is hardly distinguishable from the surrounding liver. From an imaging standpoint, this clinical complication presents itself as a leakage problem as shown in figure 2.1. Typically, a leakage problem can be addressed by incorporating prior shape information as a segmentation citerion but, extreme morphological variability such as those seen in ADPKD can often limit the applicability of such methods.

Morphological Variability

Morphological variability can be a limiting factor when attempting to use generic segmentation methods resulting in poor generalization and undesirable segmentation outcomes. In ADPKD, the increase in the cyst volume is largely individualized varying from patient to patient with no specific identified pattern in the cyst growth. The overall growth of the individual cysts causes an exponential increase in the kidney volume accompanied with highly variable morphological



(a) Normal Kidneys (b) ADPKD Kidneys

Fig. 2.1. Leakage Problem and Morphologocal Variability. ADPKD Kidneys (b) are difficult to segment due to severe morphological changes in comparison to healthy kidneys (a). White arrows show adjacent liver cysts exhibiting similar physical properties leading to a leakage problem.

changes in the polycystic kidneys, as shown in figure 2.1 (also refer figure 1.4). Previous works have suggested methods for building models that learn patterns of variability from a set of already segmented images [31]. Incorporating deformation to the data consistent with the training set seems plausible but, in case of extreme morphological variability such as those in late stages of ADPKD, many training examples would be necessary and it may still be difficult to develop satisfactory segmentation solutions solely based on shape models without additonal feature representations to guide the segmentation.

Intensity Inhomogeneity

Intensity inhomogeneity, also referred to as intensity non-uniformity (INU), shading or spatial bias is an imaging artifact perceived as a smooth variation of intensities across the image [12]. Even though it might appear inconspicuous to human observer, such an artifact can degrade different image analysis methods including feature extraction, segmentation and registration. Mainly appearing on MR images due to distortions in the magentic field [98, 149], segmentation methods typically assuming a constant intensity value per region such the piecewise constant Mumford-Shah model [88] may perform poorly in the presence of an intensity inhomogeneity artifact. Different methods have been proposed for correction of intensity inhomogeneity based artifact including those performing segmentation along with a bias field fitting [97, 156]. Other methods based on parametric bias field correction [129] or non-parametric non-uniform intensity normalization [123] have also been investigated.

Noise

A random and unwanted signal variation can be considered as noise and it is inherently present in all electronic systems. Noise can originate from different sources including electronic interference. On imaging modalities such as CT, *poisson noise* arises due to the statistical error of low photon counts causing random thin bright and dark streaks that appear along the direction of greatest attenuation. This type of noise can be reduced by using iterative reconstruction, or by combining data from multiple scans [14]. Another type of artifact known as the *speckle*, which is a noise-like variation appears as irregular granular pattern in an image making it difficult to recognise differences in contrast and can be reduced by using different filtering techniques (such as: median filter as shown in figure 2.2) [59]. In some cases, imaging artifacts may be also be caused due to external reasons such as presence of a metal implant. Although, not strictly an internal source of noise but such a noisy artifact may degrade the image quality to a high extent. These different kinds of artifacts leading to reduced quality of images have been shown in figure 2.2.

Partial Volume Effects

Partial volume effects appear when multiple tissues contribute to a single pixel causing the blurring of intensity across boundaries [97]. Partial volume artifacts are commonly seen on CT and MRI, when the resolution is not isotropic and in many cases, poor along one axis of the image. Higher resolution imaging helps to alleviate this problem and the most common approaches to deal with partial volume effects include soft segmentations. Segmentation methods generally enforce a binary decision on whether a pixel is inside or outside the object of interest, also known as hard segmentation. Instead, soft segmentation approaches allow regions or classes to overlap, thereby allowing for uncertainty in the location of object boundaries.

Fig. 2.2. **Imaging Artifacts.** CT image of ADPKD kidneys with speckle noise (left). Speckle noise reduced using median filter (centre). Imaging artifact caused by a metal implant (right).

2.3 Current Trends in Medical Image Segmentation

Different strategies for image segmentation have been proposed in literature using algorithms for partitioning an image into non-overlapping regions that are homogeneous with respect to some common characteristics and corresponding to distinct regions of interest (such as the anatomical structures) in the image. We describe some of the theoretical frameworks based on region-growing, contour evolution, graph based, or shape models that have been proposed for medical image segmentation on CT and MRI along with their recent application in ADPKD.

Thresholding based methods

Thresholding is one of the simplest traditional approaches for segmenting scalar images by creating binary partitions of the image intensities. Different structures within an image can be separated based on their contrasting intensities, also known as *threshold* values. Thus, segmentation is achieved by grouping together image pixels with intensities lying in specific threshold range into one class and other pixels into respective classes based on the threshold range of their intensities. One of the earliest attempts for segmentation in ADPKD used histogram-based statistical approach, popularly known as the "Otsu-Thresholding" [90] for automatically classifying compartments within polycystic kidneys into cysts, healthy parenchyma and residual intermediate-volume [5], However, the thresholding method for segmentation has limitations as it does not take into account spatial characteristics of the image, thus, making it sensitive to noise and intensity inhomogeneities. Such image artifacts can easily corrupt the image histogram making separation more difficult. Even though, variations on classical thresholding methods have been reported in literature [114], but the use of thresholding remains essentially in use for several image pre-processing tasks and other techniques have been investigated that incorporate other information based on local intensities and connectivity [78].

Region based methods

The most popular region based approach for segmentation is the *region-growing* algorithm which uses local neighbourhood intensity properties for aggregating pixels together [2]. Starting with initial seed points placed manually in the region of interest, the algorithm automatically examines all neighbouring pixels to determine if they have similar intensities

to these seed points and in that case, iteratively includes the new pixels thereby growing the region until intensity homegenity criteria is no longer satisfied. Region based methods are sensitive to noise, especially in case of CT images with partial volume artifacts [99]. In ADPKD, region growing method was attempted for segmentaion of polycystic kidneys on MR images [86].

Graph-Cut based methods

One of the earliest implementations using graph-cuts was based on the minimum spanning tree (MST) used for point clustering as well as image segmentation [154]. In a graph-cut based segmentation, image pixels are represented as nodes (i.e. vertices) of a graph connected via edges to neighboring pixels and a weight is associated with each edge based on a property (such as difference in image intensity) of the pixels connecting these edges. Thus, graph cut partitions a directed or undirected graph into disjoint sets and the optimality of these cuts is generally introduced by associating an energy to each cut. An automated method using graph-cuts in combination with surface model was previously used on rather unsubstantial or mildly deformed kidneys of transgenic mice with ADPKD on MRI [79].

Graph-Search based methods

A popular technique, known as the *livewire segmentation* is based on optimal graph-search problem [82] providing boundary definition using the shortest paths between nodes in a graph (as described by the Dijkstra algorithm) [38]. The livewire algorithm first convolves the image with a suitable filter such as the canny-edge detector [20] to extract the edges and then uses this filtered image as a graph where image pixels are defined as nodes and the edges are weighted according to features exhibited by the filter. This method generally relies on user interaction for placing successive anchor points on the object of interest in the image while minimum cost path is computed and drawn as the boundary between these successive anchor points. In chapter 3, an application of the livewire method is described in detail for kidney segmentation on both CT and MRI to allow TKV computation in ADPKD.

Boundary based methods

The boundary based segmentation technique is generally based on contour evolution and the most popular approach is described using *active contours*, which is an energy-minimizing model guiding contour deformation [67]. In particular, an object is described by a contour delineating its boundary and the desired configuration of the contour is modeled as a local minimum of an energy defined on the image data. Thus, starting from a manual contour initialization, the contour minimizes the energy and evolves towards the boundary of the object of interest. The performance of active contours is strongly dependent on the user-defined manual contour. In ADPKD, active contours have been previously used in combination with sub-voxel morphology based algorithm on MRI [103]. Another approach modeled a spatial prior probability map (SPPM) with evolving kidney contours incorporated into a level set framework for segmentation of polycystic kidneys from MR images in ADPKD [68].

Active Shape Models

This segmentation method involves generation of appropriate shape models built from series of data reflecting morphological properties of the object of interest in the image. The most popularly used *active shape models* are based on a statistical a-priori mean model of the object

of interest derived from series of templates to create an atlas with enough variability but without lacking specificity. The algorithm for active shape modelling (ASM) makes an initial rough guess of shape, orientation, scale and position using information such as the edges or distance criteria to find differences between the template model and actual image data and guides a deformation process that iteratively progresses until convergence criterion is satisfied. ASMs can be useful for recognizing structures with a definitive shapes but have limitations when presented with highly complex structures with insufficient information to describe possibile deformations to fit the object to be segmented. In ADPKD, active shape modelling has been described before for 3D segmentation of polycystic kidneys with limited convergence to polycystic kidney shapes due to high complexity of these kidneys [100], as shown in figure 2.3.

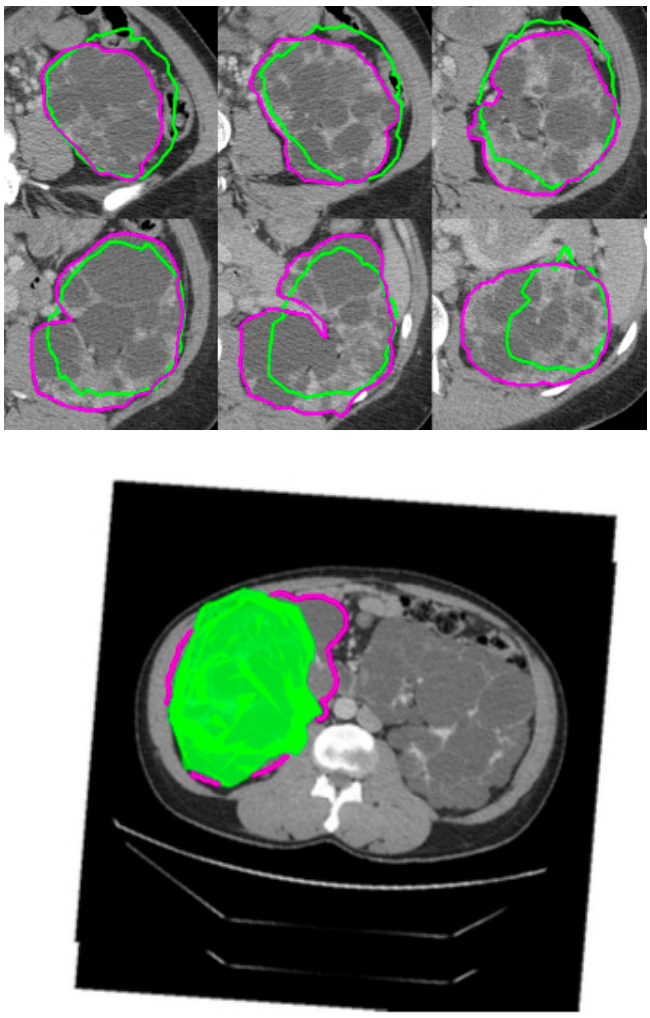


Fig. 2.3. **Active Shape Model in ADPKD.** Top: Magenta contours represent gold-standard manually outlined by an expert operator while green contours show the deformed model's intersections with each axial plane. Bottom: Green deformed model unable to reach the real kidney dimensions shown using the purple contour (Image courtesy: [100]).

2.4 Machine Learning in Medical Applications

An intrinsic quality in humans is their ability to use knowledge and experience to make decisions for solving complex tasks. For instance, in the field of medical imaging, experts use their knowledge about specific properties in the data and their vast experience of encountering similar data at previous instances for making reliable decisions on newly presented data to find optimal solutions. In a similar way, machine learning relies on its ability to learn directly from the data and generalize from past observations for performing future predictions. This simulated "human-like" behaviour of using knowledge from the data and experience from past observations for making future predictions can be very useful for desigining optimal solutions in various medical applications. However, learning-based methods gained popularity only recently as they have been notoriously known to possess a "black-box" nature that needs to be well understood prior to incorporation into any application. The potential of machine learning algorithms to model complex feature representations and their possibility to scale well to variety of data allows reproducible and accurate results for different tasks.

Since the last decade, learning-based approaches have been successfully used for solving various tasks in the medical domain. In the field of medical imaging, machine learning has been successfully used for vital tasks such as anomaly detection, organ localization, segmentation and disease prediction. Moreover, these methods have also provided improved accuracy in various image registration tasks by learning application-specific similarity measures directly from the data. Applications requiring combination of diverse information can also be incorporated into a learning based framework such as in computer aided diagnosis where decisions are based on information from multiple sources such as imaging data, patient history and current symptoms.

Learning algorithms can adopt different strategies depending on their application which can be categorized into three major learning paradigms, namely *supervised*, *unsupervised* and, *semi-supervised* learning. In *supervised learning*, the aim is to predict a desired output variable Y based on an input vector X on the assumption that both input and output variables approximately follow a predictive relationship $Y = f(X)$, or in a probabilistic manner, model a conditional distribution $P(Y|X)$. During training, the desired output is already known for incoming input and the learning algorithm approximates a mapping function f, or models a conditional distribution $P(Y|X)$, such that after training the model, it can automatically predict the output when a new unseen input observation arrives based on the previously learned examples during the learning phase. Supervised learning can be used for *classification* tasks where the new incoming observation is assigned to one of the previously defined discrete classes. It can also be used for *regression* tasks where the desired output is a continuous variable. *In this thesis, we will focus on machine learning methods using supervised learning for classification.* In *unsupervised learning*, there are no outputs associated to the input observations and the aim is to find similar groups in the input feature space of X, also known as *clustering* task, or to estimate the distribution of input observations X, also known as *density estimation*. The third major learning paradigm is known as *semi-supervised learning* where the outputs are known only for few input observations and the aim is to learn the function f or estimate the distributions $P(Y|X)$ by making use of both labelled as well as unlabelled data sets.

An important application in the domain of medical imaging includes automatic segmentation, however, finding optimal segmentation algorithms that perform similar to human experts is rather challenging. In clinical settings, methods employed for segmentation need to be highly accurate, reliable and robust against errors. However, the imaging dataset can be frequently multi-dimensional or even multi-modal and suffer from variations in image quality, resolution, signal to noise ratio and additionally consist of anatomy specific complexities. For this reason, designing automated methods that allow direct translation of a complex phenomenon into an appropriate and realistic model can be very challenging. Machine-learning based classification is among the popular approaches for image segmentation and it exploits the advantages of supervised learning by assigning to image pixels the probabilities of belonging to the region of interest. Different classifiers have been suggested in literature such as the rule-based classifiers, nearest neighbor classifiers, naïve bayes classifiers, support vector machines (SVM), decision tree classifiers, and neural networks.

In this dissertation, two separate approaches based on a random forest classifier and deep convolutional neural networks (CNNs), respectively have been evaluated for segmentation of polycystic kidneys on CT dataset acquired from clinical studies in ADPKD. Random forests are easily scaled to large training datasets, allow fast training and predictions, provide good generalisation to previously unseen data yielding a probabilistic output and they can also be used for multi-class problems. Moreover, they can decide importance of different features and are generally easier to interpret by humans. These useful properties allow their effective use for classification and several applications have successfully used random classification forests previously [34, 65, 95, 106, 120].

The concept of CNNs is known to be inspired from the initial works of Nobel laureates, David Hubel and Torsten Wiesel on information processing in the visual cortex of a cat [60]. Their experiments showed that the visual stimuli are processed by a cascading hierarchy of neurons that are arranged in a particular architecture. Comprised of simple and complex cells, these neurons extract increasingly complex information from the pattern of light cast on the retina to form an image. Overall, their work was fundamental to understanding the process of building visual perception of the world around us. Over the last years, deep learning approaches have become widely popular and in particular, CNNs have garnered special attention by achieving promising results in a variety of classification applications [30, 76, 121, 130]. CNNs are capable of encoding image specific features and can therefore be efficiently used for extracting low level features as well as automatically capture advanced abstract information from the input data which can be very useful in the context of image classification and segmentation. In subsequent chapters, a detailed description is provided on random forests and deep convolutional neural networks along with their application in ADPKD for segmentation.

2.5 Outline and Contributions

The main contributions of this dissertation are presented in the next chapters and relate to the applicability and performance of machine-learning based approaches using *Random Forests* or *Convolutional Neural Networks* for segmentation of polycystic kidneys from CT dataset of ADPKD patients at different stages of the disease.

Chapter 3: Kidney Volume measurement methods in ADPKD

In this chapter, the importance of TKV in ADPKD is highlighted. As the main contribution, a comprehensive comparison is made between different available methods for TKV computation on CT and MR images in terms of reproducibility, accuracy, precision, and time requirement. Our results help in identifying the most suitable kidney volume measurement method for clinical studies evaluating treatment efficacy on ADPKD progression.

The presented contribution has been published in: *K. Sharma, et al. "Kidney volume measurement methods for clinical studies on autosomal dominant polycystic kidney disease". In: PLoS ONE, (2017)*.

The limitations of currently employed TKV measurement methods described in this chapter provide good motivation for developing new segmentation strategies for increasing efficiency of TKV measurement routine in ADPKD clinical studies. Presented contributions in this dissertation aim at improving the segmentation of polycystic kidneys using machine learning methodologies.

Chapter 4: Random Forests for Segmentation

In this chapter, the key concepts of decision tree learning and classification using *random forests* are summarized. The applicability and performance of a random forest classifier for segmentation of polycystic kidneys on CT dataset of ADPKD patients with severe renal insufficiency is analyzed. As a novel contribution, geodesic distance volumes consisting of intensity-weighted distances to a manual outline of the respective kidney in its middle slice (for each kidney) of the CT volume are introduced as additional source of information to the random forest classifier. The segmentation performance of the proposed approach is evaluated qualitatively and quantitatively using ground truth annotations from clinical experts.

The presented contribution can be found in: *K. Sharma, et al. "Semi-Automatic Segmentation of Autosomal Dominant Polycystic Kidneys using Random Forests". In: arXiv preprint, (2015)*.

Chapter 5: Deep Learning for Automatic Segmentation

In this chapter, the main ideas behind *artificial neural networks* and theoretical concepts of deep learning using *convolutional neural networks* (CNNs) are described. As the main contribution, a fully automated method using CNNs is proposed for segmentation of polycystic kidneys on CT dataset from patients at different stages of ADPKD. The efficiency of learned features using CNNs for segmenting the complex polycystic kidneys is analyzed and finally, the performance and applicability of this approach for TKV computation in ADPKD is evaluated.

The presented contribution has been published in: *K. Sharma, et al. "Automatic Segmentation of Kidneys using Deep Learning for Total Kidney Volume Quantification in Autosomal Dominant Polycystic Kidney Disease". In: Scientific Reports, Nature (2017)*.

Chapter 6: Conclusion and Outlook

In this section, the contributions of this thesis are summarized and possible directions for further research on segmentation strategies in ADPKD are discussed.

Appendix

A brief overview is provided on additional contributions that have not been discussed in this dissertation.

3 Kidney Volume Measurement in ADPKD

3.1 Role of Total Kidney Volume (TKV) in ADPKD

In ADPKD, sustained development and expansion of bilateral renal cysts is responsible for enlargement of the kidneys. The rate of individual cyst growth and number of cysts in each kidney determines the overall rate of kidney enlargement which is expressed as the change in total kidney volume (TKV). ADPKD patients experience irreversible structural modifications in kidneys starting early in childhood, often extending to the liver over course of time and progressing during lifetime. Despite progressive structural damage, the renal function remains normal for the first few decades which is known to derive from the capacity of each kidney to compensate for the loss of functional nephrons by increasing single nephron filtration rate in the remaining functioning nephrons [51]. Therefore, measurement of GFR for monitoring ADPKD progression is unreliable especially during early phase of the disease. Previous investigations have shown an association between TKV and renal function [25, 40] and several studies have provided evidence for the use of TKV as an important imaging biomarker for assessment of disease severity as well as for predicting disease progression in ADPKD [3, 23, 45, 48]. The European Medicines Agency (EMA) and the Food & Drug Administration (FDA) now acknowledge TKV as prognostic imaging biomarker for use in clinical trials on ADPKD [39, 141]. Several studies on ADPKD investigating response of different biomarkers to new therapies, have provided evidence and supported TKV as an indicator of treatment efficacy in ADPKD. The effect of long acting somatostatin analogue to help in slowing kidney volume and kidney cyst growth has been previously studied [21, 53]. Other studies have used TKV to investigate the effect of sirolimus, an mTOR inhibitor found to inhibit cell proliferation and cysts growth in adult patients with ADPKD and normal renal function or mild to moderate renal insufficiency [94] and, on ADPKD adults with moderate/severe renal insufficiency and CKD stage 3b or 4 [110]. Several clinical studies using TKV to investigate the effect of pharmacological treatments in ADPKD patients have been reported [17, 21, 58, 94, 111, 116, 125, 127, 135, 138, 146], however, a complete review on them remains out of the scope of this dissertation. Previous investigations have also assessed the role of kidney volume fraction comprised of cysts, also known as the total cyst volume (TCV), as a useful indicator of ADPKD progression on Computed Tomography (CT) [122] and Magnetic Resonance Imaging (MRI) [48]. Certain pharmacological treatments that are known to reduce the growth of cysts in polycystic kidneys and help monitoring blood pressure [22] would benefit from investigations on TCV. In addition to TKV and TCV, measuring the change in renal blood flow as a potential surrogate biomarker calculated using phase-contrast MRI has also shown to precede GFR decline, but its application in ADPKD is still at preliminary stage and requires further verification [69, 137].

3.2 Comparison of TKV Measurement Methods

In several clinical studies, rate of GFR decline and changes in TKV are among the primary outcomes for evaluating efficacy of drug treatment on ADPKD patients. Previous studies have suggested that drug treatment in these patients could limit kidney enlargement, thus accurate and reproducible TKV measurements at different time points during drug therapy could provide crucial information about disease progression. Different ADPKD longitudinal studies on patients receiving standard of care have reported a yearly average increase in TKV of 5.3% to 5.7% per year [28, 45, 48, 138] and, this growth is estimated to become less than 3% per year in patients under treatment [21, 94, 138]. Therefore, precise measurements of TKV are necessary to effectively detect small changes over time intervals as short as 6 months or 1 year and to also limit number of patients enrolled in ADPKD clinical studies, thus making them more feasible while remaining significant.

As was previously shown in figure 1.4, the volume and shape of ADPKD kidneys can considerably vary even among patients that have a similar GFR range. Some kidneys adopt regular shape but most patients have markedly irregular shaped kidneys with prominent surface irregularities due to the presence of different sized and shaped cysts. This heterogeneity makes accurate and reliable kidney volume measurement task challenging. In particular, it requires a reliable method which can give reproducible results for each case as well as adaptable to the hetereogneity encountered in different patients. Moreover, if the method is operator-dependent, then TKV computation needs to be performed by an experienced operator aware of peculiarities of polycystic kidneys and surrounding organs affected by ADPKD which might be a confounding factor in accurate kidney segmentation.

So far, the most commonly employed methods for TKV measurement from CT or MRI include whole kidney contouring (hereafter named as planimetry) [111] and Stereology (grid point counting over the kidney) [8]. Both techniques tend to be time consuming, and thus simpler and faster methods such as those using a mid-slice approach [7] or an ellipsoid equation [52, 63] have been recently proposed for quick estimation of TKV. However, it is crucial to determine true precision and accuracy of a method that is adopted for TKV measurement such that it allows detection of small changes lying within yearly average TKV growth in ADPKD (i.e. $< 3\%$ to 5%). Moreover, it is also necessary to consider the reproducibility and time required by such methods for their effective use in clinical studies. Previous works have assessed the validity of a single or few available TKV estimation methods in comparison with either manual planimetry or stereology [7, 8, 13, 52, 63, 126]. So far, there are no comprehensive analyses comparing the precision, accuracy and reproducibility, along with the amount of time required by different methods used for TKV measurement in ADPKD. Such a comparison among different available methods is important to define the adequacy of TKV quantification strategies in clinical investigations that aim to evaluate the effect of drug treatments. In this respect, we compared different methods available for TKV quantification on CT and MR images within clinical studies on ADPKD. The methods were evaluated for reproducibility, accuracy, precision, and time requirement. Additionally, the influence of expertise required by each method and the sensitivity of these methods to detect "between-treatment" group difference in TKV change over one year was studied. The results help in identifying the most suitable kidney volume measurement method for clinical studies evaluating treatment efficacy on ADPKD progression.

	Experimental dataset		Validation dataset
	MR	CT	MR
number of acquisitions	15	15	75
Clinical Study	EuroCYST	SIRENA 2 (n=5) ALADIN 2 (n=10)	ALADIN
Age (years)	49 [38-62]	51 [35-67]	37 [20-63]
Gender (females)	7 (47%)	4 (27%)	39 (52%)
GFR (ml/min per $1.73m^2$)	62 [31-114]$^\sharp$	22 [10-35]	84 [32-137]§
Left KV (ml)	1,474 [365-3,061]	1,558 [335-3,184]	971 [186-2,634]
Right KV (ml)	1,366 [308-3,544]	1,596 [263-3,256]	877 [169-3,317]
Total KV (ml)	2,840 [707-6,605]	3,154 [598-6,002]	1,855 [404-5,577]

Tab. 3.1. **Demographic and Renal Function Parameters.** Demographics and clinical characteristics of ADPKD patients included in the experimental and validation datasets from past and on-going clinical trials. $\sharp$ missing data for $n = 3$ patients; $\S$ missing data for $n = 2$ patients. Note: All values in table are expressed as mean [range] or absolute numbers (%).

3.2.1 Patient Dataset

For the main experiment, 15 MR images from baseline examinations of 15 ADPKD patients enrolled in the EuroCYST study [96], a multi-centre longitudinal observational study on ADPKD progression in patients with estimated $GFR \geq 30ml/min$ per $1.73m^2$ (clinicaltrials.gov identifier NCT02187432) were used. These MRI exams were acquired according to the EuroCYST MRI acquisition protocol [96], including standard localizer, T2 single shot fast/turbo spin echo (coronal acquisition, 4 mm slice thickness, 0 mm spacing, $FOV = 30-35$ to avoid wrap-around, 256×256 matrix, $TE \approx 70-190$ ms based on the vendor and max TR), FISP or FIESTA 3D spoiled gradient echo (coronal acquisition, 4 mm slice thickness, 0 mm spacing, $FOV = 30-35$, 256×256 matrix, $TE \approx 2ms$, $TR \approx 7ms$, $flip-angle = 40-50°$), and T1-3D spoiled gradient echo (coronal acquisition, slice thickness of 4mm, spacing 0mm, $FOV = 30-35$, 256×256 matrix, $TE \approx 2ms$, $TR \approx 4ms$, $flip-angle \leq 15°$). Once acquired, MR images were transferred to DICOM 16-bit format from the clinical scanner on digital media, and 3D-T1 MRI sequences were used for KV computation. All the 3D-T1 MR images included in this study (n=15) were acquired at six different centres of the EuroCYST study and selected to uniformly represent large range of single KV range *(from 707 to 6,605 ml)* and different image quality.

Additionally, 15 CT images were acquired on ADPKD patients with estimated $GFR \leq 40ml/min$ per $1.73m^2$, enrolled in either ALADIN 2 (clinicaltrials.gov identifier NCT01377246) or SIRENA 2 [110] (clinicaltrials.gov identifier NCT01223755) clinical trials. These CTs were acquired in a single breath-hold scan (120 kV; 150 to 500 mAs; matrix 512×512; 2.5 mm collimation; 0.984 slice pitch; 2.5 mm increment). Each CT acquisition was initiated 80 seconds after the infusion of 100 ml non-ionic iodinated contrast agent (Iomeron 350; Bracco, Italy) at a rate of 2 ml/s, followed by 20 ml saline solution at the same infusion rate. Once acquired, CT images were transferred in DICOM 16-bit format from the clinical scanner on digital media, and resampled to 5 mm slice thickness for KV computation. The CT acquisitions used in this study (n=5 from SIRENA 2 and n=10 from ALADIN 2 clinical studies) were taken from different centres and were selected to uniformly represent large single KV range

(from 598 to 6,002 ml) and different image quality. The main socio-demographic and clinical features of ADPKD patients used in the experiments are reported in table 3.1.

3.2.2 Experimental Setup and Methods

In this experiment, we compared different methods available for TKV quantification in terms of reproducibility, accuracy, precision, and time required, on a series of MR and CT acquisitions obtained within two clinical studies on ADPKD. On the acquired MR and CT images, two independent operators with different level of experience quantified volume of single kidneys (SKV) on left and right kidneys, separately. The expert operator (KS for both MR and CT) routinely performed KV computations for ADPKD clinical trials for two years, acquiring experience on all different techniques used for our experiments, while the beginner operators (KP for MR and LVQ for CT) started performing KV computation for the purposes of the current experiment, after specific training on kidney anatomy and different computational methods. Same protocol was used by expert and beginner operators to measure SKV, which defined the kidney border at the main renal blood vessels and hilum junction using a perpendicular line to separate the kidney. Fat and vessels lying inside the kidney were included inside the kidney outline, while any abdominal fat surrounding the kidney was excluded. Special attention was paid to separate regions where kidneys and liver were adjacent. Each operator computed SKV twice, and at least two weeks apart, to eliminate potential memory from first set of measurements. The SKV (i.e. right and left) was computed on 30 kidneys for both MRI and CT dataset using six different methods described below. Additionally, the expert operator also measured the kidney length.

Method 1: Polyline Manual Tracing

In this method, the kidney contour was manually segmented using the polygon tool in ImageJ software [34] (NIH, Bethesda, MD) and has been referred to as “ImageJ polyline” method. For an accurate measure of SKV, each kidney was manually outlined by drawing a polyline composed of numerous points on all contiguous slices and then SKV was finally computed as the sum of the surface area of all the kidney outlines, multiplied by the slice thickness. The theoretical hypothesis on the accuracy of planimetry in quantifying the volume of an object with ellipsoidal shape, based on the area of serial sections, is dependent on section thickness and orientation with regard to the object size. To estimate the volume quantification error caused by sectioning, three ellipsoids of different sizes were considered, and planimetry was applied to these ellipsoids using randomly positioned but uniformly distributed serial sections of thickness and orientation typically found in MR and CT imaging. As shown in figure A.1 (refer: *Appendix: A*), the volume quantification error of theoretical planimetry, in comparison with analytical volume, is less than 0.26% and 0.10% for MR and CT sections, respectively. Based on these results, manual segmentation of kidneys on serial sections by polylines were considered to represent as the *reference method* for KV computation for our experiments.

Method 2: Free-hand manual tracing

The Free-hand manual tracing has been suggested to provide faster means of kidney segmentation. It does not require placement of consecutive points on the kidney border, and each kidney is traced by a free-hand drawing tool by outlining all contiguous kidney slices using platform dependent Osirix imaging processing software [109]. SKV was computed similarly to

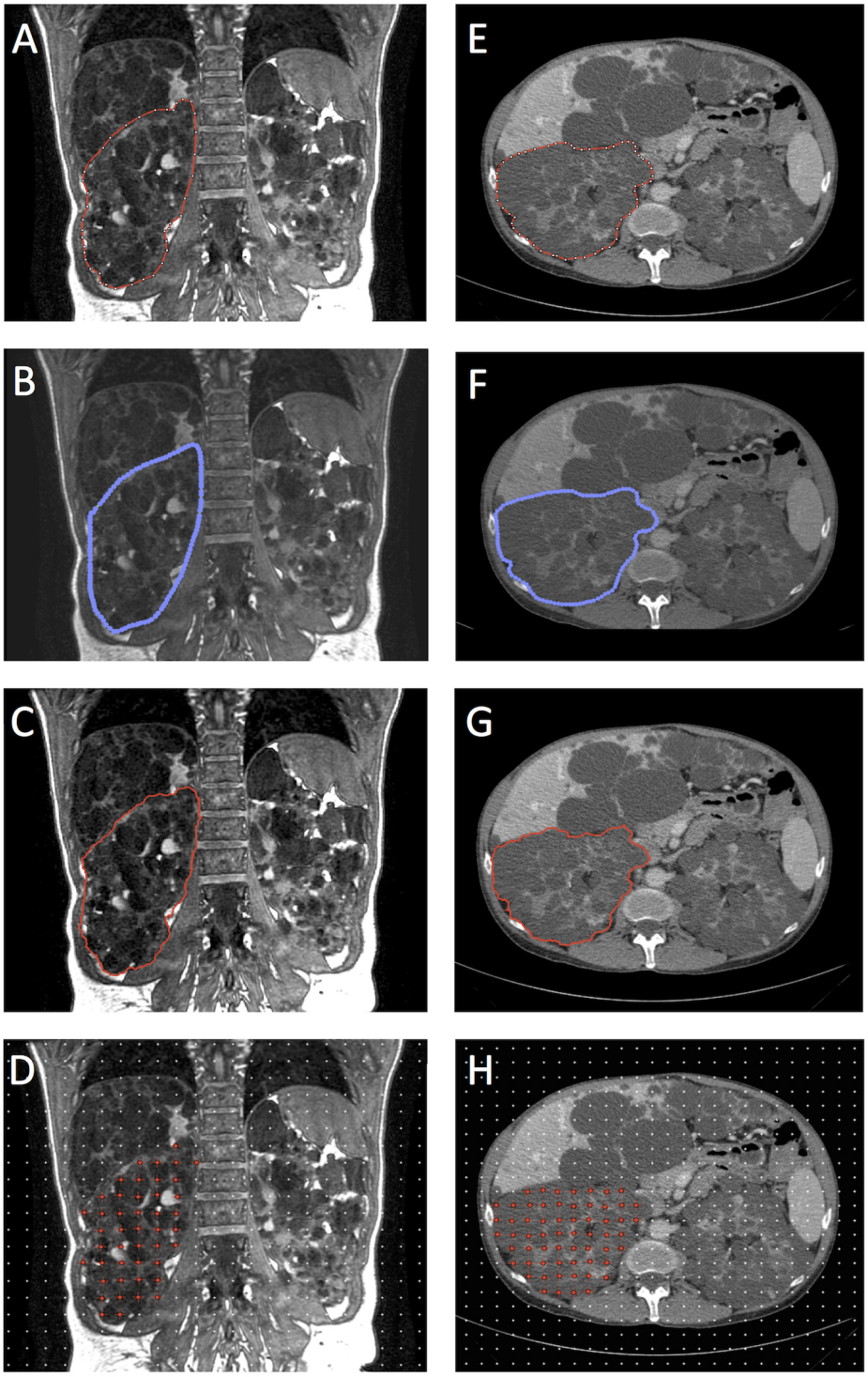


Fig. 3.1. **Representative images of polycystic kidney volume segmentations.** Representative images of polycystic kidney volume segmentations. Segmentation were performed on MRI (panels A-D) and CT image slices (panels E-H) by the expert operator using ImageJ polyline (A and E), Osirix free-hand (B and F), Livewire tool (C and G) and Stereology (D and H).

the ImageJ polyline method i.e. as the collective sum of surface area of all the kidney outlines, multiplied by slice thickness. This planimetry based method has been referred to as “Osirix free-hand”.

Method 3: Semi-automatic tracing

This semi-automatic outline tool was designed (in-house) to reduce KV quantification time. Initially customized from a plugin in ImageJ software based on the livewire segmentation (ivussnakes.sourceforge.net)[9], this tool utlizes canny-edge detection for detecting polycystic kidney outlines on all contiguous slices. Starting from a manually selected seed point, the Livewire tool automatically identifies the kidney boundary while the operator moves the mouse over the region of interest. The tool automatically recognizes correct boundary segments, and the operator places a new seed point to confirm the selection and new seed points are placed until the kidney has been completely segmented. Then SKV is also computed similarly to the above two planimetry based methods i.e. as sum of areas of the kidney outlines, multiplied by the slice thickness. This method has been referred to as the “Livewire tool”.

Method 4: Stereology

For stereology, each kidney section was extracted by counting the number of intersections of a randomly positioned grid over continous slices [8]. Stereology was performed with the ImageJ Grid plugin (rsb.info.nih.gov/ij/plugins/grid.html), using a grid comprising of crosses placed on a 3D stack with 16×16 mm spacing, 16 mm slice thickness for MR images and 15×15 mm spacing, 15 mm slice thickness for CT images. A random offset was used for grid position and the spacing was set empirically in order to reduce the time required while maintaining high accuracy. SKV was computed as point count, multiplied by grid square area and by slice thickness. Representative images of planimetry methods and Stereology, on both MR and CT images, are shown in figure 3.1.

Method 5: Mid-slice Method

We used a simplified method [7] to estimate SKV using a single slice obtained from the mid-section of left and right kidneys, separately. Thus, each kidney was outlined only on this mid-slice by manually drawing a polyline, and the SKV was estimated by multiplying the mid-slice area with total number of slices containing the kidney sections, the slice thickness, and an empirically computed factor (0.637 for right kidney, 0.624 for left kidney) [7] as shown below:

$$SKV_{right} = midSlice\,area \times slice\,thickness \times total\,slices \times 0.637, \tag{3.1}$$

$$SKV_{left} = midSlice\,area \times slice\,thickness \times total\,slices \times 0.624. \tag{3.2}$$

Method 6: Ellipsoid Equation

The ellipsoid method, has been mainly used to estimate SKV for classification purpose [63]. For left and right kidneys separately, the length (in both sagittal and coronal orientation),

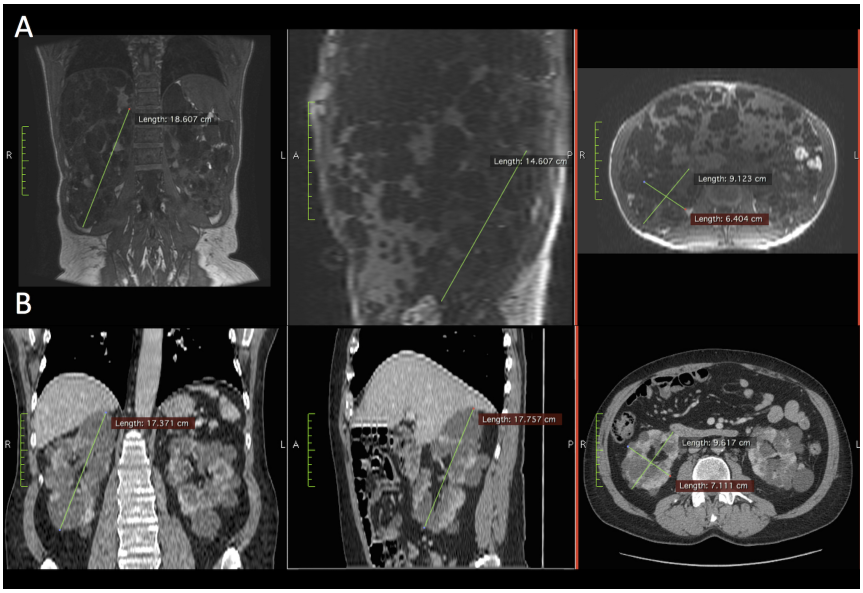


Fig. 3.2. Example single kidney volume (SKV) assessment using the Ellipsoid method. SKV assessment was performed by the expert tracer on MRI (panel A, left to right: coronal, sagittal, and axial view) and CT (panel B, left to right: coronal, sagittal, and axial view). Kidney length was assessed on both coronal and sagittal planes, while kidney depth and width were assessed on axial plane.

width and depth (in axial orientation) are measured and then SKV is estimated using the following ellipsoid formula:

$$SKV_{ellipsoid} = \frac{\pi}{6} \times length \times width \times depth, \tag{3.3}$$

where, length is the average of sagittal and coronal lengths. An example SKV measurement by ellipsoid method has been depicted in figure 3.2.

Validation Study

For selecting the reference method, a *validation study* was performed using baseline and (1-year) follow-up MR images from a separate clinical investigation [21] (ALADIN study: clinicaltrials.gov identifier NCT00309283) The validation study assessed the sensitivity of individual TKV quantification methods in detecting TKV change over 1-year period between two treatment groups was performed on a separate set of MR images from an independent clinical study [21] not utilized for the main experiments. The main socio-demographic and clinical features of patients from the ALADIN trial included in the validation dataset (n=75) are reported in table 3.1. In the original investigation of ALADIN study, TKV was computed using ImageJ polyline method. In the current validation study, the same TKV measurements were repeated using additional methods i.e. Stereology, Mid-slice and Ellipsoid method, as well as the right and left kidney length were computed. To investigate the efficacy of TKV quantification methods for detecting small changes developing over short time intervals, it was sufficient to include only 1-year follow-up dataset for the purpose of validation. We compared the sensitivity of each TKV method for detecting the difference in TKV change over 1-year period between two treatment groups. Based on results of these computed TKV changes captured by each method, we also assessed the size of patient population (sample size) that

would be necessary to use by each TKV quantification method for detecting a significant difference between two treatment groups in the same timeframe of 1-year.

3.2.3 Statistical Analyses

The statistical analyses were performed using R software (www.r-project.org), version 3.2.0. The reproducibility (intra-rater reliability for both expert and beginner operator) and inter-rater reliability were assessed by coefficients of variation (CVs) for repeated measures [66]. For MR and CT images separately, the agreement between SKV computed using different methods (within operator, first tracing) was assessed using Bland-Altman plots. The significance of the difference in SKV values and in time required for SKV computation (within operator, first tracing) was assessed by ANOVA (treatment by subjects design), followed by Tukey's honest significant difference post-hoc test. The same analysis was repeated to assess the significance of the difference in SKV values and time between first and second tracing (within operator). Root mean squared error (RMSE) was used to measure the mean difference between SKV computed by individual computation methods and the reference method (ImageJ polyline). The correlation between SKV and length was assessed using Pearson's correlation coefficient. In the validation study, for each KV quantification method, the difference in TKV change between treatment group at 1 year were assessed by ANCOVA, adjusted for baseline measurement. The difference in TKV percentage change was assessed by unpaired t-test. The sample size was computed as minimum size required to assess a significant difference between the two treatment groups based on the mean of the two treatment groups and the standard deviation of the Octreotide-LAR treatment group, assuming type I $error = 0.05$, and $power = 0.80$.

3.2.4 Results

For our experiments, six TKV computation methods were assessed for reproducibility, accuracy, precision, and time requirement. The SKV measurements using all six methods from expert and beginner operators were compared in terms of intra- and inter-operater agreement, which is particularly important for clinical studies. Descriptive statistics of SKV measurements using each method for the two operators during first and second tracings of 15 MRI and 15 CT acquisitions in the experimental dataset have been shown in table 3.2. Moreover, Intra-operator and inter-operator differences in estimating SKV with all methods are shown in table 3.3.

Time Requirement Analysis

The ImageJ polyline method required the longest time (table 3.2) with more than 30 minutes on average, on both MR and CT, and for both expert and beginner operators while Osirix free-hand was the fastest among planimetry methods for both operators, reducing the mean time to only 20 and 16 min on MR and CT, respectively. The Livewire tool required relatively shorter time than polyline for imaging modalities, but with a reduction of only 8 min on average. Stereology required shorter time, with an average of 11 and 14 minutes for MR and CT, respectively. The simplified methods (Mid-slice and Ellipsoid equation) were the fastest and required the shortest time of approximately 10 and 5 minutes, respectively. As anticipated, the time required for each KV computation using any of the above methods was consistently higher for the beginner operator (table 3.2) compared to the expert operator.

	SKV computation methods	Expert			Beginner		
		1st tracing SKV (ml)	2nd tracing SKV (ml)	Time (min)	1st tracing SKV (ml)	2nd tracing SKV (ml)	Time (min)
MRI	ImageJ Polyline	1420 ± 989	1424 ± 988	35 ± 12	1448 ± 998	1460 ± 993	40 ± 14
	Osirix free-hand	1401 ± 975	1410 ± 982	21 ± 09**	1425 ± 974	1419 ± 980	22 ± 08**
	Livewire tool	1410 ± 996	1418 ± 991	26 ± 09*	1445 ± 1009	1438 ± 993	33 ± 15
	Stereology	1373 ± 956	1365 ± 965	15 ± 09**	1407 ± 977	1391 ± 966	19 ± 16**
	Mid-slice	1401 ± 982	1421 ± 996	09 ± 01**	1462 ± 987	1476 ± 1019	11 ± 01**
	Ellipsoid	1207 ± 940**	1230 ± 970**	05 ± 01**	1207 ± 1016**	1308 ± 984*	06 ± 0**
CT	ImageJ Polyline	1577 ± 921	1572 ± 923	31 ± 11	1584 ± 931	1597 ± 936	52 ± 18
	Osirix free-hand	1579 ± 924	1587 ± 930	18 ± 07**	1572 ± 924	1572 ± 925	28 ± 10**
	Livewire tool	1551 ± 916	1541 ± 910	24 ± 08*	1563 ± 922	1558 ± 913	32 ± 11**
	Stereology	1566 ± 919	1550 ± 908	17 ± 09**	1653 ± 973	1661 ± 968	26 ± 14**
	Mid-slice	1387 ± 827**	1368 ± 833**	11 ± 01**	1368 ± 828**	1395 ± 849**	11 ± 01**
	Ellipsoid	1450 ± 909**	1480 ± 900**	04 ± 01**	1329 ± 849**	1322 ± 819**	05 ± 01**

Tab. 3.2. **Computed Single kidney volumes (SKV) and respective time required by different methods.** SKV obtained by expert and beginner operators on MR and CT images from ADPKD patients in the experimental dataset. ** p<0.001 and * p<0.05 at Tukey's honest significant difference post-hoc test (individual methods vs reference ImageJ polyline method). Number of single kidneys analyzed $n = 30$ for MR and $n = 30$ for CT. Single Kidney volumes (SKV) are expressed as mean ± SD. SKV (ml) were computed by both operators (expert and beginner), two weeks apart (1st tracing vs. 2nd tracing). Time (min) was estimated on total kidney volumes (sum of right and left SKV).

Intra- and Inter-rater Agreement Analysis

The highest intra- and inter-rater agreement for both MR and CT images were recorded for the planimetry methods (i.e. ImageJ Polyline, Osirix free-hand and Livewire tool) and Stereology. In particular, for MR, the CV was consistently lower for the expert operator compared to the beginner while for CT, both expert and beginner operator had similar CV, suggesting that reliable identification of kidney contour by a non-expert operator is easier on CT than on MRI. For the expert operator, the planimetry methods were more reproducible than Stereology and Mid-slice, on both imaging modalities MRI and CT, while the Ellipsoid method had lowest reproducibility. On evaluation of the performance of beginner operator, the Livewire tool was observed to be the most reproducible while the Ellipsoid method was the least reproducible method. There was no significant intra-operater variability in terms of computed SKV values as well as time requirement between between first and second tracings. As shown in table 3.3, the inter-rater performance on MR was worse than CT in general. On MR, the ImageJ polyline and Livewire tool had the lowest CV, followed by Stereology and Osirix free-hand. The inter-rater performance on CT was less variable and the method with the lowest CV was Osirix free-hand, followed by the Livewire tool, ImageJ polyline, Mid-slice method, and Stereology. The Ellipsoid method had the lowest inter-rater reproducibility on both MR and CT (table 3.3) and no consistent difference was found between left and right kidneys in terms of reproducibility. The intra-operator and inter-operator differences in estimating SKV using all the described methods have been summarized in table 3.3.

Method Performance Analysis

On MRI dataset, ImageJ polyline which has been adopted as the reference method in our experiments showed the highest agreement with Osirix free-hand for both expert and beginner operators (table 3.4). The accuracy of the Osirix method was high with a reported mean difference of only -0.8% and an estimated precision (percentage root mean square error, RMSE) of 3.2%. The Livewire tool also showed high accuracy and precision but lower than Osirix free-hand and followed by Stereology with reported mean difference of -3.7% indicating lower accuracy and less precision reported as percentage RMSE of 6.3%. The agreement between these different methods was evaluated using Bland-Altman plots, as shown in figure 3.3 and figure 3.4. Comparing the simplified methods with polyline method, it was observed that both mid-slice and ellipsoid equation resulted in the lowest accuracy and precision (table 3.4 and figure 3.3) and the difference in the computed SKV between Ellipsoid method and ImageJ polyline (mean of -18.8%) was found to be statistically significant ($p < 0.01$). At single kidney level, the kidney length showed an expected positive and significant correlation with SKV (ImageJ polyline, $r = 0.91$; $p < 0.01$ and $r = 0.90$; $p < 0.01$ on MR and CT, respectively). However, the residual plot of the correlation demonstrated that kidney length was not precise enough for use in clinical studies, for both imaging modalities (see figure 3.3 and figure 3.4). Additionally, no consistent difference was observed on the agreement analysis between different methods in terms of using the right or left kidney.

Validation Study Analysis

The results of validation study have been summarized in table 3.5. In the ALADIN study [21], both absolute and percentage changes in TKV at 1 year of treatment were computed using ImageJ polyline method. These TKV changes (both absolute and percentage) were significantly different for ADPKD patients under Octreotide-LAR (n=38) in comparison to

	SKV computation methods	Intra-rater (Expert)			Intra-rater (Beginner)			Inter-rater		
		SKV Difference (ml)	SKV Difference (%)	CV (%)	SKV Difference (ml)	SKV Difference (%)	CV (%)	SKV Difference (ml)	SKV Difference (%)	CV (%)
MRI	ImageJ Polyline	-04 ± 24	-0.4 ± 2.4	1.18	-12 ± 84	-3.3 ± 18.3	4.04	-28 ± 50	-2.5 ± 4.7	2.77
	Osirix free-hand	-09 ± 24	-0.3 ± 2.5	1.26	07 ± 116	0.8 ± 7.4	5.70	-24 ± 97	-3.1 ± 7.1	4.92
	Livewire tool	-08 ± 24	-1.3 ± 3.4	1.25	07 ± 47	-0.5 ± 3.7	2.28	-35 ± 46	-2.8 ± 4.6	2.83
	Stereology	07 ± 55	1.3 ± 5.5	2.80	16 ± 82	0.2 ± 11	4.13	-34 ± 89	-2.9 ± 14.6	4.75
	Mid-slice	-20 ± 60	-2.2 ± 7.7	3.11	-13 ± 115	-0.8 ± 5.1	5.48	-62 ± 252	-11.3 ± 29.8	12.61
	Ellipsoid	-23 ± 146	-3.8 ± 25.3	8.46	-101 ± 367	-19.1 ± 46	21.06	-0.79 ± 242	0.1 ± 31.9	17.73
CT	ImageJ Polyline	06 ± 25	0.6 ± 1.3	1.14	-13 ± 31	-1.2 ± 1.9	1.48	-07 ± 31	-0.03 ± 2.0	1.38
	Osirix free-hand	-08 ± 19	-0.5 ± 1.3	0.89	-0.1 ± 27	-0.3 ± 2.0	1.19	07 ± 20	0.7 ± 1.6	0.95
	Livewire tool	10 ± 21	0.7 ± 1.1	1.05	05 ± 19	-0.2 ± 1.9	0.88	-12 ± 25	-0.8 ± 1.6	1.23
	Stereology	15 ± 35	0.4 ± 2.9	1.69	-09 ± 24	-1.2 ± 3.0	1.07	-87 ± 66	-5.6 ± 3.9	4.77
	Mid-slice	19 ± 117	1.6 ± 6.3	6.00	-26 ± 58	- 2.0 ± 3.6	3.21	18 ± 73	1.8 ± 4.6	3.79
	Ellipsoid	-31 ± 193	-4.1 ± 16.9	9.28	07 ± 185	0.6 ± 12.3	9.70	121 ± 287	5.9 ± 18.0	15.65

Tab. 3.3. **Inter and intra-rater reproducibility of single kidney volumes (SKV).** SKV measured by expert and beginner operators using different quantification methods on MR and CT images from ADPKD patients in the experimental dataset. Number of single kidneys analyzed $n = 30$ for MR and $n = 30$ for CT. Single Kidney Volume (SKV) difference is expressed as mean ± SD. SKV Difference (ml) = Absolute difference between 1st and 2nd tracing; SKV Difference (%) = Percentage difference between 1st and 2nd tracing; CV = coefficient of variation for repeated measures.

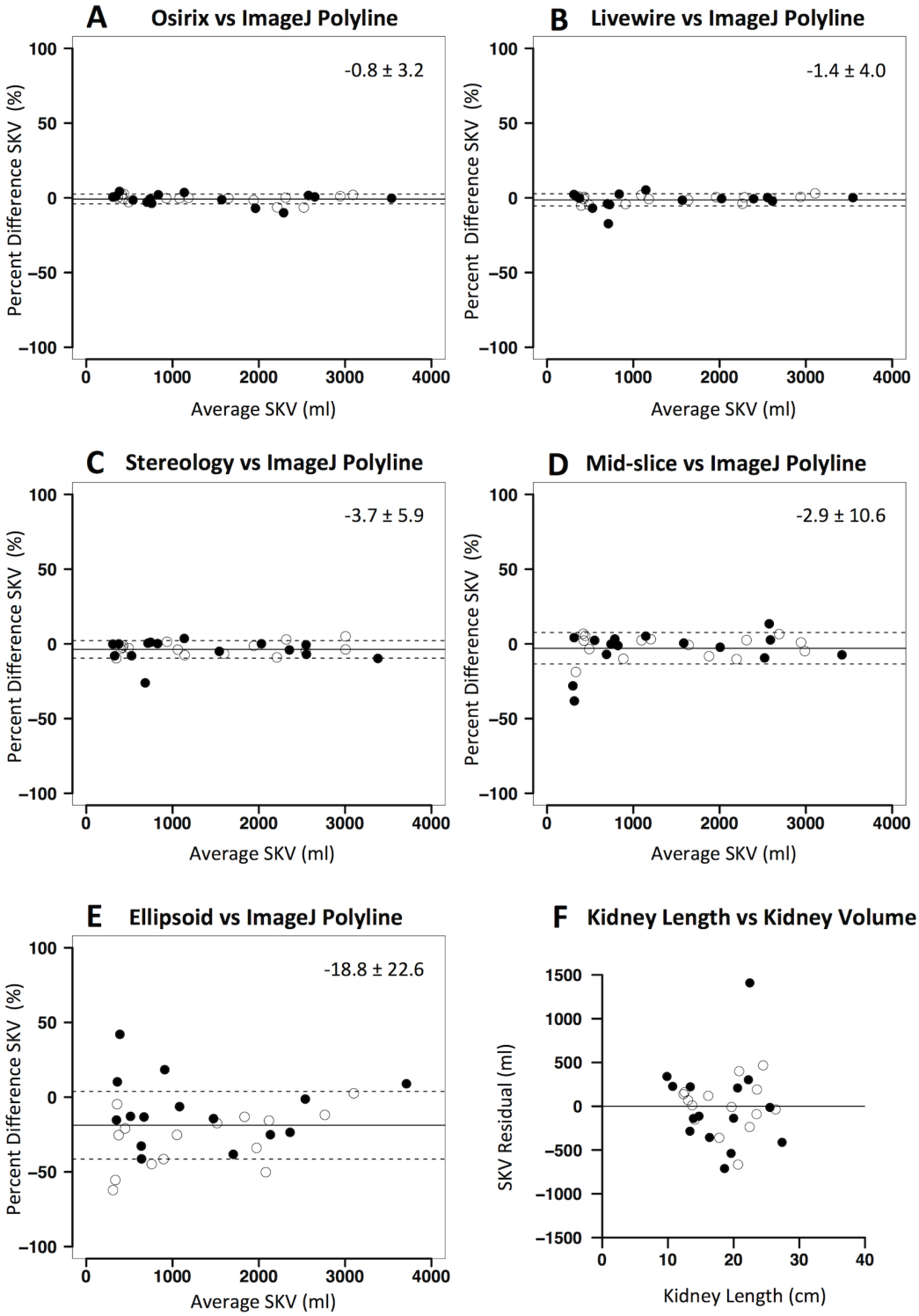


Fig. 3.3. **Agreement between kidney volume computation methods on MRI in the experimental dataset.** Panels A-E: Bland-Altman plots showing agreement between different kidney volume computation methods (A: Osirix free-hand; B: Livewire tool; C: Stereology; D: Mid-slice method; E: Ellipsoid method) versus ImageJ polyline (reference method). Percent differences in single kidney volume (SKV) are plotted against average SKV values of the two methods. Solid lines denote mean difference, while dashed lines denote ± standard deviations. Panel F: plot of the residual of the linear regression of kidney length against SKV (assessed by reference ImageJ polyline method). Black dots represent right kidneys while white dots represent left kidneys.

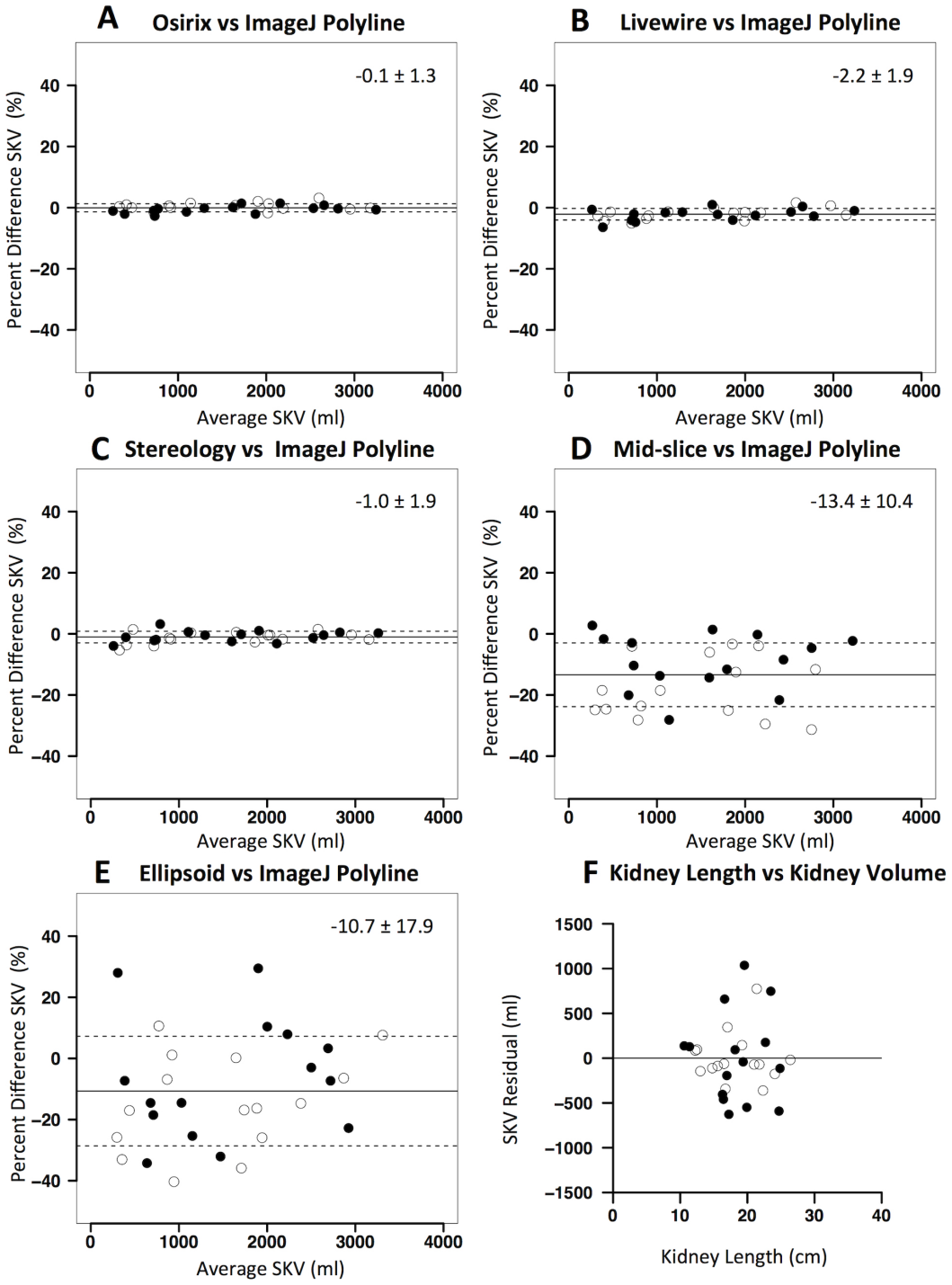


Fig. 3.4. **Agreement between kidney volume computation methods on CT in the experimental dataset.** Panels A-E: Bland-Altman plots showing agreement different kidney volume computation methods (A: Osirix free-hand; B: Livewire tool; C: Stereology; D: Mid-slice method; E: Ellipsoid method) versus ImageJ polyline (reference method). Percent differences in single kidney volume (SKV) are plotted against average SKV values of the two methods. Solid lines denote mean difference, while dashed lines denote ± standard deviations. Panel F: plot of the residual of the linear regression of kidney length against SKV (assessed by reference ImageJ polyline method). Black dots represent right kidneys while white dots represent left kidneys.

	SKV computation method	SKV Difference (ml)	RMSE (ml)	SKV Difference (%)	RMSE (%)
MRI	Osirix free-hand vs ImageJ Polyline	-19 [-229, 60]	66	-0.8 [-9.5, 4.4]	3.2
	Livewire tool vs ImageJ Polyline	-10 [-123, 93]	40	-1.4 [-15.9, 5.4]	4.0
	Stereology vs ImageJ Polyline	-47 [-329, 152]	101	-3.7 [-23.1, 5.2]	6.3
	Mid-slice vs ImageJ Polyline	-19 [-251, 346]	118	-2.9 [-32.1, 14.4]	9.7
	Ellipsoid vs ImageJ Polyline	-213 [-1044, 333]	350	-18.8 [-48.0, 53.2]	25.4
CT	Osirix free-hand vs ImageJ Polyline	02 [-40, 82]	23	-0.1 [-2.7, 3.2]	1.3
	Livewire tool vs ImageJ Polyline	-26 [-88, 42]	39	-2.2 [-6.2, 1.6]	2.8
	Stereology vs ImageJ Polyline	-11 [-67, 39]	26	-1.0 [-5.2, 3.2]	2.1
	Mid-slice vs ImageJ Polyline	-190 [-862, 23]	276	-13.4 [-27.1, 2.8]	15.0
	Ellipsoid vs ImageJ Polyline	-128 [-666, 559]	289	-10.7 [-33.6, 34.5]	19.0

Tab. 3.4. Absolute and percentage difference and root mean squared error (RMSE) between methods used to compute single kidney volume (SKV) by the expert operator on MR and CT images from ADPKD patients in the experimental dataset. Number of single kidneys analyzed $n = 30$ for MR and $n = 30$ for CT. SKV are from first tracing of expert operator. SKV difference is expressed as mean difference and [range]; RMSE = Root mean square error.

the placebo treatment group (n=37) (p=0.032 and p=0.003, respectively). This difference was also statistically significant between treatment-placebo group when TKV measurements were repeated using the Stereology technique(absolute and percentage change: p=0.018 and p=0.016, respectively). However, when TKV was estimated on treatment and placebo group using Mid-slice method or Ellipsoid equation, the difference in their respective TKV measurements was observed to be not statistically significant, as shown in table 3.5. Similarly, between-treatment difference in kidney length (computed as sum of right and left kidney lengths) was also not statistically significant (table 3.5). The above results demonstrate that simplified methods cannot capture between-treatment changes occurring over the course of 1 year time. Based on the statistics of TKV percentage changes, the sample size required by different TKV measurement methods to identify significant difference between two treatment groups is lowest for ImageJ polyline (n=34 per treatment group), followed by Stereology (n=59). The estimated sample size increases up to 4-fold for TKV measurements from Mid-slice method (n=135) or when using Ellipsoid equation (n=147).

3.2.5 Conclusion

Different methods for KV computation were evaluated in terms of reproducibility, accuracy, precision and time required on both MR and CT representative images. The dataset in the main experiment consisted of 30 kidneys each from MRI and CT scans, with a wide range of SKV. Overall, planimetry methods and stereology showed the highest reproducibility, low bias, and desired accuracy and precision. However, the reproducibility of planimetry and stereology was inferior on MR than CT dataset, likely attributed by lower image quality on MR compared to CT, making kidney identification on MR further operator-dependent. High intra-rater variability was reported for the beginner operator suggesting that KV computation on MR needs to be performed by expert operators, to reliably detect KV changes. On MRI, highest accuracy and precision were observed for planimetry based methods, while on CT stereology performed equally well which might again be attributed by higher image quality

KV computation method	Absolute change in TKV (ml)			Percentage change in TKV (%)		
	Octreotide-LAR (n=38)	Placebo (n=37)	p	Octreotide-LAR (n=38)	Placebo (n=37)	p
ImageJ Polyline	46.1 ± 112.3	143.7 ± 158.1	0.032 (<0.05)	2.57 ± 6.07	6.72 ± 5.89	0.003 (<0.01)
Stereology	45.8 ± 114.1	152.1 ± 160.4	0.018 (<0.05)	3.30 ± 7.14	7.00 ± 5.83	0.016 (<0.05)
Mid-slice	40.1 ± 129.8	127.2 ± 186.0	0.111 (NS)	2.89 ± 10.71	6.55 ± 8.06	0.098 (NS)
Ellipsoid	36.3 ± 153.9	125.4 ± 179.1	0.102 (NS)	2.54 ± 11.66	6.35 ± 9.99	0.132 (NS)
Kidney length♯	-0.25 ± 1.22	0.26 ± 1.97	0.115 (NS)	-0.69 ± 3.97	1.10 ± 6.69	0.165 (NS)

Tab. 3.5. **Validation Study: Total kidney volume changes compared with baseline at 1 year of treatment with placebo or Octreotide-LAR.** Total kidney volume was assessed by different kidney volume computation methods on MR images taken from the ALADIN clinical study. Abbreviations: LAR, long-acting release; KV, kidney volume; NS, not statistically significant; TKV, total kidney volume (sum of right and left kidney volumes). ♯ Kidney length (in cm) is computed as sum of right and left kidney lengths. *p* values from ANCOVA (absolute change) or unpaired t-test (percentage change).

and more number of axial sections compared to MRI. The mid-slice method and ellipsoid equation, despite providing quick KV estimates, were less reproducible and showed lowest precision and accuracy on both MR and CT images.

Our work and previously reported investigations provide evidence that both mid-slice and ellipsoid equation cannot detect KV changes in the range of 3 to 5% due to much lower precision ranging between 10 and 25% (i.e. SD of the difference between KV calculated by these methods and the reference method). The validation experiment in our work also showed that these simplified methods are not precise enough to be utilized in clinical studies for capturing between-treatment changes in TKV that might develop over one-year treatment period. Moreover, owing to the high variability in estimating TKV, both mid-slice and ellipsoid methods require approximately 4-fold larger sample size than ImageJ polyline to capture significant difference between TKV changes in the two treatment groups. The results also show that stereology allows detection of difference in TKV between the treated and control groups. Other than SKV measurements, kidney length is of interest since it can be easily computed on ultrasound investigations. It has been recently proposed as predictor of disease progression [13] and shows linear correlation with kidney volume, assessed on either MR or CT. However, the correlation is accompanied with very low precision and therefore, kidney length may be restricted to be used only for rough estimations of TKV. Our validation study also shows that kidney length is not accurate as desired to identify between-treatment changes, suggesting that it should not be recommended as outcome measure for clinical trials.

Despite having advantages of precision and accuracy, planimetry methods require 20 to 40 minutes on average for SKV measurement (21 to 35 min for expert operators, for two kidneys). Stereology reduces this average time for SKV measurement to 15 to 17 minutes and time required is reduced to great extent by the simplified methods (5 to 10 minutes approximately), but at the cost of reduced precision and accuracy of SKV measurements. To overcome time requirement and operator-dependency limiting the planimetry methods, it would be ideal to use completely automated approaches.

The limitations of manual segmentation and stereology for efficient TKV computation in clinical studies, provide good motivation for investigating novel strategies to improve segmentation of polycystic kidneys from acquired imaging (CT or MR) dataset. Some attempts to develop automatic segmentation tools have been reported in literature and also described in chapter 2, but achieving desired accuracy and precision required for clinical studies is a challenging task. In the next chapters, we describe two different machine learning methods based on *random forests* and *deep convolutional neural networks*, respectively for segmentation of polycystic kidneys from CT dataset of ADPKD patients. We show that by formulating the segmentation task into a pattern-recognition problem and training an efficient classification model, it is possible to identify complex patterns within the data, thereby facilitating fast and reproducible segmentation for TKV measurement in ADPKD.

Part II

Machine Learning based Approaches for Segmentation

> *The original question, "Can machines think?" I believe to be too meaningless to deserve discussion.*
>
> **— Alan Turing**
> ("Computing Machinery and Intelligence" - Mind 59 (1950): 433-460)

Random Forests for Segmentation

4

4.1 Introduction

Random Forests, or more generally *Decision Forests* are a popular ensemble learning method that have been successfully applied to a number of computer vision, machine learning, and medical image analysis tasks. One of the initial works on *decision trees* by Breiman et al.[19] describing classification and regression trees (CART) strongly influenced later developments in this field. *Decision Trees* are directed acyclic graphs consisting of a hierarchy of feature learners in an ensemble of a decision model. They use predictive modelling for making probabilistic decisions in machine learning applications. Decision trees became popular because they are computationally inexpensive, allowing fast model construction which can be also be used on very large training datasets, and can be devised to take into account the uncertainty in a probabilistic function. One of the most popular algorithms for training optimal decision trees is the *C4.5* by Quinlan [101]. For growing a decision tree, heuristic-based approaches are used to guide the decision tree algorithm in the vast hypothesis space. However, solely learning an optimal decision tree is known to be an NP-complete problem [77], that can lead to complex models which do not generalize well due to overfitting of the training dataset.
Based on the ensemble learning, weak decision trees, also known as the *Random Forests* were constituted aiming to optimize a single complex tree. They consist of an ensemble of independent *decision trees* following a divide and conquer strategy in a probablistic framework to solve regression, classification or clustering based tasks. Random Forests can achieve better generalization by averaging their predictions in a learning process over de-correlated trees. T. K. Ho [56] first introduced random decision forests for handwritten digit recognition. In subsequent work [57], random forests were shown to yield superior generalization compared to both boosting and pruned C4.5 trained decision trees. In another approach, by introducing randomness during the learning process, also known as *bagging*, it was possible to train independent trees with a random subset of the training data [18] . Random forests have since been used for several tasks including regression, classification, semi-supervised and/or manifold learning in both medical and general applications.

4.2 Decision Trees

A decision tree can be constructed using an efficient tree induction algorithm that usually employs a greedy strategy based on a series of locally optimal decisions about which attribute to use for partitioning the data and growing the tree. Following a "divide" and "conquer" approach, a decision tree defines different types of nodes based on their location in the tree. As shown in figure 4.1, starting from a *root node* which has no incoming edges and zero or more outgoing edges, a test condition is placed on it. Any branch with a satisfying outcome to this test condition leads to either an *internal node* which has exactly one incoming edge and two or more outgoing edges or to *leaf/terminal node* which instead has exactly one incoming edge and no outgoing edges. The path of a decision tree terminates with a final outcome at the leaf node.

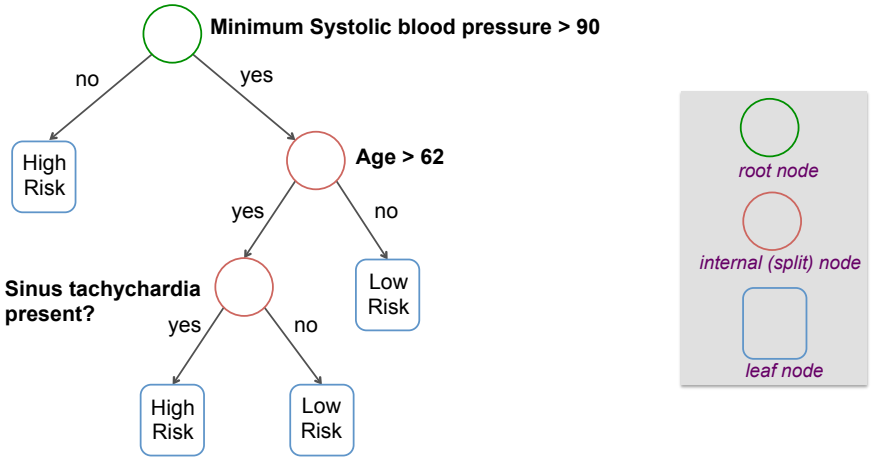


Fig. 4.1. **Decision Tree for Classification.** An oversimplified *decision tree* for survival analysis showing classification of input patient observations into high risk or low risk based on simple decisions in a hierarchical manner. Starting from the root node (green circle), a test condition is specified for each attribute (eg: systolic blood pressure, age, sinus tachycardia) leading to either an internal split node (red circle) for further node splitting or a leaf node (blue (round-edged) square) which stores the final answer i.e. final classification into low risk or high risk patient.

For decision tree induction, the learning algorithm must satisfy two criteria. Firstly, to select an attribute test condition for obtaining smaller subsets of target variable at each recursive step. To achieve this, the learning algorithm has to specify a test condition for different attributes along with an objective function for evaluating the goodness of each test condition. Secondly, a stopping criterion needs to be established to terminate the tree growing process. Three common stopping criteria have been identified for this purpose. First criterion involves the depth of a tree and thus, after a certain depth is achieved, the iterative splitting stops. The second criterion is the minimum training instance population per leaf node. If the population of training instances reaches below a certain threshold, the splitting stops. Finally, the decision function (also known as the objective function) decides whether there is additional information gained after splitting the training instances and the splitting stops once its variation becomes below a certain threshold. The most popular approach, also known as the top-down induction [61, 102], expands a node until the target variable has identical attribute values at a particular node, or until the splitting process leads to the same class and thus further splitting would not create further subsets for prediction. In this thesis, we will mainly focus on binary decision trees for classification.

4.2.1 Decision Tree Learning

The learning process in a binary decision model is defined as an iterative process of recursively selecting best attributes (or features) to split the incoming observations (or data) $\{\mathbf{X}\}_{n=1}^{N} \in \mathbb{R}^{d}$ to predefined class labels $\{\mathbf{Y}\}_{n=1}^{N} \in \mathbb{R}^{d'}$ using a *splitting-function (f)*. Therefore, on splitting at a given node k, two disjoint subsets S_k^{left} and S_k^{right} are generated that follow $S_k = S_k^{left} \cup S_k^{right}$,$S_k^{left} \cap S_k^{right} = \emptyset$, $S_k^{left} = S_{2k+1}$ and $S_k^{right} = S_{2k+2}$. These subsets S_k^{left} and S_k^{right} are sent to the left and the right children of the k^{th} node in the tree, respectively. Thus, the splitting function f_k can be defined as:

$$\begin{cases} f_k(\mathbf{X}) \in \{0,1\}\ , & \\ f_k(\mathbf{X}) = 0, & \mathbf{X} \text{ sent to the Left} \\ f_k(\mathbf{X}) = 1, & \mathbf{X} \text{ sent to the Right} \end{cases} \tag{4.1}$$

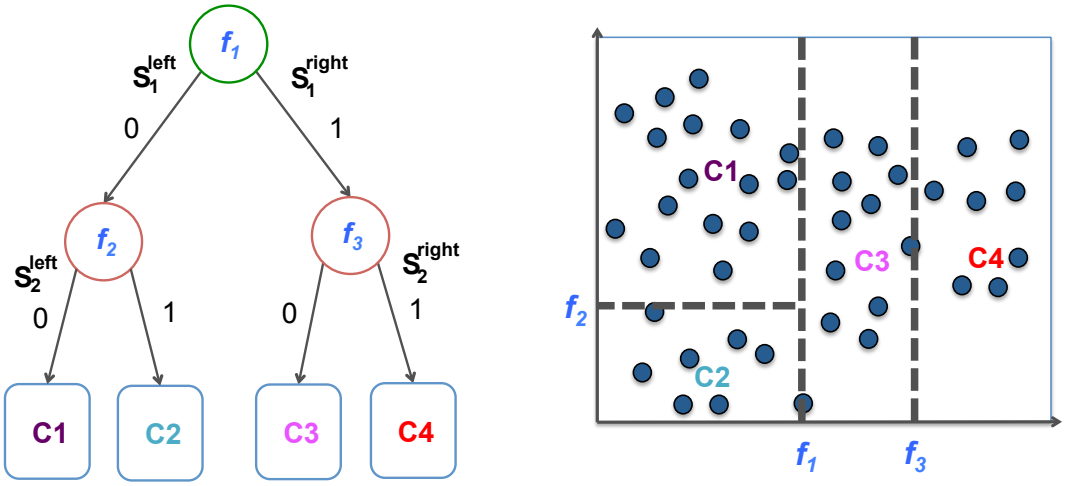


Fig. 4.2. **Decision Tree Splitting.** A simple *binary decision tree* with series of splitting functions (f_1, f_2, f_3) at different nodes partitioning the incoming observations (shown as blue dots) into output classes (shown as classes *C1, C2 C3 and, C4*) at leaf nodes.

During learning phase, the data reaching a leaf node is used to model a maximum a posteriori problem "locally". Thus, for a given leaf node l, let $p(i|l)$ denote the fraction of training data from class i associated with the node l. Then, the posterior model at this leaf node l is given by the equation 4.2.

$$\hat{i} = \mathbf{argmax_i}\ p(i|l), \tag{4.2}$$

where, the *argmax* operator returns the argument i that maximizes the expression $p(i|l)$. Besides providing the information needed to determine the class label of a leaf node, the fraction $p(i|l)$ may also be used to estimate the probability that an input observation assigned to the leaf node l belongs to class i. During the test phase, these posterior distributions allow predictions on new unseen observations reaching a given leaf.

4.2.2 Limitations of Decision Trees

Building a decision tree is computationally inexpensive and is non-parametric for building classification models. Therefore, it does not require any prior assumptions regarding the type of probability distributions satisfied by the class. Once a decision tree has been built, classifying a test instance is extremely fast. However, finding an optimal decision tree is an NP-complete problem that can lead to complex models which do not generalize well due to overfitting problem. Also, the top-down, recursive partitioning approach in the decision tree algorithm might lead to very small number of instances reaching down till the leaf node to allow a statistically significant decision regarding class representation of the nodes, also known as a data fragmentation problem. This could be potentially dealt by restricting the splitting until the number of instances falls below a certain threshold. Decision trees are susceptible to *overfitting* if they are too large. This problem can be reduced by using a tree pruning step by trimming the branches of initial tree such that it improves the generalization capability of the decision tree. In order to achieve better generalization, *Random Forests* have shown to outperform regular decision trees by replacing a single decision tree by an ensemble of decorrelated trees.

4.3 Random Forests

A random forest (or decision forest) is comprised of a group of independent decision trees with decorrelated predictions. Injecting randomness between individual trees allows greater generalization and improved robustness to noisy data and different approaches have been proposed to incorporate this into a decision forest model. Random forests can be mainly instantiated for classification, regression and clustering tasks. A key characteristic that distinguishes classification from regression is that regression forests allow predictive modeling with a final output being continuous instead of being categorical. While classification and regression tasks are associated with supervised learning and model relationship between input and output feature space, a clustering task represents an unsupervised problem where groups (or clusters) of points having similar characteristics in the input data space are required to be detected. Random forests can also be used for density estimation to model probability distribution as described in [32] In this thesis, we will mainly focus on random forests for classification.

4.3.1 Randomization Process

The randomization process is done only during training phase while the test phase is completely deterministic and it can be realized in two ways. In the first approach, also known as "bagging" (a combination of "bootstrap" and "aggregation") introduced by Breiman et al.[18], each independent tree is trained with a random subset of the whole training data thus introducing randomness during the learning process yielding greater training efficiency. In order to achieve this, from a given training set, subsets (or bootstrap) are generated, each of which consist of elements randomly sampled using a uniform distribution with or without replacement as shown in figure 4.3. The final predictions from these individual trees are aggregated by averaging the posterior probabilities generated by each independent tree. The second

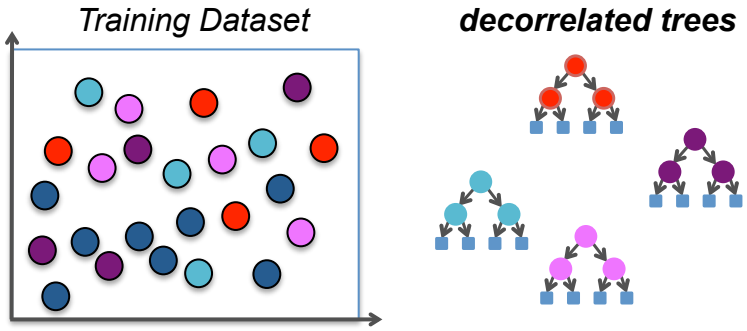


Fig. 4.3. **Bagging Process Subset Generation.** Each independent tree is trained with a random subset of the whole training dataset thereby introducing randomness.

approach is the randomized node optimization[57] which is applied while generating the splitting function. Using a greedy strategy, a set of splitting functions is generated randomly and based on a predefined objective function, the best splitting function is selected from this set. The effect of injecting randomness during training process leads to increase in the degree of decorrelation between different trees, thus increasing generalization. Also, it allows implicit feature selection and robustness against noisy data by gaining independence within the training set. The above two approaches can also be used together and are thus not mutually exclusive, although, bagging is known to achieve greater generalization.

4.3.2 Forest Training and Prediction

All trees (F_t such that $t \in \{1, 2, ..., T\}$) in a random forest are trained independently and possibly parallel to each other. The information required for making final prediction is learned during the training phase at all leaf nodes. Thus, if we consider each leaf node in the tree corresponding to a part of the input feature space, then an ensemble of these leaf nodes in a tree build respective partitions P_t over the given feature space. These leaf nodes (L_z) such that $z \in \{1, 2, ..., T\}$ model the posterior distribution from the given subset of a training set and depending on the chosen objective/decision function, each individual tree behaves as a surjective function leading the input observation X to a leaf node. A posterior model at a leaf node L is used for performing a prediction which is given by:

$$\hat{Y} = \mathbf{argmax_Y}\ P(Y|X \in L, P). \tag{4.3}$$

Eventually, combining final predictions from different trees in a single random forest is usually done by simply averaging tree posteriors at all the leaf nodes and the overall forest prediction is then computed as:

$$P(Y|X) = \frac{1}{T}\sum_{t=1}^{T} P(Y|X \in L_t^{(z_t)}, P_t). \tag{4.4}$$

It should be noted though that averaging tree posteriors is one of the several aggregation approaches that provides a good compromise between giving higher weight to the most confident tree and reducing contribution of noisy data [32]. Other approaches include weighted averaging of all trees according to their respective confidence and performing the averaging only over a fraction of the most confident predictions.

The behaviour of Random forests is also influenced by few important parameters that directly affect the forest's computational efficiency, predictive accuracy and, generalization capability. One of these parameters is the forest size which has been suggested to monotonically increase the final test accuracy [35, 120, 153] It has been shown that the prediction error monotonically decreases (thus increasing test accuracy) with an increase in forest size as accumulating the number of trees in a forest allows to average out noisy predictions that corresponds in a monotonic decrease of the prediction error. The second important parameter is the tree depth which is a crucial to optimize as it directly affects the generalization capability of the forest. On one hand, a short tree might suffer from high heterogenity in the leaf nodes which would decrease its prediction confidence and on the other hand, an extremely deep tree could contain insufficent training data in leaf nodes and thereby, start fitting noisy features leading to poor generalization capability. This is a contributing reason for decreased prediction error with tree depth reaching an optimal point and any further increase in tree depth leads to increased prediction error. Although, very deep trees are prone to overfitting, this can be mitigated by using large training dataset. Another important parameter in constructing a decision forest is the amount of randomness and its effect on the tree correlation. As shown by Criminsi et al. [32], increased randomness of each tree reduces their correlation. However, high randomness leads to much lower overall confidence and such complex weak learners make it difficult to find discriminative sets of parameter values. Apart from the above parameters, the choice of attributes (or features) employed to train the forest also influences its prediction accuracy. Lastly, the training objective function plays an important role in the forest behaviour. Different objective functions that can be employed for training a random forest have been discussed in the next section.

4.4 Classification Forests

Classification forests provide a simple, yet effective strategy of combining randomly trained classification trees. They have been most commonly employed for classification purposes where, an input observation is automatically linked with a predefined output class. The desired output class and the training labels are discrete, categorical,and unordered. Different classifiers have been suggested in literature to build models from input dataset such as the rule-based classifiers, nearest neighbor classifiers, naïve bayes classifiers, support vector machines (SVM), decision tree classifiers, and neural networks. Each classifier uses a learning algorithm to select the model that best fits the relationship between attribute set and class label of the input data and, eventually it should be able to correctly predict class labels of unseen instances. SVM have been one of the most popular choices of classifiers, particularly for binary classification tasks providing maximum-margin separation, thereby allowing good generalization even on smaller training data [142, 143]. Unfortunately, SVM can be memory-intensive, inefficient to train, and difficult to interpret. Moreover, they do not extend naturally to multi-class problems [75, 134]. Several useful properties of random forests allow their effective use for classification purposes including their scalability to large training sets, fast training and predictions, good generalisation to previously unseen data, yielding a probabilistic output, and their ability to handle multi-class problems. Also, they can provide good insight into the importance of a given feature and are generally easier to interpret by humans. A number of applications have used classification forests successfully [34, 65, 95, 106, 120]. In particular, for kidney segmentation, Kontschieder et al. [74] performed semantic image segmentation using geodesic distances as an additional criterion for the node splitting in order to ensure spatial compactness of the pixel clusters of each child node.

4.4.1 Problem Statement

Given a training set $\{X^{(n)}, Y^{(n)}\}_{n=1}^{N}$, the goal of classification is to model a posterior probability distribution $P(Y|X)$ such that for any unseen observation lying in the feature space of the input instances $X^{(n)}$, can be assigned to its label lying in the output feature space of $Y^{(n)}$ using the maximum a posteriori given by:

$$\hat{Y} = \mathbf{argmax_Y}\ P(Y|X). \tag{4.5}$$

Each tree F_t (such that $t \in \{1, 2, ..., T\}$) in a classification forest, builds a partition P_t over the input feature space which is instantiated by two main components, a predefined *decision function*: to select the best split during node optimization and, the *leaf posterior*: to recursively split the training data and reduce the class uncertainty associated with these class posteriors. Different decision functions (also referred to as a objective function or impurity function in literature) can be used during node optimization process as described ahead.

4.4.2 Decision Function

In classification tasks, the goal of node optimization is to find the best split based on a predefined decision function aiming to reduce the class uncertainty (or impurity). When

training a tree, the degree of impurity of the parent node (before splitting) is compared with the degree of impurity of the child nodes (after splitting) and, to what extent each feature decreases the impurity (sometimes weighted impurity) in a tree is assessed. For binary classification, the smaller the degree of impurity, the more skewed is the class distribution. In this way, a node with class distribution (0,1) would have zero impurity, whereas a node with uniform class distribution (0.5, 0.5) would have the highest impurity.

The most popular impurity measures for classification are *Gini impurity* and *Entropy (more specifically, Shannon entropy)*. The *Gini impurity* provides a measure of misclassification by computing the probability of an element from a set being misclassified when it is randomly picked and assigned a label from a given distribution in a subset. Thus, from a given subset of class distribution at a node, gini impurity provides the expected error at this node, if a datapoint is randomly selected and assigned to a label from the distribution at that node. Gini impurity reaches its minimum (zero) when all instances in the node fall into a single label category, making the set completely pure.

Let s_k be the subset of training observations arriving at a given node *k*. Then, the fraction of observations belonging to a class *i* can be denoted by $p(i|s_k)$. The gini impurity is given by:

$$G(s_k) = 1 - \sum_{i=1}^{c} [p(i|s_k)]^2, \tag{4.6}$$

where, c is the number of classes. On the other hand, the Shannon entropy is given by:

$$E(s_k) = -\sum_{i=1}^{c} p(i|s_k) \log_2 p(i|s_k). \tag{4.7}$$

After splitting, s_k is further divided into two subsets,s_k^{left} sent to the left child node and s_k^{right} sent to the right child node, respectively.

When using the Shannon entropy as the impurity measure, the difference between class uncertainty before and after the node splitting is known as the *information gain* (Δ_{info}), which can be used to determine the goodness of split as follows:

$$\Delta_{info} = E(s_k) - w_{left} E(s_k^{left}) - w_{right} E(s_k^{right}), \tag{4.8}$$

where, $E(s_k)$ is the entropy at the parent node *k*, while $w_{left} = |s_k|/|s_k^{left}|$ and $w_{right} = |s_k|/|s_k^{right}|$.

In figure 4.4, the values of both impurity measures, *Gini* and *Entropy* for binary classification have been shown. The fraction of records belonging to one of the two classes have been denoted by *p*. As shown in the figure, both the impurity measures attain maximum value when the class distribution is uniform (i.e., when p = 0.5) while the minimum values are attained when all the observations at the node belong to the same class (i.e., when p equals 0 or 1). Both, Gini impurity as well as Shannon entropy behave similarly, and one can expect similar results using any one of the two measures. In this thesis, we will focus on node optimization using information gain (Δ_{info}).

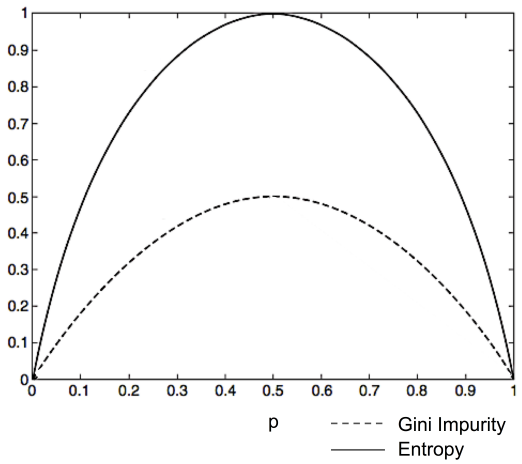


Fig. 4.4. **Impurity Measures.** Entropy and Gini impurity are shown for a binary classification problem. The *x-axis* represents the probability (p) of one class and the *y-axis* shows the value of both impurity measures. Both measures of class uncertainty reach their maximum at 0.5.

Thus, using a greedy strategy for training each tree, a set of splitting function (f) candidates are generated at every node and the best candidate is selected based on the one maximizing information gain (Δ_{info}):

$$f_k = \mathbf{argmax}_{\mathbf{f_k}} \mathbf{\Delta_{info}}(s_k, s_k^{left}, s_k^{right}). \tag{4.9}$$

By optimizing the decision function, leaf nodes consisting of observations lying in similar feature space and belonging to the same class are generated.

4.4.3 Class Posteriors

During training of each decision tree, the training data is split recursively to reduce class uncertainty associated with *class posteriors* by creating leaf nodes that are class consistent. Thus, for a random tree F_t, class posteriors can be approximated at each leaf node $L_t^{(z_t)}$ with class consistent partitions $P_t = \{L_t^{(z_t)}\}_{z_t=1}^{Z_t}$ over the given feature space as below:

$$P(i|X \in L_t^{(z_t)}, P_t) = \frac{|\{X^{(n)} \in L_t^{(z_t)}, Y^{(n)} = i\}|}{|\{X^{(n)} \in L_t^{(z_t)}\}|}. \tag{4.10}$$

4.4.4 Forest Prediction

After the training phase, predictions on unseen incoming observations can be performed by feeding them to the forest and combining all the tree posteriors in the forest. Thus, the forest prediction Y for an input observation X can be performed as the average of all the tree posteriors given by:

$$P(Y|X) = \frac{1}{T} \sum_{t=1}^{T} P(Y|X \in L_t^{(z_t)}, P_t), \tag{4.11}$$

and then using the maximum a posteriori:

$$\hat{Y} = \mathbf{argmax_Y}\ P(Y|X). \tag{4.12}$$

In the next section, we describe application of random forests for segmentation of polycystic kidneys in ADPKD.

4.5 Semi-Automatic Segmentation of Polycystic Kidneys

In this section, we describe a semi-automatic segmentation method based on random forests for 3D segmentation of kidneys from patients with ADPKD and severe renal insufficiency, using computed tomography (CT) data. As described before, ADPKD severely alters the shape of the kidneys due to non-uniform cyst formation both in the kidneys and surrounding liver. Therefore, fully automatic segmentation of such kidneys is very challenging. We present a semi-automatic segmentation approach based on a random forest classifier with minimal user interaction. The main novelty of the approach is the introduction of geodesic distance volumes as additional source of information to the random forest classifier. These volumes contain the intensity weighted distance to a manual outline of the respective kidney in only one slice (for each kidney) of the CT volume. We evaluate the performance of the proposed approach qualitatively and quantitatively on 55 CT acquisitions using ground truth annotations from clinical experts.

4.5.1 Patient Dataset

For our experiments, a total of 55 CT acquisitions from 41 ADPKD subjects were used. For image acquisition, a 64-slice CT scanner (LightSpeed VCT; GE Healthcare; Milwaukee, WI) was used, with single breath-hold scans and same scanning parameters for all patients (voltage 120 kV, current 150 - 500 mAs, collimation 2.5 mm, matrix 512x512, slice pitch 0.984 and increment 2.5 mm). All of these data sets were manually segmented by clinical experts in order to obtain ground truth annotations.

4.5.2 Method

In this work we propose a semi-automatic approach for segmenting ADPKD kidneys from CT data. Initially, the user performs an outline of each of the polycystic kidneys in it's corresponding mid-slice, i.e., the middle slice out of all the sections containing kidney. Then, as shown in figure 4.5 (right), an intensity weighted geodesic distance to the respective mid-slice segmentation is computed at all non-segmented voxels, as described by Soille [124]. This generates two 3D distance volumes, one for each kidney as additional modalities (information channels). Our goal is to formulate the segmentation task as a voxel-wise classification problem, where we assign to each voxel p a label $l(p) \in \{l_b, l_r, l_l\}$ where, l_b models the background class, l_r denotes the label for the right kidney and, l_l denotes the label for the left kidney. Based on a set of labeled output classes, we aim at training a decision rule by means of a random forest classifier.

The random forest classifier used for our experiments consists of a collection of decorrelated binary decision trees. Each decision tree in the random forest, is a hierarchically ordered set of nodes, where each node has exactly 0 or 2 children and in the former case, the node is called a leaf. Using a set of labeled output classes, these decorrelated decision trees are trained in order to infer the relationship between visual features and labels. Thus, the random forest classifier provides a piecewise approximation of each class posterior over the feature space.

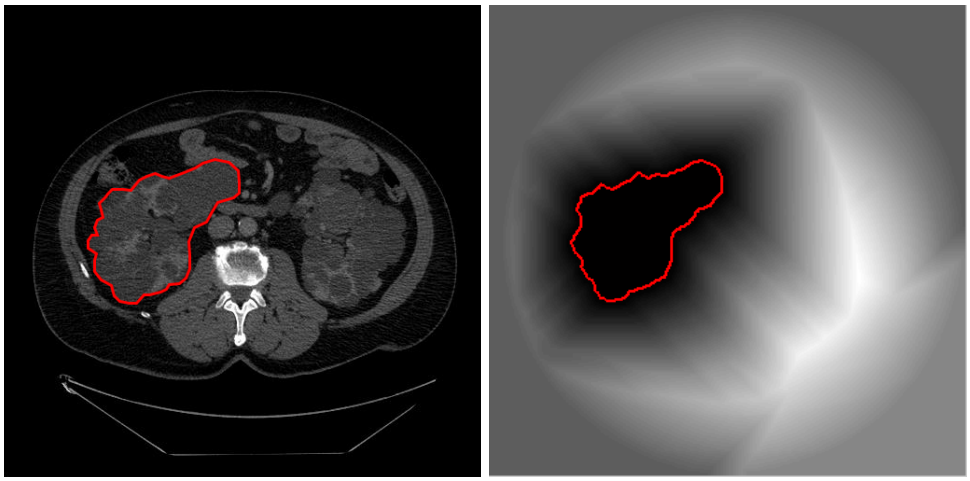


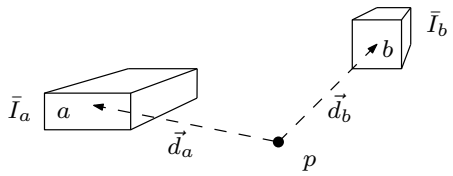
Fig. 4.5. **Geodesic Distance Map. Left:** Manually outlined original mid-slice image. **Right:** Corresponding Geodesic Distance Map.

Feature Selection

For the purpose of classification, we use so-called *box features* for classification. As depicted in figure 4.6, a box feature at a location p is defined by two offset vectors $\overrightarrow{d_a} \in R^3$ and $\overrightarrow{d_b} \in R^3$ which specify the centers of mass of two boxes a and b. These boxes are of size $x_a \times y_a \times z_a$ and $x_b \times y_b \times z_b$, respectively. In each of these boxes we calculate the mean value w.r.t. one intensity (information) channel. Thus, it is possible that the mean values $\bar{I}_a$ and $\bar{I}_b$ can be computed from different channels in order to capture inter-modality correlations. Once the mean values are computed, we select one of the six functions h_j, where $j = 1, ..., 6$, as shown in figure 4.6, for computing a scalar feature value. Therefore, one box feature can be parametrized by a vector:

$$(\overrightarrow{d_a}, \overrightarrow{d_b}, x_a, y_a, z_a, x_b, y_b, z_b, k_a, k_b, j), \tag{4.13}$$

where, k_a and k_b specify the intensity channels used for computing the respective mean value and $j = 1, ..., 6$ specifies the function used for computing the scalar feature value.



$$h_1(\bar{I}_a, \bar{I}_a) = \bar{I}_a \qquad h_4(\bar{I}_a, \bar{I}_b) = \bar{I}_a + \bar{I}_b$$
$$h_2(\bar{I}_b, \bar{I}_b) = \bar{I}_b \qquad h_5(\bar{I}_a, \bar{I}_b) = \bar{I}_a > \bar{I}_b$$
$$h_3(\bar{I}_a, \bar{I}_b) = \bar{I}_a - \bar{I}_b \qquad h_6(\bar{I}_a, \bar{I}_b) = |\bar{I}_a - \bar{I}_b|$$

Fig. 4.6. **Box Feature.** A box feature is defined by the two offset vectors $\vec{d}_a, \vec{d}_b \in \mathbb{R}^3$, the box sizes, and the choice of the function h_i used for computing a single scalar value out of the mean values $\bar{I}_a$ and $\bar{I}_b$. Note that these mean values could be computed from different information channels, i.e., the CT volume and the two geodesic distance volumes.

Forest Training

For each decision tree, we randomly select a set of training samples S, consisting of voxels with known labels and start building the tree at the root node. At each node, we select a splitting function defined by the selected box feature and a threshold value as follows:

First, we randomly select 100 features and compute the minimum and maximum values of the respective feature on all samples of the node. Then we divide the range of each feature using 10 thresholds (equally spaced between the respective maximum and minimum value) and evaluate the information gain of all considered feature-threshold combinations. The splitting function at the current node is then chosen as the combination of feature and corresponding threshold which yields the overall highest information gain. Although this strategy is a greedy one, it is still one of the most popular choices due to its computational efficiency [33]. This node splitting process is repeated recursively until either the maximum depth is reached, or if the number of samples sent to child nodes is too low. Eventually, each leaf node models the class posterior estimate using a histogram from the samples that reached this leaf node.

Forest Testing

For prediction, the goal is to classify the unseen voxels p. We feed each test sample through the tree, starting at the root node and according to the splits recorded at each node during the training, we obtain the class histogram stored at the leaf reached by the test sample in this particular decision tree. The overall prediction is then computed as the average of the output posteriors of each tree, and the prediction for each voxel p is then given by the class with highest average posterior.

4.5.3 Evaluation

We performed a 5-fold cross-validation, i.e., we selected 44 samples for the training and tested the trained classifiers on the remaining 11 samples. This process was repeated five times such that every data set has been used once for validation. Moreover, we performed two sets of experiments: In the first set we trained the random forest classifiers without the geodesic distance volumes as additional information channels. In the second round of experiments we trained the classifier with the same settings, but with the geodesic distance volumes as additional information channels.

4.5.4 Results and Conclusion

We computed the mean dice score coefficient (DSC) for the predicted volumes for all 55 cases. While the green bars in figure 4.8 depict the results computed with the geodesic distance volumes, the red bars show the results for the baseline approach, i.e., the random forest classifier trained solely using the CT volumes. For the baseline approach, the DSC for right and left kidneys was 0.67 ± 0.13 and 0.68 ± 0.14, respectively. Reported DSC for the proposed geodesic distance volumes approach for right and left kidneys was 0.70 ± 0.11 and 0.71 ± 0.13. Example predictions of ADPKD kidneys from the random forest classifier using proposed geodesic distance volumes have been shown in figure 4.7.

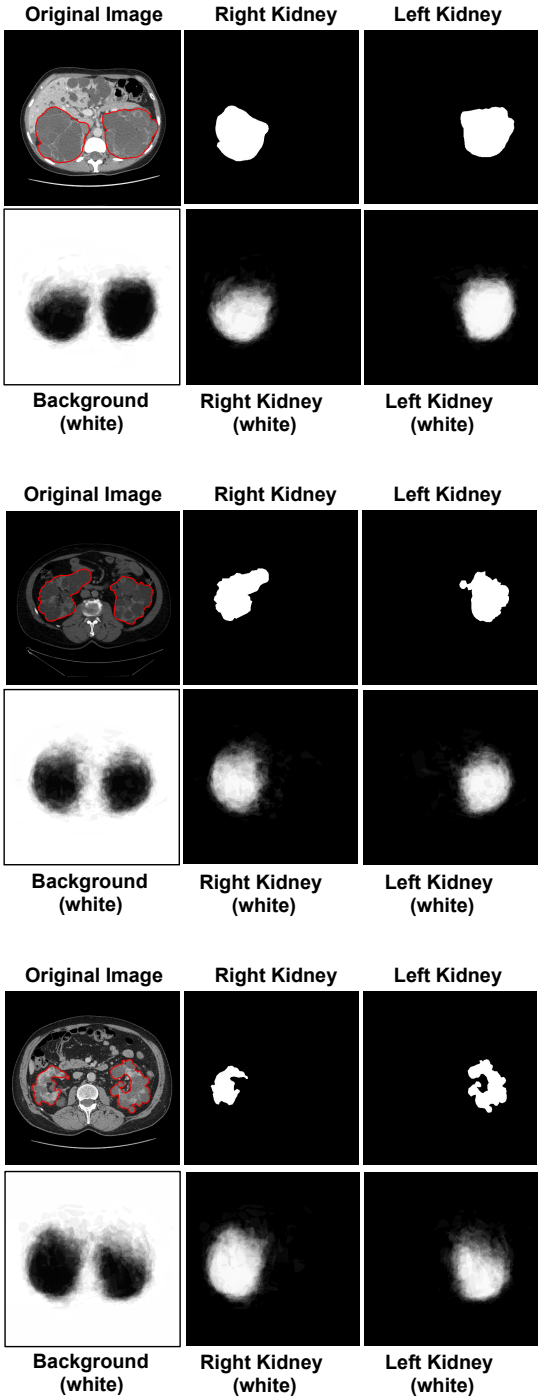


Fig. 4.7. **Random Forest Predictions. Upper Panels:** Original segmentations (red contour) of ADPKD kidneys and generated manual segmentation masks of right and left kidneys for 3 different cases. **Lower Panels:** Random forest predictions (proposed geodesic distance volumes approach) of background (bottom left), right kidney (bottom-middle) and left kidney (bottom-right) classes shown in white.

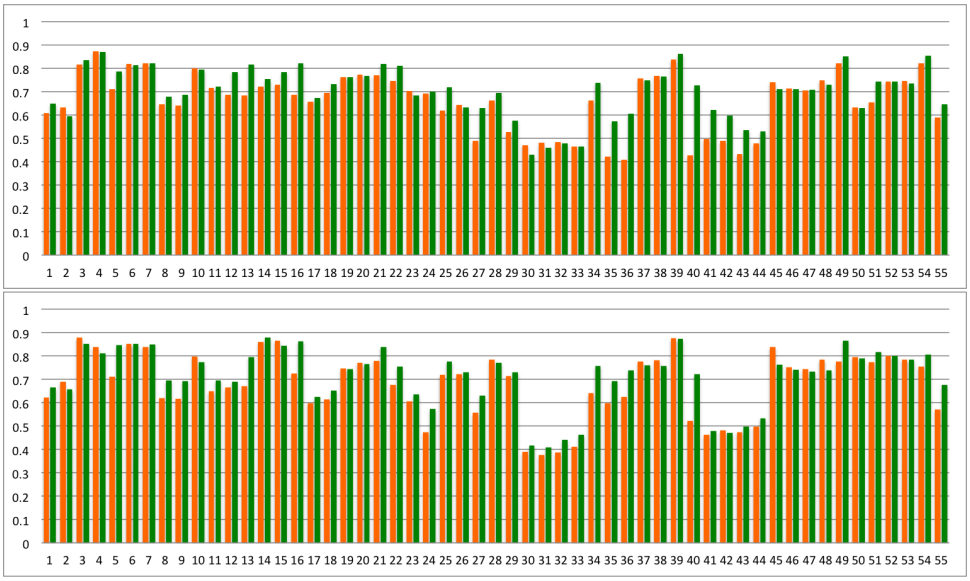


Fig. 4.8. **Dice Scores for Right Kidneys (upper panel) and Left Kidneys (lower panel).** The results obtained for all 55 acquisitions with the baseline method are shown in red, while the results of the proposed geodesic distance volume approach are shown in green.

In general, we can make the following observations: Firstly, the segmentation results tend to be better for the left kidneys which is due to the fact that the boundary between the right kidney and the liver is often hard to discriminate. Secondly, there are a considerable number of cases where the geodesic distance volumes improve the segmentation results - especially in case of the left kidneys. We improved in 35 out of 55 cases by 13.25% on average for the right kidney and in 36 out of 55 cases by 10.35% for the left kidney, respectively, while the baseline is only better by 2% for right kidney and by 2.42% for the left kidney on average for the remaining 20 and 19 cases, respectively. Therefore, the average gain of the proposed method (in the cases where it outperforms the baseline) tends to be higher than the average gain of the baseline method (in the other cases).

TKV Agreement Analysis

We performed volumetric measurement on kidney segmentations from the proposed random forest approach and compared the semi-automated TKV with the true TKV (obtained from ground truth annotations). The Mean absolute percentage TKV error (MAPE) was $77.9\% \pm 64.3\%$ and the Coefficient of Variation (COV) was 37.6%. The Bland Altman plots were used to determine agreement between the TKV computed from proposed random forest approach (with geodesic distance volumes) true TKV obtained from manual segmentations of the ADPKD kidneys as shown in figure 4.9. The lower and upper limits of agreement (LOA) for percentage difference on Bland-Altman plots were -6.7% and 106.4%, respectively.

We presented a method for segmentation of ADPKD kidneys on contrast enhanced CT. Based on the results, we may draw the following two conclusions: Firstly, the proposed usage of the random forest approach helps to improve the overall segmentation process compared to baseline approach. Secondly, all results clearly show that segmentation of ADPKD kidneys is not at all a solved task. The main reasons are: (i) the progressive cyst expansion in ADPKD leading to a significant and unpredictable deformation and enlargement of the kidneys, especially in

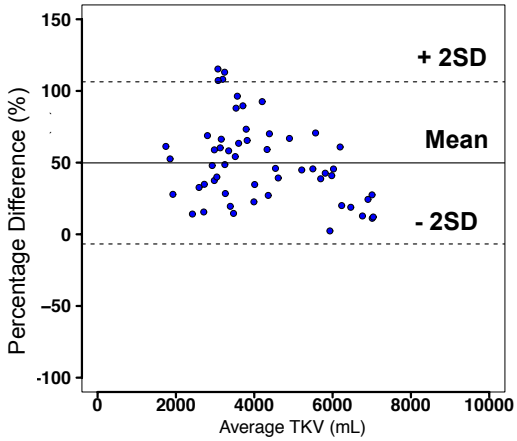


Fig. 4.9. **TKV Agreement Analysis using Bland-Altman Plots.** TKV measurements from semi-automated segmentation method using geodesic distance volumes compared with true TKV measurements from manual segmentations of ADPKD kidneys.

patients at late stage of the disease, and (ii) the aforementioned tissue inhomogeneities of surrounding organs, for instance due to neighboring liver cysts, make fully automated kidney segmentation a challenging task. As evaluating forests is computationally very efficient, it currently seems to be a good strategy to evaluate random forests. We would like to emphasize that this is to the best of our knowledge, one of the first approaches for minimally interactive segmentation of ADPKD kidneys from CT data. Both CT and MRI have been investigated for monitoring structural changes in ADPKD and for association between TKV and renal function or renal function decline. The reason to acquire contrast enhanced CT images in the current study was to perform further renal compartment measurements. But, these additional measurements are out of the scope of this work.

5 Deep Learning for Segmentation

5.1 Artificial Neural Networks

Artificial neural networks (ANN) have been inspired by the biological neural system consisting of several interconnected neurons. The input data, which can be a multidimensional vector, is fed to an input layer that is further connected to a series of hidden layers. These hidden layers generate activity patterns (or activations) that encode information about important features contained in the input data and make decisions based on the information received from previous layers. The network undergoes a learning process that is enabled by adjusting the strength of weighted connections (or weights) between neurons in the different layers. As described previously, different strategies have been used to train a neural network.

Supervised Learning: In supervised learning, the input data is fed to the network while the desired output (i.e. ground truth labels) is available to improve the learning process by updating the weights in the network. During training, the final classification error is minimized based on these ground truth labels. Examples of supervised learning include regression and classification tasks.

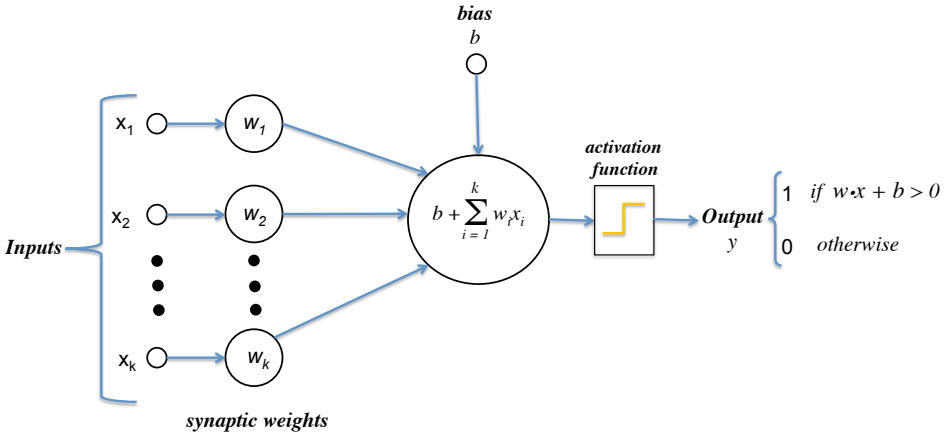
Unsupervised Learning: The training process in unsupervised learning does not contain any ground truth labels. The network improves by reducing or increasing the cost function associated with the learning process. An example of unsupervised learning is clustering i.e. dividing the entire dataset into different groups according to some unknown pattern. Another example is self-organizing maps typically used for dimensionality reduction.

Semi-Supervised Learning: Semi-supervised learning make use of large amount of unlabeled data together with small amount of labeled data for training a network. Recently, such combination of supervised and unsupervised learning was proposed for deep neural networks, known as the *Ladder Networks* trained to simultaneously minimize the sum of supervised and unsupervised loss functions [105].

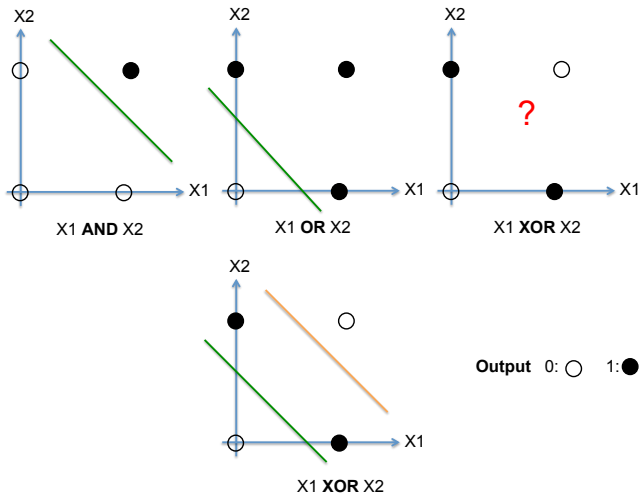
Reinforcement Learning: This type of strategy is based on observation and is similar to supervised learning since it requires some feedback. However, based on how well the neural network performs through trial and error, the feedback is supplied as a reward and the network adjusts its weights to allow better decision making in the subsequent iterations. Examples of reinforcement learning include robot navigations by adapting through negative feedback of encountering obstacles and several logic games such as chess and backgammon.

5.1.1 The Perceptron

In 1957, psychologist Frank Rosenblatt conceptualized an electronic brain model known as the *photoperceptron* [107]. The simplest type of artificial neural network known as a *single layer perceptron* (SLP) follows the mathematical modeling of a biological neuron and was first used for image recognition purpose [108]. It is a linear classifier that computes weighted sum of its inputs and learns a linear decision boundary by employing an activation function that outputs a non-zero value only when this weighted sum exceeds a certain threshold, as shown in figure 5.1a. The SLP is capable of performing only boolean logic operations (such as AND, OR operations) to solve linearly separable problems. Instead, a multi-layer perceptron (MLP) (*feed-forward neural network*) can be used for classification of linearly inseparable problems (such as XOR operation), as shown in figure 5.1b.



(a)Single Layer Perceptron



(b)Decision Boundary Learning

Fig. 5.1. **(a) Single Layer Perceptron:** Single layer perceptron learning a binary classifier by employing an activation function (unit step function) that takes a linear combination of the input values x_i and weights w_i, where i = (1,...,k) and labels a positive output *(y)* when this weighted sum exceeds a threshold. The bias *(b)* shifts the decision boundary away from the origin. **(b) Decision Boundary Learning:** Single layer perceptron is capable of learning only a linear decision boundary (such as AND, OR) while, a multi-Layer perceptron can be used for solving linearly inseparable problems (such as XOR) to generate more complex decision boundaries.

5.1.2 Learning Process: Introducing Non-Linearity

A *feed-forward* artificial neural network is composed of multiple layers such that each layer is connected to every next layer without any connection among neurons in the same layer and the information flows only in forward direction. Along with the input and output layers, it consists of one or more hidden layers and each layer contains several neurons that are interconnected by synaptic weight links. As shown in figure 5.2, the information flows in a feed-forward manner from the input via hidden layers to the output layer. The selection of optimal number of hidden layers is important as too small of a network might lack representative power of modeling useful features leading to high bias problem. On the other hand, if the number of hidden layers is too many, it may lead to over fitting of the input training data thereby modeling the noise in the training dataset, leading to a high variance problem.

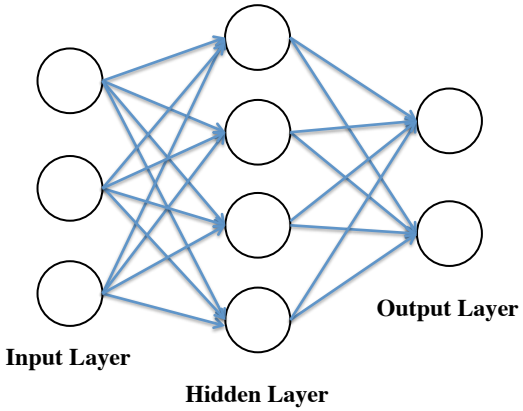


Fig. 5.2. **Feed-Forward Neural Network Architecture.** A multilayer neural network with three layers: *input layer, hidden layer and output layer*. Each layer is connected to the last one. Multiple hidden layers can be introduced to make the architecture deep.

Each neuron in the neural network is associated with an input weight *(w)* and a bias term *(b)*. In response to the received input *(x)*, the neurons in a neural network are modeled using a *non-linear* activation function *(f)*.

$$y = f(\Sigma wx + b), \tag{5.1}$$

where, x is the received input,
w is the learned weight and,
b is the associated bias parameter,
y is the output vector.

The weight w is updated during the learning process while the bias term b accounts for the possible mean shift, moving the activation function to the left or right as required for successful learning of the model. In this thesis we will mainly focus on supervised learning strategy, wherein a *loss function* measures the cost of predicting y (the true label) while parameterizing the activation function f using weight vector w. Different types of activation functions can be employed for performing this non-linear function modeling as described below.

Sigmoid: The sigmoid function has been traditionally used for introducing non-linearity into the network to generate strong classifiers and has the mathematical form shown below. It takes real valued input and the output is in the range of [0,1].

$$f(x) = \frac{1}{(1 + e^{-x})} \tag{5.2}$$

TanH: The TanH function takes real valued inputs and produces output in the range of [-1,1]. It suffers from the saturation problem as well however, the output from the TanH activation are zero-centered.

$$f(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{5.3}$$

ReLU: The rectified linear unit is one of the most commonly used activation function in deep neural networks. A rectified linear unit or ReLU eliminates negative values $(x < 0)$ and thresholds them at zero while computing the function:

$$f(x) = max(0, x) \tag{5.4}$$

Its popularity comes from the advantages of using it over sigmoid or tanH activations. Due to its non-linearity and non-saturating property, the ReLU is considered to speed up the stochastic gradient descent convergence and require low computation time in comparison to the traditional sigmoid or tanH functions as shown by Krizhevsky et al. [76].

Leaky ReLU: One disadvantage of using ReLU with high learning rate is that it might encounter large gradients that lead to weight updates that never activate the neurons and thus the gradient remains equal to zero starting this point. These ReLU units can remain "dead" and this issue can be avoided by using proper learning rate settings. In order to get rid of the above problem, Leaky ReLUs were devised wherein a small negative slope (such as 0.01) is associated with the function. However, using leaky ReLU has not been very popular due to inconsistency in its results.

$$f(x) := \begin{cases} \alpha x & \text{if, } x < 0 \\ x & \text{if, } x > 0 \end{cases} \tag{5.5}$$

here, α is a small constant.

MaxOut: The maxout function computes:

$$f(x) = max(w_1 x_1 + b_1, w_2 x_2 + b_2), \tag{5.6}$$

and thus, it is a generalization of ReLU and leaky ReLU while benefitting with the advantages of leaky ReLU with stable weight updates [42]. However, using the MaxOut function leads to increase in the number of parameters and thus, it is also not a popular choice against ReLU.

5.1.3 Training a Neural Network

ANNs can be formulated in terms of minimization of a loss function which is influenced by adaptative parameters such as the synaptic weights and biases. For training a neural network, *gradient-based* algorithms have been popularly employed as they are known to converge fast and a common method for gradient computation through application of recursive chain rule is known as *Backpropagation* [112, 113, 150, 151]. The backpropagation algorithm was first proposed by Paul Werbos [150] and became widely popular in the 1980s with the work of Rumelhart et al. [113] . In general, backpropagation has been widely discussed in context of supervised learning where it uses the desired output for each input and attempts to minimize the final loss function. However, it can also be used for unsupervised learning where the desired output is equal to the input and the network attempts to learn a compact representation of the input distribution. Examples of using backpropagation for unsupervised tasks include training of autoencoders [11], which may be typically useful for dimensionality reduction or for recently developed deep belief networks [54, 55]. The backpropagation process mainly includes two phases, a *Forward Pass* and a *Backward Pass*. For supervised learning, the input data is introduced to the network and processed through different layers of the network until it arrives at the output layer where the actual output is compared with desired output. Then, the error between the actual and desired output is computed in terms of minimization of the loss function and calculated for each neuron in the output layer. These errors computed at neurons in the output layer are then propagated backwards to each layer in the network and during this process, backpropagation utilizes these errors to compute gradients of the loss function with respect to the weights in the network. Finally, the computed gradients in conjunction with a suitable optimization method, are used to update the weights of neurons in the network with an ultimate goal to minimize the loss. In this way, backpropagation allows randomly initialized neurons in a neural network to find the right set of parameters to learn relevant features from the input dataset for successful predictions. After training, the goal of such a network is then to use these learned parameters to accurately identify similar patterns (or features) in new input data that was previously unseen during training phase and introduced to the network without any information regarding the expected output. Below, the two phases of *backpropagation* have been described in more detail.

Forward and Backward Propagation

Consider a simple ANN with two inputs, two hidden neurons and, two output neurons as shown in figure 5.3. The goal of backpropagation is to optimize associated weights with each neuron such that the network learns to correctly map arbitrary inputs to the outputs. Starting from the input layer consisting of the inputs: x_1 and x_2, respectively, the data reaches the first hidden layer consisting of neurons h_1 and h_2 and produces training outputs at o_1 and o_2. Each input data interacts with every individual neuron in the hidden layer and the total output of a neuron in the hidden layer is given by a combination of its weights and biases as shown below for the neuron h_1.

$$net_{h_1} = w_1 x_1 + w_2 x_2 + b_1. \tag{5.7}$$

Similarly, the output net_{h_2} from h_2 is computed by replacing the weights with w_3 and w_4 in the above equation. Using a sigmoid activation function, the output for h_1 is given by:

$$out_{h_1} = \frac{1}{(1 + e^{-net_{h_1}})}. \tag{5.8}$$

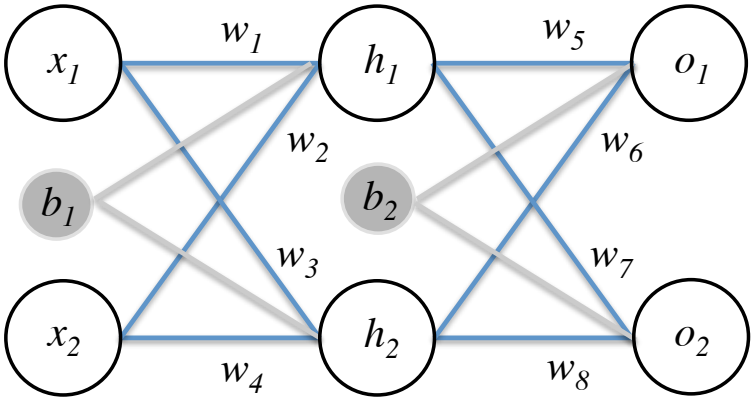


Fig. 5.3. **Artificial Neural Network Training.** A simple ANN representation consisting of two inputs, two hidden neurons and two output neurons.

In the same way, the activation is applied on net_{h_2}. The above process is then repeated for the output layer neurons, by utilizing the outputs out_{h_1} and out_{h_2} from hidden layer neurons as inputs to the final output layer. Thus, the output out_{o_1} at o_1 is given by:

$$net_{o_1} = w_5\ out_{h_1} + w_6\ out_{h_2} + b_2. \tag{5.9}$$

With application of the sigmoid activation function, we get:

$$out_{o_1} = \frac{1}{(1 + e^{-net_{o_1}})}. \tag{5.10}$$

After computing the output out_{o_2} at o_2, we now compute the error using an appropriate loss function for both output neurons o_1 and o_2. One of the choices for loss function such as the *sum-of-squared errors* would be computed as:

$$E = \mathbf{\Sigma}\frac{1}{2}(output_{desired} - output_{actual})^2. \tag{5.11}$$

In the above equations, $output_{desired}$ is given by the ground truth labels while $output_{actual}$ is the network output given by out_{o_1} or out_{o_2}. The total error from the output neurons o_1 and o_2 is given by:

$$E = E_{o_1} + E_{o_2}. \tag{5.12}$$

This leads to the next phase of *backward propagation* where the goal is to update each of the weights in the network such that the actual output from the training is brought closer to desired output by minimizing the loss function for each neuron in the entire network. Thus, using the chain rule, we compute the change in error due to contribution of the weight w_5 as partial derivative of E with respect to w_5:

$$\frac{\partial E}{\partial w_5} = \frac{\partial E}{\partial out_{o_1}} \frac{\partial out_{o_1}}{\partial net_{o_1}} \frac{\partial net_{o_1}}{\partial w_5}. \tag{5.13}$$

Taking partial derivative at equation 5.11 w.r.t out_{o_1}, we get:

$$\frac{\partial E}{\partial out_{o_1}} = -(desired_{o1} - out_{o_1}), \tag{5.14}$$

where, $desired_{o1}$ is the ground truth output value at o_1 and, partial derivative of out_{o_1} w.r.t net_{o_1}, is computed as:

$$\frac{\partial out_{o_1}}{\partial net_{o_1}} = out_{o_1}(1 - out_{o_1}). \tag{5.15}$$

While considering equation 5.9, the partial derivative of net_{o_1} w.r.t w_5 is equal to out_{h_1}.

$$\frac{\partial net_{o_1}}{\partial w_5} = out_{h_1}. \tag{5.16}$$

Finally,

$$\frac{\partial E}{\partial w_5} = -(desired_{o1} - out_{o_1})\, out_{o_1}(1 - out_{o_1}) out_{h_1}. \tag{5.17}$$

Similarly, chain rule is applied to compute changes in error with respect to w_6, w_7 and w_8. The backward pass is then propagated backwards to compute change in the error associated with neurons in the *hidden layer*, such as for w_1 given by:

$$\frac{\partial E}{\partial w_1} = \frac{\partial E}{\partial out_{h_1}} \frac{\partial out_{h_1}}{\partial net_{h_1}} \frac{\partial net_{h_1}}{\partial w_1}. \tag{5.18}$$

Since the output of each hidden layer neuron contributes to the error of both output neurons out_{o_1} and out_{o_2}, therefore in equation 5.18:

$$\frac{\partial E}{\partial out_{h_1}} = \frac{\partial E_{out_{o_1}}}{\partial out_{h_1}} + \frac{\partial E_{out_{o_2}}}{\partial out_{h_1}}. \tag{5.19}$$

Now, in above equation

$$\frac{\partial E_{out_{o_1}}}{\partial out_{h_1}} = \frac{\partial E_{out_{o_1}}}{\partial net_{o_1}} \frac{\partial net_{o_1}}{\partial out_{h_1}}, \tag{5.20}$$

where,

$$\frac{\partial E_{out_{o_1}}}{\partial net_{o_1}} = \frac{\partial E_{out_{o_1}}}{\partial out_{o_1}} \frac{\partial out_{o_1}}{\partial net_{o_1}}. \tag{5.21}$$

Since,

$$E_{out_{o_1}} = \frac{1}{2}(desired_{o_1} - out_{o_1})^2, \tag{5.22}$$

following equation 5.14, the partial derivative of $E_{out_{o_1}}$ w.r.t out_{o_1} is given by:

$$\frac{\partial E_{out_{o_1}}}{\partial out_{o_1}} = -(desired_{o1} - out_{o_1}), \tag{5.23}$$

and from equation 5.15, we know that:

$$\frac{\partial out_{o_1}}{\partial net_{o_1}} = out_{o_1}(1 - out_{o_1}). \tag{5.24}$$

Now, using equation 5.9, we find the partial derivative of net_{o_1} w.r.t out_{h_1}. Therefore, in equation 5.20:

$$\frac{\partial net_{o_1}}{\partial out_{h_1}} = w_5. \tag{5.25}$$

Finally,

$$\frac{\partial E_{out_{o_1}}}{\partial out_{h_1}} = -(desired_{o1} - out_{o_1}) out_{o_1}(1 - out_{o_1}) w_5. \tag{5.26}$$

In equation 5.19, the process is similarly repeated for $\frac{\partial E_{out_{o2}}}{\partial out_{h_1}}$ giving:

$$\frac{\partial E_{out_{o2}}}{\partial out_{h_1}} = -(desired_{o2} - out_{o_2}) out_{o_2}(1 - out_{o_2}) w_7. \tag{5.27}$$

In equation 5.18, $\frac{\partial E}{\partial out_{h_1}}$ (denoted as E_{tot})can be calculated by adding above two equations (5.26 and 5.27) while remaining terms are given by:

$$\frac{\partial out_{h_1}}{\partial net_{h_1}} = out_{h1}(1 - out_{h1}), \tag{5.28}$$

and (using equation 5.7),

$$\frac{\partial net_{h_1}}{\partial w_1} = x_1. \tag{5.29}$$

Finally, equation 5.18 can be solved as:

$$\frac{\partial E}{\partial w_1} = E_{tot}\, out_{h1}(1 - out_{h1})\, x_1. \tag{5.30}$$

Changes in error with respect to w_2, w_3 and w_4 can be computed similarly. This way, by calculating partial derivatives starting from the output layer through the hidden layer, all weights in the network are updated while minimizing the loss function and this process is repeated recursively to achieve the closest optimal solution to the desired output.

5.2 Deep Learning

During early development years of ANNs, only shallow network architectures could be trained successfully owing to restricted computational resources and training strategies at the time. However, such shallow networks could not efficiently learn from low level features to high level concepts automatically to capture relevant patterns in high complex problems. Thus, other simpler models such as SVMs and stacked Autoencoders gained more attention in machine learning tasks. With recent improvements in hardware and development of more efficient training algorithms, it became possible to model complex and abstract non-linear features using deeper neural network architectures constructed using several hidden layers of neurons, and came to be known as *deep neural networks* (DNNs). In 2006, Hinton et al. presented Deep Belief Networks (DBN) [55] where DNNs were pre-trained with Restricted Boltzmann Machines (RBMs) by greedily training each layer of the network using RBM before finetuning the whole network at once. Their work is seen as a breakthrough leading to regained interest in neural networks. DNNs particularly benefit from their innate ability of learning low level features and then automatically capturing advanced abstract information from the input data, however, proper initialization of DNNs is very crucial for training as randomly initialized parameters may result in the training stopping at a local minima which is far away from the optimal solution, thereby resulting in poor performance of the network. For image based recognition and segmentation tasks, Convolutional Neural Networks (CNNs) have recently become very popular. Other interesting deep learning techniques such as Recurrent Neural Networks (RNN) [49] and deep Q-Networks [87] have also found considerable interest. In this thesis, we will focus on the application of CNNs for supervised classification, details of which have been described in the next sections.

5.3 Convolutional Neural Networks

CNNs are a sub-class of DNNs consisting of neurons that perform non-linear operations on input and predict the output in a way similar to ANNs. Starting from an input raw image, a CNN finds learnable weights to optimize the final prediction score. However, the main difference between a CNN and ANN is that CNNs particularly use images as input. Regular Neural Networks are known to not scale well to full images and lead to high computational complexity due to large number of neurons required for modeling full-scale images, and are thus prone to over-fitting due to such large number of parameters. Instead, CNNs encode image specific features into the network architecture and can be efficiently used in the context of images for classification and segmentation tasks. One main advantage of using CNN is parameter sharing i.e. the CNN particularly makes use of the same features across the entire image region. The fact that features found in one part of the image can also be located in other parts of the image is utilized by the CNN. This mainly helps in reducing the overall number of parameters required for training the network by sharing the same weights and bias and limiting the feature set to image focused tasks. Also, CNNs overcome the limitation of traditional ANNs associated with the computational complexity by sharing these learnable weights across the depth (or channels) of the image.

5.3.1 Convolutional Neural Network Architecture

A convolutional neural network consists of several layers that learn a hierarchical representation of features. Different layers commonly employed in convolutional neural networks and their respective function have been described below.

Convolutional Layer

In image processing, a convolution operation is performed by sliding a kernel of specific size (e.g. 3×3 matrix) over image pixels covering the receptive field of the kernel and computing the dot product of each pixel with the corresponding entry of the kernel. The final output is the summation of these dot product entries of the kernel and corresponding pixels. The kernel is then moved one (or more) pixel forward in the image and this process is repeated for each pixel in the entire image region as shown below 5.4. Convolution requires to first flip the kernel *(K)* and then the convolution operation is performed on the input image *(I)*, as shown below.

$$(I * K)(x, y) = \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} I(i, j) K(x - i, y - j), \tag{5.31}$$

where, *w* and *h* are the width and height of the kernel. It should be noted though, that several machine learning libraries utilize the cross-correlation operation which essentially has the same effect as convolution but without flipping the kernel. Based on the kernel type, the convolution process could lead to different effects on the input image such as edge detection, blurring, etc. In context of convolutional neural networks, the convolutional layers exploit this property by employing a set of learnable filters (i.e. kernels) with small receptive fields that produce activation maps as a result of capturing specific features at different spatial positions in the input image. Output of the initial layer captures only low-level features such as the

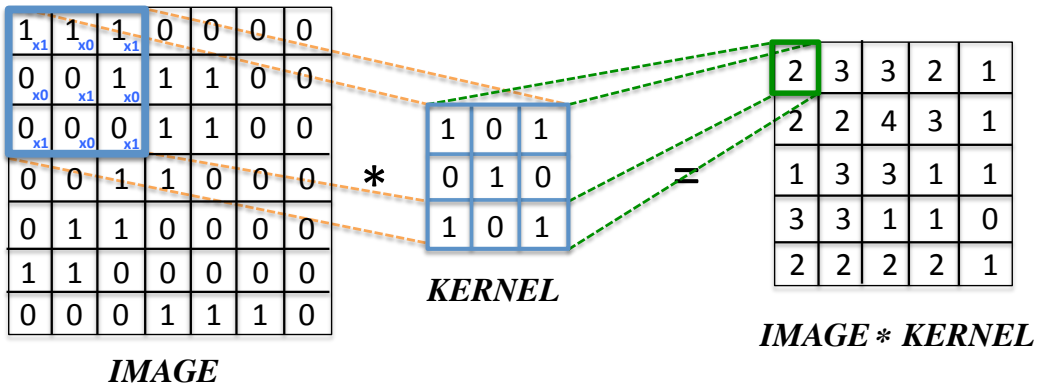


Fig. 5.4. **Convolution Operation.** For a two-dimensional image, I, and a convolution kernel, K of size $h \times w$, by overlaying the kernel on the image and computing sum of the elementwise products between them, image features can be extracted.

edges and to learn a hierarchical representation of the input, this output from the initial layer is fed to convolutions in deeper layers that further extract higher-level abstract features. In contrast to the regular ANNs, the CNNs have essential properties, namely *sparse connectivity* and *parameter sharing* that enable efficient feature learning.

Sparse (Local) Connectivity: When using a kernel size smaller than the input image, the connectivity of pixels becomes local. The center pixel within the receptive field of the kernel looks around only its immediate neighbors. This allows detection of small meaningful features in local vicinity of the pixel. In terms of the neuron architecture, this is equivalent to local connectivity between neurons in adjacent layers.

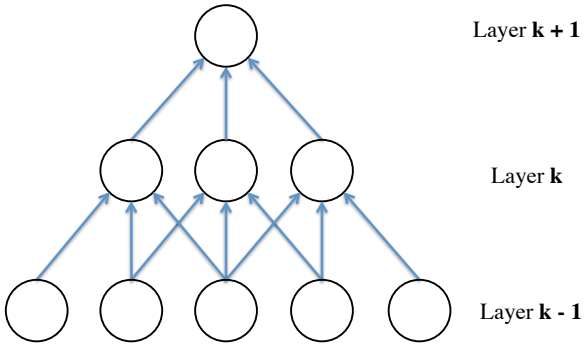


Fig. 5.5. **Local Connectivity.** The neurons in hidden layer k receive their input only from a subset of neurons that are spatially adjacent in layer k-1 (input layer). The overall connectivity of neurons in layer k+1 w.r.t to the input layer k-1 is larger (i.e. with the width of 5) compared to their local connectivity to neurons in layer k (with the width of only 3).

As shown in figure 5.5, the neurons in hidden layer k receive their input only from a subset of neurons that are spatially adjacent in layer k-1 (input layer). Thus, if layer k-1 is the input layer, and the layer k consists of units with a receptive field of width 3, then units in layer k will only be connected to maximum of 3 nearest neurons in layer k-1. Similarly, all neighboring layers will follow this pattern and each unit will ignore any variations outside its own receptive field allowing strong local feature detection. It should be noted that by stacking several such layers also preserves global connectivity. As shown in the figure 5.5, the overall connectivity of neurons in layer k+1 w.r.t to the input layer k-1 is larger (i.e. with the width

of 5) compared to their local connectivity to neurons in layer k (with the width of only 3). This way the overall response of the kernels becomes increasingly global. In comparison to the ANNs where each neuron is connected to every neuron in the next layer, this property of the CNNs leads to fewer parameters and thus lower memory requirement leading to improved efficiency.

Parameter Sharing: In a convolution layer, all spatial locations across the width and height of an input share the same convolution kernel which helps to greatly reduce the number of parameters required by that convolution layer during training and also creates translation invariance for the CNN. These replicated units form a feature map (also known as activation maps) by sharing the same weight vector and bias for all neurons in each input slice. Thus, by accumulating feature maps along the height (i.e. the depth dimension) of the input volume, the final output of the convolution layer is generated. The size of these output feature maps is decided by some hyper-parameters such as the input depth, the stride and the zero padding.

Batch Normalization Layer

Ioffe et al. [62] first introduced Batch Normalization for properly initializing deep neural networks. During training the network, change in internal parameters of every layer leads to a change in the distribution of the inputs supplied to successive layers. This leads to an internal covariance shift problem, which makes it particularly hard to train deep networks even when the input data has been normalized before feeding to the network. In order to counteract this issue, the activations in the network need to follow a Gaussian distribution (with zero mean and unit variance) allowing robust initialization. Thus, by inserting Batch Normalization after every Convolution Layer, normalization is integrated within the network architecture. It has also shown to have beneficial effects on training by allowing higher learning rates for different models and additionally behaving as a regularizer. Lately, another normalization has been introduced in the work of Salimans et al. using Weight Normalization to improve training of deep neural networks [115]. Their method has shown to improve the optimization process and to accelerate convergence of stochastic gradient descent. Their normalization method can be incorporated in recently introduced recurrent models (LSTMs) and other applications such as deep reinforcement learning.

Activation Function

An activation function is required to break the linearity of a network to model complex functions in real world non-linear tasks. A neural network consisting of neurons without an activation function is equivalent to a linear network performing transformations (linear) that are incapable of dealing with non-linear problems even with large number of layers stacked together.

Among different activation functions described previously, the sigmoid function has become rather unfavorable as it can saturate and kill gradients when the sigmoid neuron activates at 0 or 1. The gradient for such regions is very small and almost no signal flows out in this case. Alternatively, if the initial weights are too large then it may lead to saturation of the neurons and the network is unable to learn. Another undesirable property of the sigmoid function is that the subsequent layers of the neural network are not zero centered which affects the

gradient descent process during back propagation making the gradients either all positive or all negative leading to unfavorable gradient updates.

One of the most popular choices for non-linear activation is the ReLU function [76], which can be used to efficiently introduce desired non-linearity into the network. Additionally, the ReLU function is capable of counteracting the "vanishing gradients" problem during back-propagation [41]. When the errors are back propagated, the gradients tend to get smaller further up in the hidden layers leading to slower training. This can be avoided by using the ReLU as the activation function which avoid the vanishing gradients problem because when the input is greater (or equal) to zero, the output of the ReLU is the input, and thus on back-propagation the derivative is equal to one.

Pooling Layer

The pooling operation is used to reduce total number of parameters and amount of computation required for training the network by reducing the spatial size of input feature map, which also helps to control overfitting and provide translation invariance. The most commonly used pooling operation is the max-pooling which sub-divides the input into a set of non-overlapping regions originating from maximum activation positions of the input feature maps as shown in figure 5.6. Other pooling operations such as average pooling [80] and L2-norm pooling have also been suggested previously but max-pooling remains the most popularly used down-sampling operation as it is shown to perform better than the former approaches. One drawback of using the pooling layer is immense reduction in the size of informative feature maps and hence recent attempts have been made to either remove pooling layers, use smaller kernels or by fractional pooling [43].

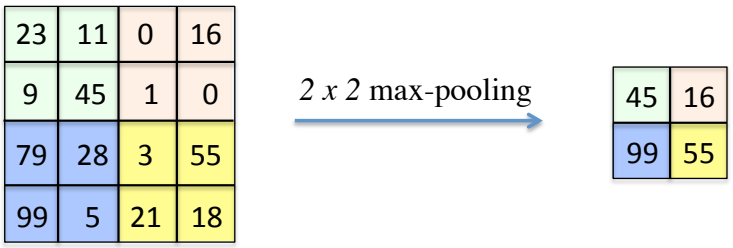


Fig. 5.6. **Pooling Operation.** A 2×2 max-pooling operation sub-dividing the input feature map into a set of non-overlapping output regions that originate from maximum activation positions in the input feature map 2×2 region.

Fully Connected Layer

The final layer in the architecture before computing the loss/error function is the fully connected layer, which consists of full connections between neurons of successive layers. This way, a fully connected layer behaves similar to the ANNs and such a configuration is generally employed for classification or regression tasks. It is possible to convert these fully connected layers into *fully convolutional layers* by using the kernel size to be the same as the size of the input and thus converting full neuron connections to local region in the input.

Loss Layer: Minimizing Error Propagation

CNNs use similar strategy of backpropagation that was previously explained for ANNs. A popular choice for classifying multiple labels is given by a *softmax classifier function* which provides more intuitive output in terms of normalized class probabilities. Consider a training set of input images $x_k \in R^D$, each associated with output label $y_k \in \{1, ..., C\}$. A softmax function f can be written as:

$$f_c(z) = \frac{e^{z_c}}{\Sigma_{i=1}^{C} e^{z_i}}, \tag{5.32}$$

where, it transforms its input such that it can be interpreted as a probability distribution of C class labels. This transformation from the softmax function can be utilized by the *cross-entropy loss* given by:

$$E_{crossEntropy}(t, q) = -\boldsymbol{\Sigma}\, t(x)\, \log\, q(x), \tag{5.33}$$

where, the t is the true distribution while, q is the estimated distribution predicted by the network. Thus, the softmax classifier minimizes the cross-entropy between the estimated class probabilities q given by:

$$q = \frac{e^{f_{y_c}}}{\Sigma_{k=1}^{C} e^{f_{y_k}}}, \tag{5.34}$$

and the true distribution t of the correct class labels given by $t_{k=1}^{c} \in \{0, 1\}$.

5.3.2 Training a CNN

After designing the appropriate architecture, training the CNN requires few checks before initiating the training process. These include input data preparation, regularization and optimization schemes as describe below.

Data Preparation

Before feeding the input images to the CNN, as a pre-processing step, it is common to randomly shuffle the training set to avoid any meaningful order which may bias the optimization algorithm. It is also beneficial to perform mean subtraction and normalization on the dataset. By subtracting the mean of training dataset from all the pixels in each input image, the dataset is essentially centered around the origin. Furthermore, if the input features have different scales, it is beneficial to perform data normalization i.e. shifting scale in the range from 0 to 255.

Regularization

When a network cannot effectively learn due to reduced generalization of features it leads to an overfitting problem wherein the network starts to fit the noise along with the training data. Such a network with poor generalization of the features performs poorly when testing on unseen dataset. In order to overcome this, different regularization techniques such as $L1/L2$ regularization, dropout, Max Norm or also simple techniques such as data augmentation can been used.

Optimization

One of the most popular way to train a neural network involves gradient-based training algorithms that are usually known to converge fast. The training set performance is given by an empirical risk measure while the expected performance provides the performance on the future examples. Minimizing the empirical risk has been suggested over expected risk in the statistical learning theory. Using the gradient descent to minimize the empirical risk updates weights of the network.

Recently, the *Stochastic Gradient Descent* (SGD) has been proposed for use over the traditional gradient descent particularly when the training dataset is large. For each iteration, the SGD estimates the gradient using randomly picked examples. The stark difference between gradient descent and SGD is that the gradient descent is run through all training examples for a single update of a parameter, however, in case of SGD, only a subset of the training examples is used to update a parameter in each iteration. SGD often converges faster in comparison with gradient descent. However, the error function might not be well minimized for SGD but the close approximation is generally enough to optimal values.

5.4 Automatic Segmentation in ADPKD using Convolutional Neural Networks

In recent years, CNNs have shown superior performance in several computer vision tasks such as image classifcation, object detection and semantic segmentation. The main advantage of CNNs in comparison to many other machine-learning-based methods, such as random forests, is that they do not require hand-crafted features. In the domain of medical imaging, CNNs have previously been proposed for localization and segmentation of kidneys with mild morphological changes using patch-wise approaches on CT [132, 157]. In this work, a novel method is presented for automated segmentation of ADPKD kidneys using fully convolutional neural networks, trained end-to-end, on slicewise axial-CT sections. The method has been assessed for its qualitative and quantitative accuracy and precision to measure TKV on large CT dataset of patients at different stages of ADPKD. The proposed approach facilitates fast and reproducible measurements of TKV in agreement with manual segmentations from clinical experts.

5.4.1 Patients: Clinical Characteristics

The dataset for our experiments consisted of 244 CT acquisitions from ADPKD patients enrolled in three independent clinical trials on ADPKD. These acquisitions comprised of baseline and follow-up CT images from study 1 (SIRENA), study 2 (SIRENA 2), and study 3 (ALADIN 2). The main clinical and demographic characterisitics of the patients are summarized in table 5.1.

The SIRENA clinical trial (study 1) [94] (ClinicalTrials.gov Identifier: NCT00491517) was a randomized cross-over study that compared changes in kidney volume and its compartments in adult ADPKD patients (> 18 years) with normal renal function or mild to moderate renal insufficiency over 6-month treatment of sirolimus or conventional therapy. Sirolimus (also known as Rapamycin) is known to exert antiproliferative effects due to inhibition of mTOR that regulates cellular metabolism and growth. This could prove to be pertinent for inhibiting cyst progression in ADPKD, eventually halting kidney disease progression in ADPKD. Initially, 21 ADPKD patients were recruited for the study but only 15 patients (12 Male and 3 Female) between 28 and 46 years of age completed the study. Among these patients, 7 were randomly

Clinical Study	**Gender**	**Number of Acquisitions**	**Age (years)**	**Estimated GFR (eGFR)**	**Total Kidney Volume (ml)**
	Female/Male	*Training Set* *Test Set*	*[Range]*	*(ml/min per 1.73 m^2)*	*Mean ± SD* *[Range]*
SIRENA	3/12	26	39.1	eGFR $\geq$ 40	1,891.4 ± 1,073.2
(study 1)		26	[28 - 46]		[501.9 - 5,093.2]
SIRENA 2	24/17	45	53.8	15 $\leq$ eGFR $\leq$ 40	3,139.1 ± 1,485.5
(study 2)		15	[41 - 70]		[1,197.1 - 6,634.1]
ALADIN 2	32/37	94	53.6	15 $\leq$ eGFR $\leq$ 40	3,132.7 ± 2,152.2
(study 3)		38	[33 -74]		[321.2 - 14,670.7]

Tab. 5.1. **Demographics and Clinical Characteristics of ADPKD Patients.** ADPKD patients (n=125) with baseline and follow-up CT acquisitions (training set = 165, test set = 79) included in our study.

assigned to sirolimus followed by conventional treatment and 8 to conventional followed by sirolimus therapy. The enrollment criteria was $eGFR \geq 40ml/min$ per 1.73 m^2, and 24-hour urinary protein excretion rate of 0.3 g. The average TKV of patients in our experiments from SIRENA study was 1,891.7$\pm$ 1073.2 ml (TKV range: 501.9 ml - 5,093.2 ml).

The SIRENA 2 clinical trial (study 2) [110] (ClinicalTrials.gov Identifier: NCT01223755) was a randomized and parallel group trial that compared changes in GFR on 3-year treatment with sirolimus added on to conventional therapy (21 patients) or conventional treatment alone (20 patients) in ADPKD patients with moderate/severe renal insufficiency ($15 \leq eGFR \leq 40ml/min$ per 1.73 m^2) in 1 and 3 years versus baseline. The average TKV of patients from SIRENA 2 included in our experiments was 3,139.1 $\pm$ 1,485.5 ml (TKV range: 1,197.1 ml - 6,634.1 ml).

The ALADIN 2 Study (ClinicalTrials.gov Identifier: NCT01377246) was a multicentric, randomized longitudinal study (3-years) that assessed the efficacy of treatment with long-acting somatostatin analogue (Octreotide LAR) compared with placebo in slowing kidney and liver growth rate in the ADPKD patients with moderate/severe renal insufficiency ($15 \leq eGFR \leq 40ml/min$ per 1.73 m^2). The average TKV of patients from ALADIN 2 included in the experiments is 3,132.7 $\pm$ 2,152.2 ml (TKV range: 321.2 ml - 14,670.7 ml).

5.4.2 CT Image Acquisition

The CT images of study 1 and study 2 were acquired at single centre in Bergamo, while CT images of study 3 were acquired at four different centres (Bergamo, Naples, Agrigento and Treviso) in Italy. All the above CT images from the three clinical trials were acquired with a 64-slice CT scanner (LightSpeed VCT; GE Healthcare, Milwaukee, WI) using a single breath-hold scan (120 kV; 150 to 500 mAs; matrix 512x512; collimation 2.5 mm; slice pitch 0.984; increment 2.5 mm) initiated 80 seconds after the injection of 100 ml non-ionic iodinated contrast agent (Iomeron 350; Bracco, Italy) at a rate of 2 ml/s, followed by 20 ml physiologic solution at the same injection rate.

5.4.3 Data Annotation and Experimental Setup

The image sequence for each CT acquisition was accessed using ImageJ software (version 1.48v) [1] and manually delineated along the border of right and left kidney seperately by clinical experts and trained personnel. The boundary delineation was performed using a standard protocol for all kidneys with respect to the hilum and liver cysts. Final manual segmentations were checked and corrected by a single operator to avoid inter-rater bias during the segmentation process.

For our main experiment, the CT acquisitions (n=244) were manually divided into the training (n=165) and test set (n=79), trying to achieve a similar distribution in both sets based on the available TKV range (321.2 ml - 14,670.7 ml), see table 5.2. We performed an additional 3-fold cross validation on same dataset by sorting the dataset according to ascending TKV range and then randomly partitioning into 3 sub-sets (n=242: 80, 81, 81) splitting it uniformly into the 3 cross-validation sets. From the original dataset of 244 acquisitions, 2 cases (TKV>13,000 ml) were removed from the cross-validation set due to their non-representative nature in the

entire patient population providing inadequate number of images for learning such rare cases. This was confirmed by our feature visualization experiment for one of these patients, shown in figure 5.10 (bottom). For the remaining cases (n=242) in the 3 cross-validation sets, 2 sets were used for training and the remaining set for testing. This process was repeated three times such that every data set was used once for testing.

5.4.4 Data Augmentation

Two separate augmentation methods were applied on the training dataset to mitigate overfitting and achieve good generalisation. Firstly, each CT image was shifted in x-y direction (rigid translation of 32 pixels each in x and y direction), and secondly, by non-rigidly deforming the respective slice and applying a low frequent intensity variation. Both augmentation methods were performed using commercial software package Matlab [83]. This increased the training dataset from 16,000 CT slices to 48,000 CT slices in case of main experiment. Both augmentation steps are shown in figure 5.7

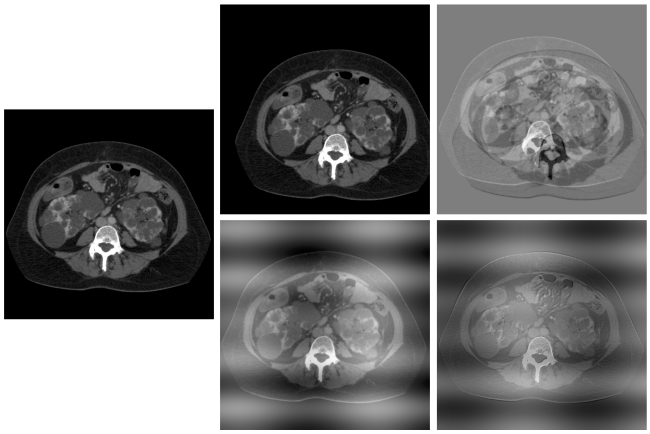


Fig. 5.7. **Data Augmentation.** Left: Original patient CT image; Top Centre: Image obtained by first augmentation strategy, Top Right: Difference image from orginal and shifted image; Image obtained by second augmentation strategy: Deformation Image, Bottom Right: Difference Image from original and transformed image.

5.4.5 Convolutional Neural Network Architecture

As shown in figure 5.8, our CNN architecture follows the VGG-16 representation [121] but consists of only the first 10 layers of convolution filters with a receptive field of 3×3 and a spatial padding of 1 pixel for every convolution layer. Due to the internal covariance shift problem, it is rather difficult to train deep neural networks with saturating non-linearities as explained by Ioffe et al.[62]. Thus, a batch normalization layer was employed after every convolution to properly initialize the network by compelling activations to follow standard Gaussian distribution and normalizing the inputs to zero mean and unit variance. Application of batch normalization was crucial to improve the overall accuracy of the network in our experiments. After batch normalization, a layer of neurons with the Rectified Linear Unit (ReLU) [76] activation function is used. In order to reduce the number of parameters, and thus the computation complexity of the network, max-pooling layers with a 2×2 pixel window

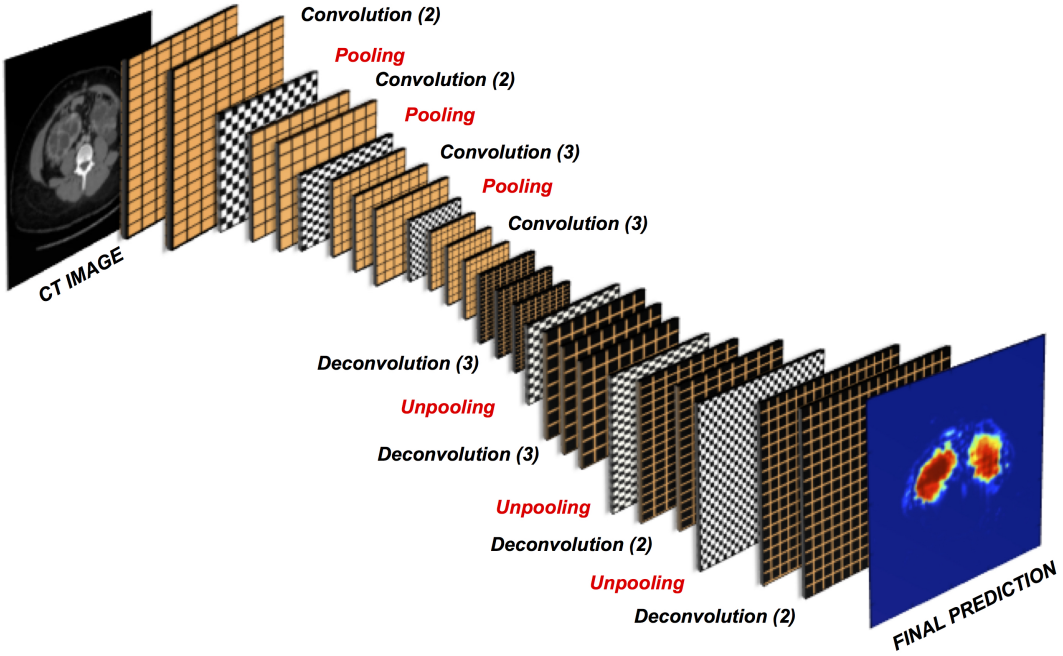


Fig. 5.8. **Fully Convolutional Neural Network Architecture.** For feature extraction step, we used 10 layers of convolution filters with a receptive field of 3×3 and spatial padding of 1 pixel followed by max pooling layers with 2×2 pixel window and stride of 2 pixels to progressively reduce the spatial size of the input after convolution step. To achieve pixelwise segmentation, deconvolution and unpooling layers were used for upsampling the feature maps.

and a stride of 2 pixels were used to progressively reduce the spatial size of the input to half the original size along both the height and the width. The above layers correspond to the feature extraction step that are followed by a series of deconvolution and unpooling layers for up-sampling the feature maps for pixelwise segmentation, following Zeiler et al.[155] and Noh et al.[89]. The deconvolution layers perform convolution like operation but in the opposite way leading to upsampling of coarse feature map into the reconstructed shape of the input. But, performing only deconvolution leads to coarse reconstructions and therefore, unpooling layers help to refine the output. These unpooling layers reverse the pooling operation by preserving the locations of maximum activation that were extracted during the max-pooling step, and re-utilising these locations to place the maximum activations back to their original spatial position [155]. The unpooling layers and the deconvolution layers are finally connected to a 1×1 convolutional layer that maps the last feature vector to the desired foreground (kidney) and background (non-kidney) classes. The output of the final convolutional layer is connected to the cross-entropy loss to optimize the weights by penalizing deviation between true and predicted labels.

5.4.6 Training and Testing

All experiments were performed using the Caffe[64] framework. Before feeding to the CT images to the network for training, th original size of the CT slices 512×512 was re-sampled to 224×224 and the image range was normalized to [0,255] in order to reduce the computational complexity of the network. Additionally, a mean subtraction from the training dataset was done as a pre-processing step and finally the CT slices were randomly shuffled before training

Threshold	Sensitivity	Specificity	Youden Index	Accuracy	Precision	F1 Score
0.1	0.97	0.95	0.92	0.95	0.60	0.72
0.2	0.95	0.98	0.93	0.98	0.75	0.83
0.3	0.93	0.98	0.92	0.98	0.80	0.85
0.4	0.92	0.99	0.90	0.98	0.83	0.86
0.5	**0.90**	**0.99**	**0.89**	**0.98**	**0.85**	**0.87**
0.6	0.88	0.99	0.87	0.98	0.87	0.87
0.7	0.86	0.99	0.85	0.98	0.89	0.87
0.8	0.82	1.00	0.82	0.98	0.91	0.86
0.9	0.71	1.00	0.71	0.97	0.91	0.79

Fig. 5.9. **Threshold Selection.** Qualitative metrics for different thresholds. As shown in the figure, 0.5 provides the optimal cut-off for threshold selection.

the network. Using a sample size of 8 (batch-size) for each iteration, training was performed on a workstation with an Intel Xeon 8-core 2.40 GHz CPU and a NVIDIA GeForce TITANX (12 GB) GPU. Training the network took approximately 24 hours. The optimization of the network during training was done using the adaptive gradient ("AdaGrad") method with learning rate set of 0.0001, and a weight decay of 0.0005, respectively. The weights were initialized in both convolution and the deconvolution layers using the xavier initialization [41].

In the test phase, previously unseen 9,000 CT slices (79 CT acquisitions) that were not included in the original training phase were used for prediction from the pre-trained network. The output predictions consisted of the foreground (kidney) and the background (non-kidney) pixels, where pixels with a probability higher than 0.5 were regarded as foreground (kidney) pixels. The threshold was selected based on the analysis of the Receiver Operating Characteristic (ROC) space along with the Accuracy, Precision, F1 Score and the Youden-Index (to maximize both sensitivity and specificity). The results of threshold analysis indicate that 0.5 yields the best compromise of the metrics and therefore it is selected for generating the final segmentation results. The results from the analysis on different thresholds have been summarised in figure 5.9.

5.4.7 Feature Visualization

We adapted the occlusion method from Zeiler and Fergus [155] to understand the importance of context and the corresponding learned feature in the image segmentation domain. This approach has been previously used for measuring the change in classification while occluding part of image with a gray square of fixed size and constant intensity value in a sliding window fashion in order to obtain the importance of each image region. We retrieve full segmentation maps from the CNN network instead of prediction scores for classification. Thus, we compute the change in DSC with respect to the original unoccluded image as shown in figure 5.10.

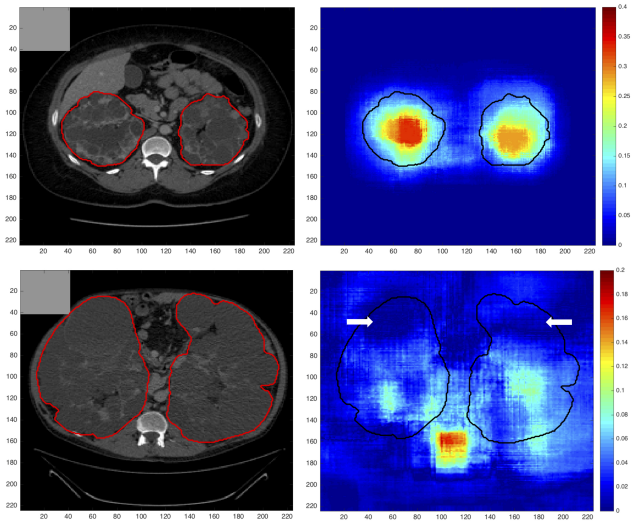


Fig. 5.10. **Feature Visualization.** Segmentation maps measuring change in DSC while occluding (shown with gray square) different parts of the image with respect to original unoccluded image as a measure of importance for the respective image region. The manually generated outline of the kidney is shown in red and black, respectively. Top: The largest change occurs when the kidneys themselves are occluded. Bottom: Same experiment for an ADPKD patient with high TKV (>13,000 ml).

5.4.8 Statistical Analyses

The performance of our automated segmentation method was evaulated by computing the dice score coefficient (DSC)[37] to assess the spatial overlap accuracy of the predicted and true manual segmentation labels. The Concordance Coefficient Correlation (CCC) measure was used to evaluate the reliability and reproducibility of the automated method with respect to the standard manual method. Bland–Altman analysis was used to assess the agreement between TKV estimated from the automated segmentation method and the corresponding manual segmentation. The COV for repeated measures[66] (computed as the ratio of standard deviation to mean of the measurements) between true and automated TKV was also computed. The non-parametric Wilcoxon signed rank test for paired samples was used to measure the statistical significance of correlation between automated and true TKV measurements. Additionally, Spearman's rank correlation coefficient (ρ) was employed as a non-parametric test to measure the strength of association between the automated and true TKV measurements. Finally, we computed the MAPE and RMSE with respect to TKV in order to provide error measures which are not sensitive to under-estimations and over-estimations cancelling out each other. All statistical analyses were performed using R studio [128] version 0.98.953.

5.4.9 Total Kidney Volume Quantification

First, we resampled the 224×224 segmentation predictions obtained from the CNN back to original size of 512×512 using bicubic interpolation method. Then, we performed a morphological closing operation to recover potential holes within predicted kidney regions and to remove any small isolated noise pixels wrongly predicted as foreground (kidney) pixels. Finally, TKV was computed as the product of number of foreground pixels multiplied with the pixel spacing in x and y direction and the corresponding slice thickness.

5.4.10 Results

For our experiments, baseline and follow-up CT acquisitions (training set = 165; test set = 79) of ADPKD patients ($n = 125$) with a wide range of TKV (321.2 ml – 14,670.7 ml) and an estimated glomerular filtration rate $(eGFR) \geq 40ml/min$ per 1.73 m^2 (study 1) or $15 \leq eGFR \leq 40ml/min$ per 1.73 m^2 (study 2 and study 3) were used. The studies 2 (SIRENA 2) and 3 (Aladin 2) depict similar clinical characteristics, therefore results of segmentation performance and TKV computation were combined together for these two studies. The proposed method was assessed for its segmentation performance with respect to ground truth manual annotations from experts. Additionally, volumetric measurements were made on the kidney segmentations from the CNN (hereafter referred as: *automated TKV*) and compared with *true TKV* measurements (obtained from ground truth annotations) to assess their agreement, accuracy and precision.

Segmentation Performance Analysis

The overall mean DSC between segmentations from the automated method and ground truth kidney segmentations from clinical experts was 0.86 ± 0.07 (mean ± SD) for the entire test set (n=79). In particular, in study 1, consisting of patients with mild to moderate renal insufficiency (n=26, table 5.1), the mean DSC was 0.86 ± 0.06. For studies 2 and 3 (combined) consisting of patients with moderate/severe renal insufficiency (n=53, table 5.1), the mean DSC was 0.86 ± 0.08. The segmentation predictions from automated CNN method from 4 different patients have been shown in figure 5.11. The time required for prediction of segmentation using automated CNN method was only few seconds per patient CT acquisition, while the manual segmentation from experts required approximately 30 minutes per patient. In addition, three example cases of final segmentations generated using automated CNN method have been shown (figure 5.12) in comparison with the manual segmentation masks used as gold-standard. The segmentation masks from automated CNN method were generated using optimal threshold of 0.5 as explained in section 5.4.6.

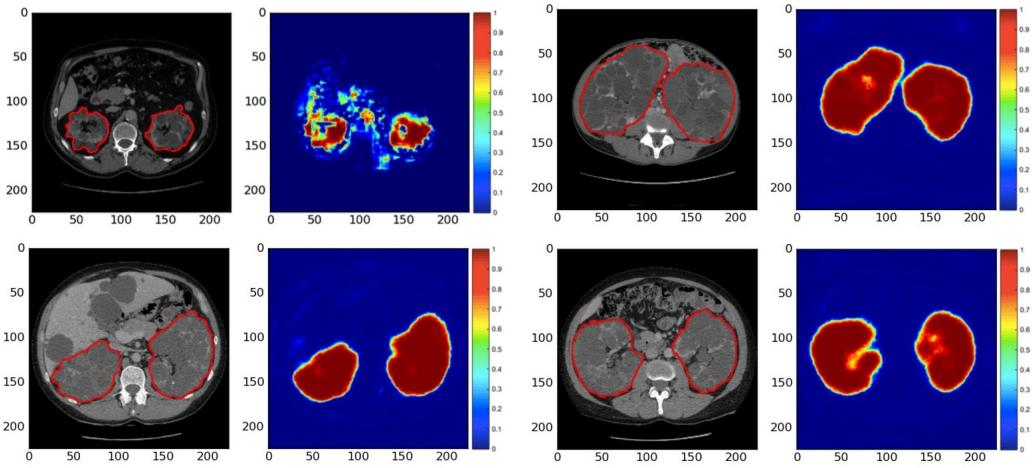


Fig. 5.11. **CNN Predictions of ADPKD Kidneys.** Four segmentations (red contour) of ADPKD kidneys from CT acquisitions of different patients are shown. The corresponding CNN-generated probability maps are shown in pseudo colors.

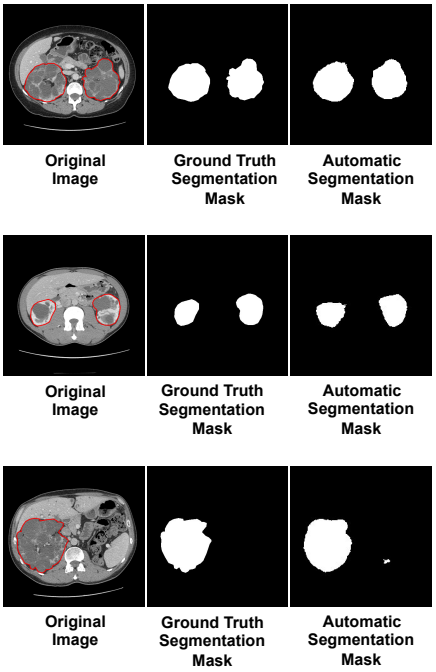


Fig. 5.12. **CNN Segmentation Masks of ADPKD Kidneys.** Segmentation masks generated from CNN predictions (i.e. threshold > 0.5) in comparison with ground truth masks generated from manual segmentations (i.e. gold-standard) of ADPKD Kidneys (red contour) from three different cases (shown original images). Foreground (kidney) pixels are denoted as white while the background (non-kidney) pixels are denoted as black.

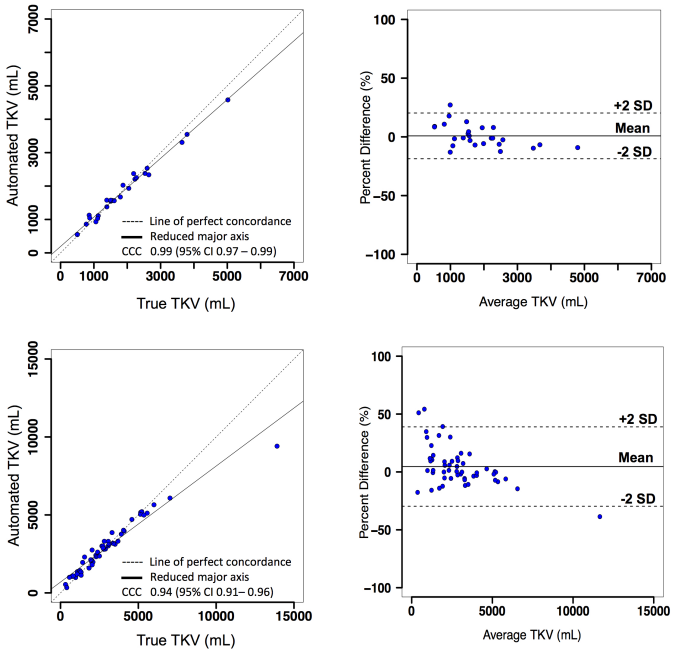


Fig. 5.13. **Left: Concordance Correlation Coefficient (CCC) plots showing strength of association; Right: Bland-Altman plots showing agreement between TKV measurements.** TKV measurements from automated segmentation method (main experiment) are compared with true TKV measurements from manual segmentations for study 1 (top, n=26) and, studies 2 and 3 (bottom, n=53).

TKV Concordance Analysis

We performed volumetric measurement on kidney segmentations from the CNN and compared the automated TKV with the true TKV (obtained from ground truth annotations) in terms of accuracy and precision of the measurement. As shown in figure 5.13 (top-left), for study 1 (test set = 26), there is substantial strength of association between the automated and true TKV with a concordance correlation coefficient (CCC) of 0.99 [95% Confidence Interval (CI): 0.97 - 0.99]. The mean TKV error between automated and true measurements was -32.9 $\pm$ 170.8 ml and the mean percentage TKV error was 1.3% $\pm$ 10.3% while the mean absolute percentage error (MAPE) was 7.8% $\pm$ 6.7%. The Bland Altman plots used to determine the agreement between the two methods are shown in figure 5.13 (top-right) with the lower and upper limits of agreement (LOA) for percentage difference of -18.6% and 20.3%, respectively. The coefficient of variation (COV) between true and automated TKV was 6.5%.

For clinical studies 2 and 3 (n=53, table 5.1), the automated and true TKV measurements showed moderate strength of association with a CCC of 0.94 [95% CI: 0.91 - 0.96] between automated and true measurements. The mean TKV error was -44.1$\pm$ 694.5 ml while the mean percentage TKV error was 6.5% $\pm$ 20.1% and MAPE was 13.3% $\pm$ 16.4%. The COV between automated and true TKV was 17%. On the Bland-Altman plots, figure 5.13 (bottom-right), the lower and upper LOA were -29.6% and 38.9%. The overall COV between true and automated TKV was 17%. The difference in the TKV measurements was found to be statistically insignificant for all the three clinical studies ($p > 0.05$). High positive correlation was observed between automated and true measurements for all the clinical studies. Study 1 showed the mean correlation coefficient (Spearman's rho) ρ of 0.97 ($p < 0.001$) and ρ of 0.98 ($p < 0.001$) for studies 2 and 3, respectively. In some cases with severe liver cysts in close proximity of the kidney, the total kidney volume was over-estimated due to inclusion of these cysts as false positive regions in the kidney segmentation. Example predictions of such mislabeled regions by the CNN are in figure 5.14.

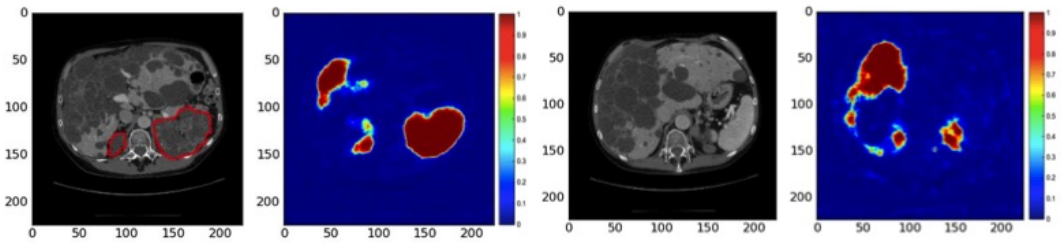


Fig. 5.14. **Mislabelled Predictions by CNN.** Left: Liver cysts predicted as foreground along with kidney region; Right: Cystic liver mislabelled as Kidney.

Cross Validation Analysis

Additonally, the 3-fold cross-validation confirmed the performance of CNN based TKV estimation model. For the three cross-validation sets DSC scores were recorded as 0.86 $\pm$ 0.1, 0.83 $\pm$ 0.8 and 0.87 $\pm$ 0.6, respectively. The COV for all three sets ranged from 14 to 15 and the root mean squared percentage error (RMSPE) ranged from 19 to 21. Results have been compiled in table 5.2.

	Main Experiment	Cross Validation 1	Cross Validation 2	Cross Validation 3
TRAINING				
Total Acquisitions	165	162	161	161
Total Original Images	16,000	16,079	15,706	15,544
Training Images: Including Augmentation	48,000	48,237	47,118	46,632
TESTING				
Total Acquisitions	79	80	81	81
Total Original Images	9,000	8,634	9,007	9,169
Mean TKV (ml)				
(mean ± SD)	2,549.0 ± 1,951.1	2,729.0 ± 1,499.3	2,780.0 ± 1,560.0	2,816.4 ± 1,669.7
[Range]	[321.2 - 13,913.6]	[399.0 - 7,034.5]	[501.9 - 7,480.6]	[321.0 - 9,605.6]
Dice Score Coefficient *(DSC)* (mean ± SD)	0.86 ± 0.1	0.86 ± 0.1	0.83 ± 0.8	0.87 ± 0.6
Mean Percentage Error *(MPE)* (mean ± SD)	4.8 ± 17.6	0.2 ± 21.1	6.8 ± 20.8	-1.3 ± 18.9
Mean Absolute Percentage Error *(MAPE)* (mean ± SD)	11.5 ± 14.1	15.1 ± 14.5	15.0 ± 15.8	13.5 ± 13.1
Concordance Correlation Coefficient *(CCC)*	0.95	0.93	0.91	0.94
Coefficient of Variation *(COV)*	16	14	15	15
Root Mean Square Error *(RMSE)*	573	527	587	564
Root Mean Square Percentage Error *(RMSPE)*	18	21	22	19

Tab. 5.2. **Cross Validation Experiments.** 3-fold cross-validation to asses the performance of our fully convolutional neural network.

5.4.11 Conclusion and Discussion

In this study, a novel method was presented to automatically segment polycystic kidneys, and its qualitative and quantitative accuracy and precision to measure TKV was investigated on a large dataset of CT acquisitions from ADPKD patients. The annual increase in TKV has been estimated to be around 5% [45, 48] per year, suggesting that TKV measurement should be accurate to capture small changes over time. As described in a previous chapter of this thesis, the most commonly used methods for kidney volume computation such as manual delineation and stereology[8] are simple but time consuming and subject to intra/inter-observer variability. Alternatively, the mid-slice method[7] and the ellipsoid equation[52, 63] serve to provide quick TKV measurement but lead to low accuracy and precision compared to whole kidney segmentation.

On MRI, Racimora et al.[103] proposed a segmentation approach yielding mean percentage TKV error of 22.0% ± 8.6% with an automated active contour algorithm that reduced to 3.2% ± 0.8% after manual post-editing efforts. Another semi-automatic approach with geodesic active contours and watershed edge detection[73] achieved high accuracy with mean TKV difference of 0.19% ± 6.96%. Mignani et al.[86] compared their results with stereology and reported mean percentage TKV error of – 0.6 ± 5.8%, while, Turco et al.[140] reported MAPE of 4.4% ± 4.1% and 4.2% ± 4.0% for the left and right kidneys, respectively. Other supervised segmentation methods based on stereology[147] on MRI and random forests on CT[119] have also been reported previously. However, these semi-automatic techniques are subject to intra/inter-observer variability and mostly require post-processing efforts to achieve higher accuracy leading to increase in overall processing time of TKV. Kline et al.[72] proposed automatic segmentation on follow-up MR images, however, their method essentially requires previously performed manual segmentations of kidneys on baseline images as initialization

for the segmentation process. In the work of Kim et al.[68], a level set framework has been proposed for automatic segmentation in ADPKD. Even though their method shows good correlation between automated and manual TKV measurements, the results indicate high variability (LOA higher than $\pm$ 25%) when compared with the manual method. Zheng et al.[157] used patch-based CNN in combination with marginal space learning for localization of pathological kidneys prior to an active shape model for segmentation. Their results show good segmentation accuracy (DSC $>$ 0.88) but there is substantial increase in segmentation error without CNN initialisation. Also, the presented dataset in their work appears to contain kidneys with milder morphological changes.

In this work, the performance of the proposed automated segmentation method was assessed both quantitatively and qualitatively on a large CT dataset (n=244) of patients at different stages of ADPKD, using manual segmentations from clinical experts as gold standard. For study 1 with ADPKD patients at early stage of the disease and TKV range between 500 ml and 6,000 ml, the automated TKV shows very high strength of association (CCC = 0.99) with true TKV, however, for studies 2 and 3, with ADPKD patients at more advanced stage of the disease and the TKV range between 300 ml and 15,000 ml, there is moderate strength of association (CCC = 0.94) between the two methods. Similarly, the overall accuracy and precision of the TKV measurements from automated method is higher for study 1 (MAPE = 7.8% $\pm$ 6.7%; COV = 6.5%), compared to studies 2 and 3 (MAPE = 13.3% $\pm$ 16.4%; COV = 17.0%). The performance of the automated method is decreased particularly for very low TKV ($<$ 500 ml) and for extremely high TKV ($>$ 6000 ml). This can be attributed to availability of very few instances of such small or very huge kidneys leading to poor predictions by the CNN during testing phase. However, the overall difference in TKV measurements was found to be statistically insignificant ($p > 0.05$) for all three clinical studies and the automated TKV measurements show high positive correlation with true TKV measurements (ρ = 0.98, $p < 0.001$). Moreover, the proposed method takes only few seconds for prediction of segmentation on each patient acquisition and avoids any intra/inter operator segmentation bias.

Despite the promising results, our study has some limitations. In some cases with several liver cysts in close proximity of the kidney (figure 5.14), the automated segmentation method over-estimated the kidney volume due to inclusion of liver cysts in the segmented kidney region. To potentially overcome this problem, the proposed method can be trained on 3D volumes of polycystic kidneys. Regarding the importance visualization in figure 5.10, the importance of context can be visualized for the segmentation-especially for very rare subjects with extremely high TKV: In case of a typical patient, as seen in (figure 5.10 (top)), the largest change occurs when the kidneys themselves are occluded. This is not only intuitive but also confirms that the network is not confused by changes far away from the regions of interest. This highlights robustness against changes far away from the region of interest. Nonetheless, the visualized influence region extends over the object boundaries which indicates that not only the kidneys themselves, but also local context is used to find the final segmentation. For very rare cases with extremely high TKV (>13,000 ml) though, the spatial context changes entirely due to both kidneys occupying most of the abdominal region. As a consequence, the CNN cannot exploit context information leading to poor segmentation results which is also confirmed by the feature visualization experiment (figure 5.10 (bottom)): Consider particularly the upper areas (indicated by white arrows) in the annotated kidneys which

exhibit low variation as kidney tissue does typically not appear in these areas indicating that the CNN is not expecting kidney tissues in this area.

In conclusion, a fully automated method was presented for the segmentation of polycystic kidneys from patients at different stages and severity of ADPKD using CT data. In comparison with majority of the methods previously reported on TKV computation in ADPKD, our method has been evaluated on a larger TKV range ($>$ 300 ml and $<$ 15,000 ml) and, it allows fast and reproducible TKV measurements in good agreement with manual segmentations from clinical experts. Our method can be reliably used on a TKV range of $>$ 500 and $<$ 10,000 ml, facilitating fast and reproducible measurements of kidney volumes in agreement with manual segmentations from clinical experts. The overall segmentation can be further improved by incorporating user interaction to correct mislabelled sections of CT. Particularly for high resolution CT images, this can significantly reduce the TKV computation time compared to manually tracing every section of the kidney and also, capture smaller changes in TKV over time. As a future work, the automated method can be trained on other affected organs such as the polycystic liver for computation of the liver volume in ADPKD. Moreover, the proposed method may be extended to MRI by specifically tuning the parameters used during training the CNN for MRI images.

Part III

Conclusion and Outlook

6 Conclusion and Outlook

As a conclusion of this dissertation, we summarize the main contributions and discuss possible directions for future research on segmentation strategies in ADPKD. To gain insight into the segmentation problem, a detailed overview was provided in the first chapter on normal kidney anatomy and the morphological modifications appearing in the kidney as a consequence of ADPKD. The irregular and sustained renal cyst development causing these morphological alterations can provide crucial information about disease severity and progression by assessing changes in the kidney volume. In this respect, TKV has been acknowledged as an important imaging biomarker and employed in several clinical studies as a primary end-point for investigating potential drug therapies in ADPKD. Imaging techniques such as CT or MRI can be used for adequately monitoring and assessing TKV changes in patients at different stages of ADPKD. As discussed in chapter 2, traditional segmentation approaches can prove to be insufficient in modelling high complexity of the polycystic kidneys owing to irregular variations in kidney shape, size, intensity inhomogeneities within the kidney and unclear boundaries in the presence of liver cysts. In chapter 3, we compared different methods currently being used or recently proposed for TKV computation in terms of reproducibility, reliability, and time required in order to determine the most reliable method for use in the clinical trials. Our results suggest that planimetry based methods relying on whole kidney contouring in each 2D slice of CT or MRI should be preferred over fast and simplified techniques such as the mid-slice method or the ellipsoid equations to accurately monitor TKV changes in ADPKD clinical trials. Moreover, we found that expert operators are required for performing reliable estimation of kidney volume, especially on MR images and, using efficient TKV quantification methods considerably reduces the number of patients required for enrolment in clinical investigations thereby making such studies more feasible and significant. The main disadvantage of currently employed methods based on planimetry or stereology for TKV measurement is that they tend to be rather time consuming, especially when using slices with high spatial resolution for manual segmentation or when employing very finely spaced grids for point counting in stereology. Furthermore, these methods are prone to intra-rater or inter-rater variability. The limitations of these TKV computation methods provide good motivation for finding new strategies that can either assist delineation of the polycystic kidneys or completely automate the segmentation process to aid TKV computation.

In addition to the clinical relevance of this segmentation problem, it also presents as an interesting and challenging case of pattern recognition in machine learning. Since the last decade, learning-based approaches have been successfully used in the domain of medical imaging and algorithms for pattern recognition and classification have become widely popular in improving machine intelligence for different tasks including medical image segmentation. In this dissertation, we investigated the applicability and performance of two separate machine-learning approaches, based on efficient feature engineering and representation learning, respectively, for identifying the underlying patterns within the CT imaging dataset of ADPKD patients for segmentation of polycystic kidneys.

In the first approach, as discussed in chapter 4, a random forest classifier was used in a divide and conquer partitioning approach for segmentation of polycystic kidneys using CT dataset acquired on ADPKD patients at late stage of the disease. Thus, a piece-wise posterior model was created by partitioning over the full feature space using simple binary decisions, and the posterior distribution was modelled in each leaf of this feature space. The features employed in the classification forest consisted of additional information from geodesic distance volumes that contained intensity-weighted distances to a manual outline of the respective kidney in its middle slice (for each kidney) of the CT volume. Therefore, by defining an objective function and designing the posterior model within the leaf, we aimed at training a decision rule by means of a random forest classifier to label separate classes based on kidney voxels or the background.

Mostly, a classification formulation of random forests seems to be a natural choice, however, simple classification is not always the most appropriate option as it often relies on local visual context information or suffers from unsuitable choice of hand-crafted features for segmentation. However, other efficient feature learning techniques such as the convolutional neural networks can improve the learning capability as they do not rely on manually designed features and can thus provide good generalization and better segmentation accuracy. The huge prior information on the global and local context can be used effectively by training an appropriate deep neural network model that can extract important features and combine them in a heirarchical manner for classification. Therefore, as detailed in chapter 5, we assessed the performance of a fully automated method based on deep convolutional neural network for segmentation of polycystic kidneys using CT dataset from patients at early stage and late stages of the disease. The proposed method was trained and tested on a wide range of TKV achieving a good mean Dice Similarity Coefficient between automated and manual segmentations from clinical experts and excellent mean correlation coefficient for segmented kidney volume measurements in the entire test set. Our method facilitates fast and reproducible measurements of kidney volumes in agreement with manual segmentations from clinical experts. A limitation of the proposed method based on deep learning is the inclusion of liver cysts in the segmented kidney region for some cases with highly cystic liver. To potentially overcome this problem, the proposed method may be trained on 3D volumes of polycystic kidneys thereby providing additional information about the kidney shape not captured currently in the 2D slices.

In this disseration, the proposed machine learning methods have been investigated only on CT dataset and as a future work, the segmentation strategy could be extended to other imaging modalities such as MRI or to other affected organs in ADPKD such as the polycystic liver. Similar to several other tasks in medical imaging, the machine learning based approaches used for our experiments required to be defined in a supervised way, therefore keeping human expert annotations neccessary. Training deep learning models using sparse annotations of polycystic kidneys in 3D with minimal user interaction on few equally spaced slices (or slices with greatest change with respect to the previous or next slice) could provide interesting segmentation results but this hypothesis needs to be validated in the future. Finally, for improving current procedures of TKV measurement in ADPKD clinical trials, a good strategy could involve human-machine interactive frameworks where the initial segmentation is performed using an automated segmentation method such as employing a trained deep learning model and then integrating the segmentation outcomes into a user-friendly interface that allows human-expert

interaction for fine-tuning the final segmentation to achieve desired results. This strategy may allow reduction in human effort and time requirement for performing TKV measurements while maintaining the desired level of accuracy required in ADPKD clinical trials.

We demonstrate that machine learning can be successfully used for complex medical image segmentation tasks. Future research on machine learning and its applications in the medical domain might not only lead to improved algorithms for classical computer vision problems such as image segmentation, but can also facilitate holistic physical and biological models integrating heterogeneous clinical data from various sources that foster a thorough understanding of disease development, progression and treatment possibilities.

Part IV

Appendix

A

Supplementary Information for Chapter 3: Kidney Volume Measurement in ADPKD

Estimation of ellipsoid volume by planimetry

The estimation of the volume of an ellipsoid, the idealized volume of a kidney, by planimetry is based on the calculation of the area of equispaced sections perpendicular to one axis. As shown in figure A.1, the volume of an ellipsoid can be estimated by three semi-axes (*a, b, c*) by the equation:

$$V = \frac{4}{3}\pi abc. \tag{A.1}$$

Using planimetry, the volume (V_p) can be estimated by the equation:

$$V_p = \sum_{1}^{N} A_i d, \tag{A.2}$$

where, A_i is the area of section i and N is the number of equally spaced sections of thickness d. The estimation of the ellipsoid volume can depend on section thickness and on section orientation. Since estimation of kidney volume by planimetry is related to slice thickness, as well as to orientation of kidney sections, we estimated the planimetry error based on ellipsoids of volumes comparable to ADPKD kidney volumes, using slice thickness and orientation corresponding to MR and CT images. In detail, we divided each of the two groups of ADPKD patients studied with MR and CT, respectively, in 3 subgroups based on kidney volumes estimated by polyline manual tracing method (as reported in table A.1). We then computed the mean major axis (length) of the three kidney volume classes, and the maximum area

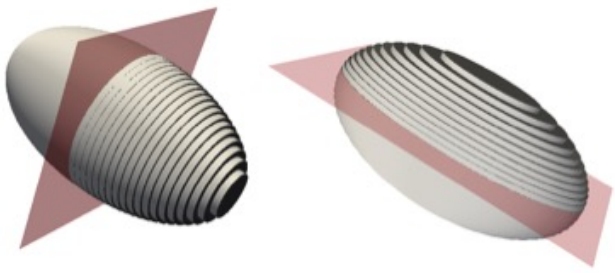


Fig. A.1. Ellipsoid volume assessment using planimetry. The ellipsoid is divided in slices (axial and sagittal slices for CT and MR, respectively), and the volume is computed as sum of the slice areas multiplied by the slice thickness.

		SKV *(ml)*	Max Area (mm^2)	Length *(mm)*
MRI	Small (n=10)	370	4,049	135.6
	Medium (n=10)	973	7,364	192.4
	Large (n=10)	2277	13,541	247.2
CT	Small (n=10)	503	5,192	150.5
	Medium (n=10)	1297	9,973	204.0
	Large (n=10)	2354	14,752	252.5

Tab. A.1. Single kidney volume (SKV), maximum area and length of average kidneys of different size.

perpendicular to the major axis (table A.1). We then considered 6 ellipsoids representative of the three average kidney volumes derived from MR and CT images, assuming length and semi-axes reported in table A.2. Ellipsoid volumes computed using analytical equation are reported in table A.2.

To simulate the effect of planimetry tracing on the estimation of ellipsoid volume, we computed the area of ellipsoid sections, with transversal or longitudinal orientation (representative of MR and CT image sequences) and section thickness of 4 and 5 mm for MRI and CT, respectively (as shown in figure A.1).

To calculate the radius of the circumferences of the hypothetical ellipsoid sections, for each slice we computed the y coordinate of the ellipse equation for a given *x* coordinate as:

$$y_T = \sqrt{(1 - \frac{x^2}{a^2})b^2} \qquad \text{(A.3)} \qquad\qquad y_L = \sqrt{(1 - \frac{x^2}{b^2})a^2} \qquad \text{(A.4)}$$

for *transverse* and *longitudinal* sections, respectively and calculated the slice area $A_i = \pi y_i^2$. Thereafter, we computed the ellipsoid volume as sum of the areas of all ellipsoid sections multiplied by the slice thickness. Ellipsoid volumes computed by planimetry are reported in table A.3. Finally we calculated the error between analytical ellipsoid volume and the volume

	Max Area (mm^2)	a *(mm)*	b, c *(mm)*	Volume *(ml)*
MRI	4,071	67.5	36	366
	7,854	95	50	994
	13,685	125	66	2,280
CT	5,024	75	40	503
	9,498	100	55	1,266
	14,306	125	67.5	2,384

Tab. A.2. Geometrical parameters of ellipsoids assumed to be representative of ADPKD kidneys of different size.

		Volume by planimetry	Volume difference	Error (%)
		(ml)	*analytical vs planimetry (ml)*	*min/max*
MRI	Small	367.0	-0.75	-.20 / +0.26
	Medium	993.2	1.08	-0.13 / +0.11
	Large	2,278.7	0.94	-0.10 / +0.04
CT	Small	502.1	0.30	-0.10 / +0.06
	Medium	1,266.3	0.15	-0.08 / +0.01
	Large	2,384.7	-0.25	-0.07 /-0.01

Tab. A.3. **Example ellipsoid volumes computed by planimetry, and percentage errors with respect to analytical volumes.** Since errors slightly change with the slicing offset, minimum and maximum errors are reported.

estimated by planimetry (see table A.3). We repeated the calculation using different slicing offsets, in order to quantify the error due to random slice positioning.

The difference between analytical and planimetry volume is very small, with a percentage error less than 0.3% (table A.3), suggesting that the estimation of the volume of ellipsoids representative of ADPKD kidneys of different sizes can be reliably obtained by planimetry, both for orientation and section thickness of MR and CT image sequences.

List of Authored and Co-authored Publications

B

2017

[117] **Kanishka Sharma**, Christian Rupprecht, Anna Caroli, Maria Carolina Aparicio, Andrea Remuzzi, Maximilian Baust, and Nassir Navab. "Automatic Segmentation of Kidneys using Deep Learning for Total Kidney Volume Quantification in Autosomal Dominant Polycystic Kidney Disease". *Scientific Reports 7, doi: 10.1038/s41598-017-01779-0*

[118] **Kanishka Sharma**, Anna Caroli, Le Van Quach, Katja Petzold, Michela Bozzetto, Andreas L. Serra, Giuseppe Remuzzi, Andrea Remuzzi. "Kidney volume measurement methods for clinical studies on autosomal dominant polycystic kidney disease". *PLoS ONE, doi: 10.1371/journal.pone.0178488*

2016

[110] Piero Ruggenenti, Giorgio Gentile, Norberto Perico, Annalisa Perna, Luca Barcella, Matias Trillini, Monica Cortinovis, Claudia Patricia Ferrer Siles, Jorge Arturo Reyes Loaeza, Maria Carolina Aparicio, Giorgio Fasolini, Flavio Gaspari, Davide Martinetti, Fabiola Carrara, Nadia Rubis, Silvia Prandini, Anna Caroli, **Kanishka Sharma**, Luca Antiga, Andrea Remuzzi, and Giuseppe Remuzzi, on behalf of the SIRENA 2 Study Group. "Effect of Sirolimus on Disease Progression in Patients with Autosomal Dominant Polycystic Kidney Disease and CKD Stages 3b-4". *Clinical Journal of the American Society of Nephrology, doi: 10.2215/CJN.09900915*

2015

[119] **Kanishka Sharma**, Loïc Peter, Christian Rupprecht, Anna Caroli, Lichao Wang, Andrea Remuzzi, Maximilian Baust, Nassir Navab. "Semi-Automatic Segmentation of Autosomal Dominant Polycystic Kidneys using Random Forests". *arXiv preprint, arXiv:1510.06915*

2013

[84] Philipp Matthies, **Kanishka Sharma**, Aslı Okur, José Gardiazabal, Jakob Vogel, Tobias Lasser, Nassir Navab. "First use of mini gamma cameras for intra-operative robotic SPECT reconstruction". *In: International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 163-170) - Springer Berlin Heidelberg, doi: 10.1007/978-3-642-40811-3_21*

[148] Wolfgang Wein, Alexander Ladikos, Bernhard Fuerst, Amit Shah, **Kanishka Sharma**, Nassir Navab. "Global Registration of Ultrasound to MRI Using the LC2 Metric for Enabling Neurosurgical Guidance". *In: International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 34-41) - Springer Berlin Heidelberg, doi: 10.1007/978-3-642-40811-3_5*

C

Abstract of Contributions not Discussed in this Thesis

Intravoxel Incoherent Motion Magnetic Resonance Imaging in ADPKD

Purpose: Autosomal dominant polycystic kidney disease (ADPKD) is characterized by the development of fluid-filled cysts leading to progressive kidney volume enlargement. Non-cystic tissue is denoted by regions of interstitial fibrosis, while residual parenchyma is limited. The aim of this study was to investigate intravoxel incoherent motion (IVIM)–based parameters in ADPKD kidneys as compared to normal kidneys, using diffusion-weighted magnetic resonance imaging (DWI).

Method: A normal control and a patient with ADPKD underwent DWI on a 1.5 T scanner using a single-shot echo planar imaging sequence with 9 diffusion weightings (b=0, 15, 50, 100, 200, 350, 500, 700, 1000). DWI images were quantified using a segmented piecewise exponential fitting to calculate IVIM parameters: F (perfusion fraction), ADC_{slow} and ADC_{fast} ("slow" and "fast" diffusion coefficients). Mean and standard deviation values derived from multiple regions of interest were computed. Figure C.1 shows the IVIM parameters computed on DWI for normal and ADPKD kidneys.

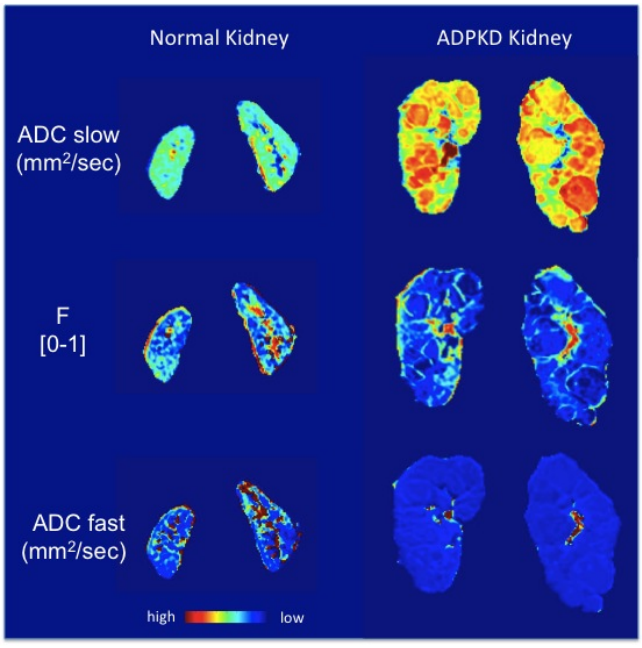


Fig. C.1. **IVIM parameters computed on Diffusion Weighted Imaging for normal and ADPKD kidneys.**

Results: In normal kidneys, an important fraction of volume is occupied by fluid in movement ($F = 0.393 \pm 0.049$ in the cortex and $F = 0.455 \pm 0.183$ in the blood vessels). Volume occupied by blood vessels and renal pelvis has the highest ADC_{fast} values ($ADC_{fast} = 0.0158 \pm 0.0070$ mm^2/sec). In ADPKD kidney, as compared to normal kidney, the fraction of fluid in movement is importantly reduced ($F = 0.101 \pm 0.053$ in the cysts) and the fluid movement in this volume fraction is lower ($ADC_{fast} = 0.0035 \pm 0.0003$ mm^2/sec). As expected, ADC_{slow} is higher in renal cysts due to water molecule free diffusion ($ADC_{slow} = 0.0028 \pm 0.0003$ mm^2/sec in the cysts and $ADC_{slow} = 0.0017 \pm 0.0001$ mm^2/sec in the normal kidney cortex).

Conclusion: Despite preliminary, our results suggest that IVIM-DWI based parameters have great potential as truly non-invasive biomarkers to obtain quantitative information about ADPKD kidneys. Our approach can be used to study ADPKD patients at different disease stage, and to investigate the structural and functional relation in non-cystic tissue. The same technique can be applied to other nephropathies to quantify renal tissue perfusion and function.

Effect of Sirolimus on Disease Progression in Patients with Autosomal Dominant Polycystic Kidney Disease and CKD Stages 3b-4

Ruggenenti, P., et al. *Effect of Sirolimus on Disease Progression in Patients with Autosomal Dominant Polycystic Kidney Disease and CKD Stages 3b-4. Clinical Journal of the American Society of Nephrology, pp.CJN-09900915* [110]

Purpose and Method: The effect of mammalian target of rapamycin (mTOR) inhibitors has never been tested in patients with autosomal dominant polycystic kidney disease (ADPKD) and severe renal insufficiency. In this academic, prospective, randomized, open label, blinded end point, parallel group trial (ClinicalTrials.gov no. NCT01223755), 41 adults with ADPKD, CKD stage 3b or 4, and proteinuria $\leq 0.5g/24h$ were randomized between September of 2010 and March of 2012 to sirolimus (3 mg/d; serum target levels of 5–10 ng/ml) added on to conventional therapy (n=21) or conventional treatment alone (n=20). Primary outcome was GFR (iohexol plasma clearance) change at 1 and 3 years versus baseline.

Results: At the 1-year preplanned interim analysis, GFR fell from 26.7 ± 5.8 to 21.3 ± 6.3 ml/min per 1.73 m^2 (P<0.001) and from 29.6 ± 5.6 to 24.9 ± 6.2 ml/min per 1.73 m^2 (P<0.001) in the sirolimus and conventional treatment groups, respectively. Albuminuria (73.8 ± 81.8 versus 154.9 ± 152.9 $\mu g/min$; P=0.02) and proteinuria (0.3 ± 0.2 versus 06 ± 0.4 g/24 h; P<0.01) increased with sirolimus. Seven patients on sirolimus versus one control had de novo proteinuria (P=0.04), ten versus three patients doubled proteinuria (P=0.02), 18 versus 11 patients had peripheral edema (P=0.04), and 14 versus six patients had upper respiratory tract infections (P=0.03). Three patients on sirolimus had angioedema, 14 patients had aphthous stomatitis, and seven patients had acne (P<0.01 for both versus controls). Two patients progressed to ESRD, and two patients withdrew because of worsening of proteinuria. These events were not observed in controls. Thus, the independent data and safety monitoring board recommend early trial termination for safety reasons. At 1 year, total kidney volume (assessed by contrast–enhanced computed tomography imaging) increased by 9.0% from 2857.7 ± 1447.3

to 3094.6 ± 1519.5 ml on sirolimus and 4.3% from 3123.4 ± 1695.3 to 3222.6 ± 1651.4 ml on conventional therapy (P=0.12). On follow-up, 37% and 7% of serum sirolimus levels fell below or exceeded the therapeutic range, respectively.

Conclusion: Finding that sirolimus was unsafe and ineffective in patients with ADPKD and renal insufficiency suggests that mTOR inhibitor therapy may be contraindicated in this context.

First Use of Mini Gamma Cameras for Intra-operative Robotic SPECT Reconstruction

Matthies, P., Sharma, K., Okur, A., Gardiazabal, J., Vogel, J., Lasser, T. and Navab, N. "*First use of mini gamma cameras for intra-operative robotic SPECT reconstruction*". *In: International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 163-170). Springer Berlin Heidelberg* [84]

Purpose and Method: Different types of nuclear imaging systems have been used in the past, starting with pre-operative gantry-based SPECT systems and gamma cameras for 2D imaging of radioactive distributions. The main applications are concentrated on diagnostic imaging, since traditional SPECT systems and gamma cameras are bulky and heavy. With the development of compact gamma cameras with good resolution and high sensitivity, it is now possible to use them without a fixed imaging gantry. Mounting the camera onto a robot arm solves the weight issue, while also providing a highly repeatable and reliable acquisition platform. In this work we introduce a novel robotic setup performing scans with a mini gamma camera, along with the required calibration steps, and show the first SPECT reconstructions.

Results and Conclusion: The results are extremely promising, both in terms of image quality as well as reproducibility. In our experiments, the novel setup outperformed a commercial fhSPECT system, reaching accuracies comparable to state-of-the-art SPECT systems.

Global Registration of Ultrasound to MRI Using the LC2 Metric for Enabling Neurosurgical Guidance

Wein, W., Ladikos, A., Fuerst, B., Shah, A., Sharma, K. and Navab, N. *"Global Registration of Ultrasound to MRI Using the LC2 Metric for Enabling Neurosurgical Guidance". In: International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 34-41) - Springer Berlin Heidelberg* [148]

Purpose and Method: Automatic and robust registration of pre-operative magnetic resonance imaging (MRI) and intra-operative ultrasound (US) is essential to neurosurgery. We reformulate and extend an approach which uses a Linear Correlation of Linear Combination (LC^2)-based similarity metric, yielding a novel algorithm which allows for fully automatic US-MRI registration in the matter of seconds. In addition, we systematically study the accuracy, precision, and capture range of the algorithm, as well as its sensitivity to different choices of parameters.

Results and Conclusion: The algorithm is evaluated on 14 clinical neurosurgical cases with tumors, with an average landmark-based error of 2.52mm for the rigid transformation. It is invariant with respect to the unknown and locally varying relationship between US image intensities and both MRI intensity and its gradient. The overall method based on this both recovers global rigid alignment, as well as the parameters of a free-form-deformation (FFD) model.

Bibliography

[1] M. D. Abràmoff, P. J. Magalhães, and S. J. Ram. “Image processing with ImageJ”. In: *Biophotonics international* 11.7 (2004), pp. 36–42 (cit. on p. 70).

[2] R. Adams and L. Bischof. “Seeded region growing”. In: *IEEE Transactions on pattern analysis and machine intelligence* 16.6 (1994), pp. 641–647 (cit. on p. 14).

[3] A. Alam, N. K. Dahl, J. H. Lipschutz, et al. “Total kidney volume in autosomal dominant polycystic kidney disease: a biomarker of disease progression and therapeutic efficacy”. In: *American Journal of Kidney Diseases* 66.4 (2015), pp. 564–576 (cit. on p. 21).

[4] S. Almeida, E. Almeida, D. Peters, et al. “Autosomal dominant polycystic kidney disease: evidence for the existence of a third locus in a Portuguese family”. In: *Human genetics* 96.1 (1995), pp. 83–88 (cit. on pp. 5, 6).

[5] L. Antiga, M. Piccinelli, G. Fasolini, et al. “Computed tomography evaluation of autosomal dominant polycystic kidney disease progression: a progress report”. In: *Clinical Journal of the American Society of Nephrology* 1.4 (2006), pp. 754–760 (cit. on pp. 8, 14).

[6] K. T. Bae and J. J. Grantham. “Imaging for the prognosis of autosomal dominant polycystic kidney disease”. In: *Nature Reviews Nephrology* 6.2 (2010), pp. 96–106 (cit. on p. 8).

[7] K. T. Bae, C. Tao, J. Wang, et al. “Novel approach to estimate kidney and cyst volumes using mid-slice magnetic resonance images in polycystic kidney disease”. In: *American journal of nephrology* 38.4 (2013), pp. 333–341 (cit. on pp. 22, 26, 78).

[8] K. T. Bae, P. K. Commean, and J. Lee. “Volumetric measurement of renal cysts and parenchyma using MRI: phantoms and patients with polycystic kidney disease”. In: *Journal of computer assisted tomography* 24.4 (2000), pp. 614–619 (cit. on pp. 22, 26, 78).

[9] W. A. Barrett and E. N. Mortensen. “Interactive live-wire boundary extraction”. In: *Medical image analysis* 1.4 (1997), pp. 331–341 (cit. on p. 26).

[10] J. C. Bear, P. S. Parfrey, J. M. Morgan, C. J. Martin, and B. C. Cramer. “Autosomal dominant polycystic kidney disease: new information for genetic counselling”. In: *American journal of medical genetics* 43.3 (1992), pp. 548–553 (cit. on p. 6).

[11] Y. Bengio et al. “Learning deep architectures for AI”. In: *Foundations and trends® in Machine Learning* 2.1 (2009), pp. 1–127 (cit. on p. 59).

[12] M. A. Bernstein, J. Huston, and H. A. Ward. “Imaging artifacts at 3.0 T”. In: *Journal of Magnetic Resonance Imaging* 24.4 (2006), pp. 735–746 (cit. on p. 13).

[13] H. Bhutani, V. Smith, F. Rahbari-Oskoui, et al. “A comparison of ultrasound and magnetic resonance imaging shows that kidney length predicts chronic kidney disease in autosomal dominant polycystic kidney disease”. In: *Kidney international* 88.1 (2015), pp. 146–151 (cit. on pp. 22, 36).

[14] F. E. Boas and D. Fleischmann. “CT artifacts: causes and reduction techniques”. In: *Imaging in Medicine* 4.2 (2012), pp. 229–240 (cit. on p. 13).

[15] N Bogdanova, B Dworniczak, D Dragova, et al. “Genetic heterogeneity of polycystic kidney disease in Bulgaria”. In: *Human genetics* 95.6 (1995), pp. 645–650 (cit. on p. 5).

[16] C. Boucher and R. Sandford. “Autosomal dominant polycystic kidney disease (ADPKD, MIM 173900, PKD1 and PKD2 genes, protein products known as polycystin-1 and polycystin-2)”. In: *European journal of human genetics* 12.5 (2004), pp. 347–354 (cit. on p. 7).

[17] W. E. Braun, J. D. Schold, B. R. Stephany, R. A. Spirko, and B. R. Herts. “Low-dose rapamycin (sirolimus) effects in autosomal dominant polycystic kidney disease: an open-label randomized controlled pilot study”. In: *Clinical Journal of the American Society of Nephrology* (2014), CJN–02650313 (cit. on p. 21).

[18] L. Breiman. “Bagging predictors”. In: *Machine learning* 24.2 (1996), pp. 123–140 (cit. on pp. 39, 42).

[19] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984 (cit. on p. 39).

[20] J. Canny. “A computational approach to edge detection”. In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), pp. 679–698 (cit. on p. 15).

[21] A. Caroli, N. Perico, A. Perna, et al. “Effect of longacting somatostatin analogue on kidney and cyst growth in autosomal dominant polycystic kidney disease (ALADIN): a randomised, placebo-controlled, multicentre trial”. In: *The Lancet* 382.9903 (2013), pp. 1485–1495 (cit. on pp. 21, 22, 27, 30).

[22] M.-Y. Chang and A. Ong. “New treatments for autosomal dominant polycystic kidney disease”. In: *British journal of clinical pharmacology* 76.4 (2013), pp. 524–535 (cit. on p. 21).

[23] A. B. Chapman and W. Wei. “Imaging approaches to patients with polycystic kidney disease”. In: *Seminars in nephrology*. Vol. 31. 3. Elsevier. 2011, pp. 237–244 (cit. on p. 21).

[24] A. B. Chapman, D. Rubinstein, R. Hughes, et al. “Intracranial aneurysms in autosomal dominant polycystic kidney disease”. In: *New England Journal of Medicine* 327.13 (1992), pp. 916–920 (cit. on p. 7).

[25] A. B. Chapman, J. E. Bost, V. E. Torres, et al. “Kidney volume and functional outcomes in autosomal dominant polycystic kidney disease”. In: *Clinical Journal of the American Society of Nephrology* 7.3 (2012), pp. 479–486 (cit. on pp. 3, 21).

[26] A. B. Chapman, L. M. Guay-Woodford, J. J. Grantham, et al. “Renal structure in early autosomal-dominant polycystic kidney disease (ADPKD): The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) cohort”. In: *Kidney international* 64.3 (2003), pp. 1035–1045 (cit. on p. 8).

[27] A. B. Chapman, V. E. Torres, R. D. Perrone, et al. “The HALT polycystic kidney disease trials: design and implementation”. In: *Clinical Journal of the American Society of Nephrology* 5.1 (2010), pp. 102–109 (cit. on p. 8).

[28] D. Chen, Y. Ma, X. Wang, et al. “Clinical characteristics and disease predictors of a large Chinese cohort of patients with autosomal dominant polycystic kidney disease”. In: *PloS one* 9.3 (2014), e92232 (cit. on p. 22).

[29] B. Cheong, R. Muthupillai, M. F. Rubin, and S. D. Flamm. “Normal values for renal length and volume as measured by magnetic resonance imaging”. In: *Clinical journal of the American Society of Nephrology* 2.1 (2007), pp. 38–45 (cit. on p. 4).

[30] D. Ciregan, U. Meier, and J. Schmidhuber. “Multi-column deep neural networks for image classification”. In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE. 2012, pp. 3642–3649 (cit. on p. 18).

[31] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. “Active shape models-their training and application”. In: *Computer vision and image understanding* 61.1 (1995), pp. 38–59 (cit. on p. 13).

[32] A Criminisi, J Shotton, and E Konukoglu. “Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning”. In: *Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114* 5.6 (2011), p. 12 (cit. on pp. 42–44).

[33] A. Criminisi and J. Shotton. *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media, 2013 (cit. on p. 51).

[34] A. Criminisi, J. Shotton, and S. Bucciarelli. “Decision forests with long-range spatial context for organ localization in CT volumes”. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Citeseer. 2009, pp. 69–80 (cit. on pp. 18, 45).

[35] A. Criminisi, J. Shotton, D. Robertson, and E. Konukoglu. “Regression forests for efficient anatomy detection and localization in CT studies”. In: *International MICCAI Workshop on Medical Computer Vision*. Springer. 2010, pp. 106–117 (cit. on p. 44).

[36] M. C. Daoust, D. M. Reynolds, D. G. Bichet, and S. Somlo. “Evidence for a third genetic locus for autosomal dominant polycystic kidney disease”. In: *Genomics* 25.3 (1995), pp. 733–736 (cit. on pp. 5, 6).

[37] L. R. Dice. “Measures of the amount of ecologic association between species”. In: *Ecology* 26.3 (1945), pp. 297–302 (cit. on p. 74).

[38] E. W. Dijkstra. “A note on two problems in connexion with graphs”. In: *Numerische mathematik* 1.1 (1959), pp. 269–271 (cit. on p. 15).

[39] *European Medicines Agency (EMA) Qualification opinion: Total Kidney Volume (TKV) as a prognostic biomarker for use in clinical trials evaluating patients with Autosomal Dominant Polycystic Kidney Disease (ADPKD)*. www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2015/11/WC500196569.pdf. (2015) (cit. on p. 21).

[40] G. M. Fick-Brosnahan, M. M. Belz, K. K. McFann, A. M. Johnson, and R. W. Schrier. “Relationship between renal volume growth and renal function in autosomal dominant polycystic kidney disease: a longitudinal study”. In: *American journal of kidney diseases* 39.6 (2002), pp. 1127–1134 (cit. on p. 21).

[41] X. Glorot and Y. Bengio. “Understanding the difficulty of training deep feedforward neural networks.” In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*. Vol. 9. Society for Artificial Intelligence and Statistics, 2010, pp. 249–256 (cit. on pp. 66, 73).

[42] I. Goodfellow, H. Lee, Q. V. Le, A. Saxe, and A. Y. Ng. “Measuring invariances in deep networks”. In: *Advances in neural information processing systems*. 2009, pp. 646–654 (cit. on p. 58).

[43] B. Graham. “Fractional max-pooling”. In: *arXiv preprint arXiv:1412.6071* (2014) (cit. on p. 66).

[44] J. J. Grantham. “Autosomal dominant polycystic kidney disease”. In: *New England Journal of Medicine* 359.14 (2008), pp. 1477–1485 (cit. on p. 3).

[45] J. J. Grantham and V. E. Torres. “The importance of total kidney volume in evaluating progression of polycystic kidney disease”. In: *Nature Reviews Nephrology* (2016) (cit. on pp. 3, 6, 21, 22, 78).

[46] J. J. Grantham, J. L. Geiser, and A. P. Evan. “Cyst formation and growth in autosomal dominant polycystic kidney disease”. In: *Kidney international* 31.5 (1987), pp. 1145–1152 (cit. on p. 6).

[47] J. J. Grantham, A. B. Chapman, and V. E. Torres. “Volume progression in autosomal dominant polycystic kidney disease: the major factor determining clinical outcomes”. In: *Clinical Journal of the American Society of Nephrology* 1.1 (2006), pp. 148–157 (cit. on p. 8).

[48] J. J. Grantham, V. E. Torres, A. B. Chapman, et al. “Volume progression in polycystic kidney disease”. In: *New England Journal of Medicine* 354.20 (2006), pp. 2122–2130 (cit. on pp. 8, 21, 22, 78).

[49] A. Graves, S. Fernández, and J. Schmidhuber. “Multi-dimensional Recurrent Neural Networks”. In: *Artificial Neural Networks – ICANN 2007: 17th International Conference, Porto, Portugal, September 9-13, 2007, Proceedings, Part I*. Ed. by J. M. de Sá, L. A. Alexandre, W. Duch, and D. Mandic. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 549–558 (cit. on p. 62).

[50] N. Hateboer, M. A. v Dijk, N. Bogdanova, et al. “Comparison of phenotypes of polycystic kidney disease types 1 and 2”. In: *The Lancet* 353.9147 (1999), pp. 103–107 (cit. on p. 6).

[51] J. P. Hayslett, M. Kashgarian, and F. H. Epstein. “Functional correlates of compensatory renal hypertrophy”. In: *Journal of Clinical Investigation* 47.4 (1968), p. 774 (cit. on p. 21).

[52] E. Higashihara, K. Nutahara, T. Okegawa, et al. “Kidney Volume Estimations with Ellipsoid Equations by Magnetic Resonance Imaging in Autosomal Dominant Polycystic Kidney Disease”. In: *Nephron* 129.4 (2015), pp. 253–262 (cit. on pp. 22, 78).

[53] E. Higashihara, K. Nutahara, T. Okegawa, et al. “Safety study of somatostatin analogue octreotide for autosomal dominant polycystic kidney disease in Japan”. In: *Clinical and experimental nephrology* 19.4 (2015), pp. 746–752 (cit. on p. 21).

[54] G. E. Hinton. “Deep belief networks”. In: *Scholarpedia* 4.5 (2009). revision #91189, p. 5947 (cit. on p. 59).

[55] G. E. Hinton, S. Osindero, and Y.-W. Teh. “A fast learning algorithm for deep belief nets”. In: *Neural computation* 18.7 (2006), pp. 1527–1554 (cit. on pp. 59, 62).

[56] T. K. Ho. “Random decision forests”. In: *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*. Vol. 1. IEEE. 1995, pp. 278–282 (cit. on p. 39).

[57] T. K. Ho. “The random subspace method for constructing decision forests”. In: *IEEE transactions on pattern analysis and machine intelligence* 20.8 (1998), pp. 832–844 (cit. on pp. 39, 43).

[58] M. C. Hogan, T. V. Masyuk, L. J. Page, et al. “Randomized clinical trial of long-acting somatostatin for autosomal dominant polycystic kidney and liver disease”. In: *Journal of the American Society of Nephrology* 21.6 (2010), pp. 1052–1061 (cit. on p. 21).

[59] T Huang, G. Yang, and G Tang. “A fast two-dimensional median filtering algorithm”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 27.1 (1979), pp. 13–18 (cit. on p. 13).

[60] D. H. Hubel and T. N. Wiesel. “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”. In: *The Journal of physiology* 160.1 (1962), pp. 106–154 (cit. on p. 18).

[61] E. B. Hunt, J. Marin, and P. J. Stone. “Experiments in induction.” In: (1966) (cit. on p. 40).

[62] S. Ioffe and C. Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *arXiv preprint arXiv:1502.03167* (2015) (cit. on pp. 65, 71).

[63] M. V. Irazabal, L. J. Rangel, E. J. Bergstralh, et al. “Imaging Classification of Autosomal Dominant Polycystic Kidney Disease: A Simple Model for Selecting Patients for Clinical Trials”. In: *Journal of the American Society of Nephrology* 26.1 (2014), pp. 160–172 (cit. on pp. 22, 26, 78).

[64] Y. Jia, E. Shelhamer, J. Donahue, et al. “Caffe: Convolutional architecture for fast feature embedding”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 675–678 (cit. on p. 72).

[65] M. Johnson, J. Shotton, and R. Cipolla. “Semantic Texton Forests for Image Categorization and Segmentation”. In: *Decision Forests for Computer Vision and Medical Image Analysis*. Springer, 2013, pp. 211–227 (cit. on pp. 18, 45).

[66] R. G. Jones and R. B. Payne. *Clinical investigation and statistics in laboratory medicine*. American Association for Clinical Chemistry, 1997 (cit. on pp. 28, 74).

[67] M. Kass, A. Witkin, and D. Terzopoulos. “Snakes: Active contour models”. In: *International journal of computer vision* 1.4 (1988), pp. 321–331 (cit. on p. 15).

[68] Y. Kim, Y. Ge, C. Tao, et al. “Automated Segmentation of Kidneys from MR Images in Patients with Autosomal Dominant Polycystic Kidney Disease”. In: *Clinical Journal of the American Society of Nephrology* 11.4 (2016), pp. 576–584 (cit. on pp. 15, 79).

[69] B. F. King, V. E. Torres, M. E. Brummer, et al. “Magnetic resonance measurements of renal blood flow as a marker of disease severity in autosomal-dominant polycystic kidney disease”. In: *Kidney international* 64.6 (2003), pp. 2214–2221 (cit. on p. 21).

[70] B. F. King, J. E. Reed, E. J. Bergstralh, P. F. Sheedy, and V. E. Torres. “Quantification and longitudinal trends of kidney, renal cyst, and renal parenchyma volumes in autosomal dominant polycystic kidney disease”. In: *Journal of the American Society of Nephrology* 11.8 (2000), pp. 1505–1511 (cit. on p. 8).

[71] A. D. Kistler, D. Poster, F. Krauer, et al. “Increases in kidney volume in autosomal dominant polycystic kidney disease can be detected within 6 months”. In: *Kidney international* 75.2 (2009), pp. 235–241 (cit. on p. 8).

[72] T. L. Kline, P. Korfiatis, M. E. Edwards, et al. “Automatic total kidney volume measurement on follow-up magnetic resonance images to facilitate monitoring of autosomal dominant polycystic kidney disease progression”. In: *Nephrology Dialysis Transplantation* (2015), gfv314 (cit. on p. 78).

[73] T. L. Kline, M. E. Edwards, P. Korfiatis, Z. Akkus, V. E. Torres, and B. J. Erickson. “Semiautomated Segmentation of Polycystic Kidneys in T2-Weighted MR Images”. In: *American Journal of Roentgenology* 207.3 (2016), pp. 605–613 (cit. on p. 78).

[74] P. Kontschieder, P. Kohli, J. Shotton, and A. Criminisi. “GeoF: Geodesic forests for learning coupled predictors”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 65–72 (cit. on p. 45).

[75] K Krammer and Y Singer. “On the algorithmic implementation of multi-class SVMs”. In: *Proc. of JMLR* (2001), pp. 265–292 (cit. on p. 45).

[76] A. Krizhevsky, I. Sutskever, and G. E. Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 1097–1105 (cit. on pp. 18, 58, 66, 71).

[77] H. Laurent and R. L. RIVEST. “CONSTRUCTING OPTIMAL BINARY DECISION TREES IS NP-COMPLETE?=”. In: *Information Processing Letters* (1976) (cit. on p. 39).

[78] C. Lee, S. Huh, T. A. Ketter, and M. Unser. “Unsupervised connectivity-based thresholding segmentation of midsagittal brain MR images”. In: *Computers in biology and medicine* 28.3 (1998), pp. 309–338 (cit. on p. 14).

[79] K. Li and B. Fei. “A new 3D model-based minimal path segmentation method for kidney MR images”. In: *Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on*. IEEE. 2008, pp. 2342–2344 (cit. on p. 15).

[80] M. Lin, Q. Chen, and S. Yan. “Network in network”. In: *arXiv preprint arXiv:1312.4400* (2013) (cit. on p. 66).

[81] E. N. Marieb and K. Hoehn. *Human anatomy & physiology*. Pearson Education, 2007 (cit. on p. 4).

[82] A. Martelli. “An application of heuristic search methods to edge and contour detection”. In: *Communications of the ACM* 19.2 (1976), pp. 73–83 (cit. on p. 15).

[83] MATLAB. *version 8.6.0 (R2015b)*. Natick, Massachusetts: The MathWorks Inc., 2015 (cit. on p. 71).

[84] P. Matthies, K. Sharma, A. Okur, et al. “First use of mini gamma cameras for intra-operative robotic SPECT reconstruction”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2013, pp. 163–170 (cit. on pp. 94, 97).

[85] R. McConnell, D. Rubinsztein, T. Fannin, et al. “Autosomal dominant polycystic kidney disease unlinked to thePKD1 and PKD2loci presenting as familial cerebral aneurysm”. In: *Journal of medical genetics* 38.4 (2001), pp. 238–240 (cit. on p. 5).

[86] R. Mignani, C. Corsi, M. De Marco, et al. “Assessment of kidney volume in polycystic kidney disease using magnetic resonance imaging without contrast medium”. In: *American journal of nephrology* 33.2 (2011), pp. 176–184 (cit. on pp. 15, 78).

[87] V. Mnih, K. Kavukcuoglu, D. Silver, et al. “Human-level control through deep reinforcement learning”. In: *Nature* 518.7540 (2015), pp. 529–533 (cit. on p. 62).

[88] D. Mumford and J. Shah. “Optimal approximations by piecewise smooth functions and associated variational problems”. In: *Communications on pure and applied mathematics* 42.5 (1989), pp. 577–685 (cit. on p. 13).

[89] H. Noh, S. Hong, and B. Han. “Learning deconvolution network for semantic segmentation”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 1520–1528 (cit. on p. 72).

[90] N. Otsu. “A threshold selection method from gray-level histograms”. In: *Automatica* 11.285-296 (1975), pp. 23–27 (cit. on p. 14).

[91] A. D. Paterson, K. R. Wang, D. Lupea, P. S. George-Hyslop, and Y. Pei. “Recurrent fetal loss associated with bilineal inheritance of type 1 autosomal dominant polycystic kidney disease”. In: *American journal of kidney diseases* 40.1 (2002), pp. 16–20 (cit. on p. 5).

[92] Y. Pei. “Diagnostic approach in autosomal dominant polycystic kidney disease”. In: *Clinical Journal of the American Society of Nephrology* 1.5 (2006), pp. 1108–1114 (cit. on p. 8).

[93] Y. Pei, A. D. Paterson, K. R. Wang, et al. “Bilineal disease and trans-heterozygotes in autosomal dominant polycystic kidney disease”. In: *The American Journal of Human Genetics* 68.2 (2001), pp. 355–363 (cit. on p. 6).

[94] N. Perico, L. Antiga, A. Caroli, et al. “Sirolimus therapy to halt the progression of ADPKD”. In: *Journal of the American Society of Nephrology* 21.6 (2010), pp. 1031–1040 (cit. on pp. 21, 22, 69).

[95] L. Peter, O. Pauly, P. Chatelain, D. Mateus, and N. Navab. “Scale-adaptive forest training via an efficient feature sampling scheme”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2015, pp. 637–644 (cit. on pp. 18, 45).

[96] K. Petzold, R. T. Gansevoort, A. C. Ong, et al. “Building a network of ADPKD reference centres across Europe: the EuroCYST initiative”. In: *Nephrology Dialysis Transplantation* 29.suppl 4 (2014), pp. iv26–iv32 (cit. on p. 23).

[97] D. L. Pham and J. L. Prince. “Adaptive fuzzy segmentation of magnetic resonance images”. In: *IEEE transactions on medical imaging* 18.9 (1999), pp. 737–752 (cit. on p. 13).

[98] D. L. Pham, C. Xu, and J. L. Prince. “Current methods in medical image segmentation 1”. In: *Annual review of biomedical engineering* 2.1 (2000), pp. 315–337 (cit. on pp. 11, 13).

[99] R. Pohle and K. D. Toennies. “Segmentation of medical images using adaptive region growing”. In: *Medical Imaging 2001*. International Society for Optics and Photonics. 2001, pp. 1337–1346 (cit. on p. 15).

[100] L. V. Quach. “Active shape modeling for the 3D segmentation of left kidneys in patients with polycystic disease (ADPKD) from CT and MR images”. In: (2016) (cit. on p. 16).

[101] J. R. Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014 (cit. on p. 39).

[102] J. R. Quinlan. “Induction of decision trees”. In: *Machine learning* 1.1 (1986), pp. 81–106 (cit. on p. 40).

[103] D. Racimora, P.-H. Vivier, H. Chandarana, and H. Rusinek. “Segmentation of polycystic kidneys from MR images”. In: *Medical Imaging 2010: Computer-Aided Diagnosis*. Ed. by N. Karssemeijer and R. M. Summers. SPIE-Intl Soc Optical Eng, 2010, 76241W (cit. on pp. 15, 78).

[104] V. P. Raikar and D. M. Kwartowitz. “Statistical shape modeling based renal volume measurement using tracked ultrasound”. In: *SPIE Medical Imaging*. International Society for Optics and Photonics. 2017, 101352Q–101352Q (cit. on p. 8).

[105] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko. “Semi-supervised learning with ladder networks”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 3546–3554 (cit. on p. 55).

[106] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite, and P. H. Torr. “Randomized trees for human pose detection”. In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE. 2008, pp. 1–8 (cit. on pp. 18, 45).

[107] F. Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957 (cit. on p. 56).

[108] F. Rosenblatt. “The perceptron: A probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386 (cit. on p. 56).

[109] A. Rosset, L. Spadola, and O. Ratib. “OsiriX: an open-source software for navigating in multidimensional DICOM images”. In: *Journal of digital imaging* 17.3 (2004), pp. 205–216 (cit. on p. 24).

[110] P. Ruggenenti, G. Gentile, N. Perico, et al. “Effect of Sirolimus on Disease Progression in Patients with Autosomal Dominant Polycystic Kidney Disease and CKD Stages 3b-4”. In: *Clinical Journal of the American Society of Nephrology* 11.5 (2016), pp. 785–794 (cit. on pp. 8, 21, 23, 70, 93, 96).

[111] P. Ruggenenti, A. Remuzzi, P. Ondei, et al. “Safety and efficacy of long-acting somatostatin treatment in autosomal-dominant polycystic kidney disease”. In: *Kidney international* 68.1 (2005), pp. 206–216 (cit. on pp. 21, 22).

[112] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. “Learning representations by back-propagating errors”. In: *Cognitive modeling* 5.3 (1988), p. 1 (cit. on p. 59).

[113] D. Rumelhart, G. Hinton, and R. Williams. “Learning internal representation by back propagation”. In: *Parallel distributed processing: exploration in the microstructure of cognition* 1 (1986) (cit. on p. 59).

[114] P. K. Sahoo, S. Soltani, and A. K. Wong. “A survey of thresholding techniques”. In: *Computer vision, graphics, and image processing* 41.2 (1988), pp. 233–260 (cit. on p. 14).

[115] T. Salimans and D. P. Kingma. “Weight normalization: A simple reparameterization to accelerate training of deep neural networks”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 901–901 (cit. on p. 65).

[116] A. L. Serra, D. Poster, A. D. Kistler, et al. “Sirolimus and kidney growth in autosomal dominant polycystic kidney disease”. In: *New England Journal of Medicine* 363.9 (2010), pp. 820–829 (cit. on p. 21).

[117] K. Sharma, C. Rupprecht, A. Caroli, et al. “Automatic Segmentation of Kidneys using Deep Learning for Total Kidney Volume Quantification in Autosomal Dominant Polycystic Kidney Disease”. In: *Scientific Reports* 7 (2017) (cit. on p. 93).

[118] K. Sharma, A. Caroli, L. Van Quach, et al. “Kidney volume measurement methods for clinical studies on autosomal dominant polycystic kidney disease”. In: *PloS one* 12.5 (2017), e0178488 (cit. on p. 93).

[119] K. Sharma, L. Peter, C. Rupprecht, et al. “Semi-Automatic Segmentation of Autosomal Dominant Polycystic Kidneys using Random Forests”. In: *arXiv preprint arXiv:1510.06915* (2015) (cit. on pp. 78, 93).

[120] J. Shotton, M. Johnson, and R. Cipolla. “Semantic texton forests for image categorization and segmentation”. In: *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE. 2008, pp. 1–8 (cit. on pp. 18, 44, 45).

[121] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014) (cit. on pp. 18, 71).

[122] C. Sise, M. Kusaka, L. H. Wetzel, et al. “Volumetric determination of progression in autosomal dominant polycystic kidney disease by computed tomography”. In: *Kidney international* 58.6 (2000), pp. 2492–2501 (cit. on pp. 8, 21).

[123] J. G. Sled, A. P. Zijdenbos, and A. C. Evans. “A nonparametric method for automatic correction of intensity nonuniformity in MRI data”. In: *IEEE transactions on medical imaging* 17.1 (1998), pp. 87–97 (cit. on p. 13).

[124] P. Soille. “Generalized geodesy via geodesic time”. In: *Pattern Recognition Letters* 15.12 (1994), pp. 1235–1240 (cit. on p. 49).

[125] A Soliman, S Zamil, A Lotfy, and E Ismail. “Sirolimus produced S-shaped effect on adult polycystic kidneys after 2-year treatment”. In: *Transplantation proceedings*. Vol. 44. 10. Elsevier. 2012, pp. 2936–2939 (cit. on p. 21).

[126] E. M. Spithoven, M. D. Van Gastel, A. L. Messchendorp, et al. “Estimation of total kidney volume in autosomal dominant polycystic kidney disease”. In: *American Journal of Kidney Diseases* 66.5 (2015), pp. 792–801 (cit. on p. 22).

[127] G. Stallone, B. Infante, G. Grandaliano, et al. “Rapamycin for treatment of type I autosomal dominant polycystic kidney disease (RAPYD-study): a randomized, controlled study”. In: *Nephrology Dialysis Transplantation* (2012), gfs264 (cit. on p. 21).

[128] R Studio. “RStudio: integrated development environment for R”. In: *RStudio Inc, Boston, Massachusetts* (2012) (cit. on p. 74).

[129] M. Styner, C. Brechbuhler, G Szckely, and G. Gerig. “Parametric estimate of intensity inhomogeneities applied to MRI”. In: *IEEE transactions on medical imaging* 19.3 (2000), pp. 153–165 (cit. on p. 13).

[130] C. Szegedy, W. Liu, Y. Jia, et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1–9 (cit. on p. 18).

[131] H. S. Thomsen, J. K. Madsen, J. H. Thaysen, and K. Damgaard-Petersen. “Volume of polycystic kidneys during reduction of renal function”. In: *Urologic radiology* 3.2 (1981), pp. 85–89 (cit. on p. 8).

[132] W. Thong, S. Kadoury, N. Piché, and C. J. Pal. “Convolutional networks for kidney segmentation in contrast-enhanced CT scans”. In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 0 (2016), pp. 1–6 (cit. on p. 69).

[133] R. Torra, C. Nicolau, C. Badenas, et al. “Abdominal aortic aneurysms and autosomal dominant polycystic kidney disease.” In: *Journal of the American Society of Nephrology* 7.11 (1996), pp. 2483–2486 (cit. on p. 7).

[134] A. Torralba, K. P. Murphy, and W. T. Freeman. “Sharing visual features for multiclass and multiview object detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.5 (2007) (cit. on p. 45).

[135] V. E. Torres, A. B. Chapman, R. D. Perrone, et al. “Analysis of baseline parameters in the HALT polycystic kidney disease trials”. In: *Kidney international* 81.6 (2012), pp. 577–585 (cit. on p. 21).

[136] V. E. Torres, P. C. Harris, and Y. Pirson. “Autosomal dominant polycystic kidney disease”. In: *The Lancet* 369.9569 (2007), pp. 1287–1301 (cit. on p. 7).

[137] V. E. Torres, B. F. King, A. B. Chapman, et al. “Magnetic resonance measurements of renal blood flow and disease progression in autosomal dominant polycystic kidney disease”. In: *Clinical Journal of the American Society of Nephrology* 2.1 (2007), pp. 112–120 (cit. on p. 21).

[138] V. E. Torres, A. B. Chapman, O. Devuyst, et al. “Tolvaptan in patients with autosomal dominant polycystic kidney disease”. In: *New England Journal of Medicine* 367.25 (2012), pp. 2407–2418 (cit. on pp. 21, 22).

[139] A. E. Turco, M. Clementi, S. Rossetti, R. Tenconi, and P. F. Pignatti. “An Italian family with autosomal dominant polycystic kidney disease unlinked to either the PKD1 or PKD2 gene”. In: *American journal of kidney diseases* 28.5 (1996), pp. 759–761 (cit. on p. 5).

[140] D. Turco, S. Severi, R. Mignani, V. Aiello, R. Magistroni, and C. Corsi. “Reliability of Total Renal Volume Computation in Polycystic Kidney Disease From Magnetic Resonance Imaging”. In: *Academic Radiology* 22.11 (2015), pp. 1376–1384 (cit. on p. 78).

[141] *U.S. Food & Drug Administration (FDA) Guidance for Industry: Qualification of Biomarker - Total Kidney Volume in Studies for Treatment of Autosomal Dominant Polycystic Kidney Disease*. `www.fda.gov/downloads/Drugs/Guidances/UCM458483.pdf`. (2015) (cit. on p. 21).

[142] V. Vapnik and C. Cortes. *Support Vector Networks, machine learning 20, 273-297*. 1995 (cit. on p. 45).

[143] V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013 (cit. on p. 45).

[144] G. Verswijvel, R Oyen, et al. “Magnetic resonance imaging in the detection and characterization of renal diseases”. In: *Saudi Journal of Kidney Diseases and Transplantation* 15.3 (2004), p. 283 (cit. on p. 9).

[145] D. Wallace, Y.-P. Hou, Z. Huang, et al. “Tracking kidney volume in mice with polycystic kidney disease by magnetic resonance imaging”. In: *Kidney international* 73.6 (2008), pp. 778–781 (cit. on p. 9).

[146] G. Walz, K. Budde, M. Mannaa, et al. “Everolimus in patients with autosomal dominant polycystic kidney disease”. In: *New England Journal of Medicine* 363.9 (2010), pp. 830–840 (cit. on p. 21).

[147] J. D. Warner, M. V. Irazabal, G. Krishnamurthi, B. F. King, V. E. Torres, and B. J. Erickson. “Supervised segmentation of polycystic kidneys: a new application for stereology data”. In: *Journal of digital imaging* 27.4 (2014), pp. 514–519 (cit. on p. 78).

[148] W. Wein, A. Ladikos, B. Fuerst, A. Shah, K. Sharma, and N. Navab. “Global registration of ultrasound to mri using the LC2 metric for enabling neurosurgical guidance”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2013, pp. 34–41 (cit. on pp. 94, 97).

[149] W. M. Wells, W. E. L. Grimson, R. Kikinis, and F. A. Jolesz. “Adaptive segmentation of MRI data”. In: *IEEE transactions on medical imaging* 15.4 (1996), pp. 429–442 (cit. on p. 13).

[150] P. J. Werbos. “Beyond regression: new tools for prediction and analysis in the behavioral science”. In: *Ph. D. Thesis, Harvard University* (1974) (cit. on p. 59).

[151] D. Williams and G. Hinton. “Learning representations by back-propagating errors”. In: *Nature* 323.6088 (1986), pp. 533–538 (cit. on p. 59).

[152] W. Wołyniec, M. M. Jankowska, E. Król, P. Czarniak, and B. Rutkowski. “Current diagnostic evaluation of autosomal dominant polycystic kidney disease”. In: *Pol Arch Med Wewn* 118.12 (2008), pp. 767–73 (cit. on p. 8).

[153] P. Yin, A. Criminisi, J. Winn, and I. Essa. “Tree-based classifiers for bilayer video segmentation”. In: *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*. IEEE. 2007, pp. 1–8 (cit. on p. 44).

[154] C. T. Zahn. “Graph-theoretical methods for detecting and describing gestalt clusters”. In: *IEEE Transactions on computers* 100.1 (1971), pp. 68–86 (cit. on p. 15).

[155] M. D. Zeiler and R. Fergus. “Visualizing and understanding convolutional networks”. In: *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 818–833 (cit. on pp. 72, 73).

[156] Y. Zhang, J. M. Brady, and S. Smith. “Hidden Markov random field model for segmentation of brain MR image”. In: *Medical Imaging 2000*. International Society for Optics and Photonics. 2000, pp. 1126–1137 (cit. on p. 13).

[157] Y. Zheng, D. Liu, B. Georgescu, D. Xu, and D. Comaniciu. “Deep Learning Based Automatic Segmentation of Pathological Kidney in CT: Local vs. Global Image Context”. In: *Deep Learning and Convolutional Neural Networks for Medical Image Computing*. Springer, 2016 (cit. on pp. 69, 79).

List of Figures

List of Tables