Aus dem Lehrstuhl für Computerunterstützte Klinische Medizin

der Medizinischen Fakultät Mannheim

(Direktor: Prof. Dr. Ing. Frank G. Zöllner)

Deep Learning based Medical Image Analysis using Small Datasets

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)
der
Medizinischen Fakultät Mannheim
der Ruprecht-Karls-Universität zu Heidelberg

vorgelegt von

Anish Raj, M.Sc.

aus

Jabalpur, Indien

2024

Dekan: Prof. Dr. med. Sergij Goerdt
Referent: Prof. Dr. Ing. Frank G. Zöllner

Deep Learning based Medical Image Analysis using Small Datasets

The emergence of deep learning (DL) has significantly improved the capabilities of medical image analysis, providing a basis for automated and accurate interpretation of imaging data. This work centers on the challenges and solutions related to the use of deep learning for medical image analysis in scenarios constrained by small datasets. DL algorithms exhibit remarkable performance when trained on large datasets. Yet, within the medical domain, the assembly of large, annotated datasets poses a big challenge. Obtaining such datasets is significantly impeded by privacy concerns, the rarity of medical conditions, and the time-intensive nature of gathering large annotated datasets. This thesis addresses these challenges by leveraging DL techniques designed to maximize learning from limited data.

This work presents a framework that employs attention mechanisms, suitable sampling strategies, modified loss functions, and optimizers to improve the effectiveness of Convolutional Neural Networks (CNNs) on limited datasets. In addition, preprocessing techniques such as resampling and histogram matching are used to mitigate shifts in the data distribution, enhancing model generalizability across different medical imaging sources.

In the first study, a deep learning methodology was devised for precise segmentation of Total Kidney Volume (TKV) in Autosomal Dominant Polycystic Kidney Disease (ADPKD) using MRI data, addressing the challenge of accurately estimating TKV, crucial for monitoring disease progression. This methodology incorporated attention mechanisms, the cosine loss function, and Sharpness-Aware Minimization (SAM) within a U-Net architecture to improve focus on relevant features, tackle small dataset limitations, and enhance model generalizability. Validated on 100 MRI scans, it demonstrated significant accuracy improvements, achieving a Dice similarity coefficient of 0.918, and showcased the efficacy of ensemble models for further accuracy enhancement.

The next study was an extension of the first study, where a generalizable algorithm for kidney segmentation was developed. Incorporating Nyul normalization, resampling, and attention mechanisms, this CNN framework demonstrated high generalizability and accuracy across varied patient datasets. Validated on two separate cohorts, it achieved significant improvements over the baseline model, underscoring its clinical potential for precise TKV calculation.

In the next study, an approach was developed for the automated prognosis of renal function decline in ADPKD patients using MRI data. Employing a dual-model strategy that integrates a CNN for kidney volume segmentation with attention mechanisms and an MLP for disease progression prediction, this method combines image and biomarker features. Validated on 135 patients, it achieved prognostic accuracies with area under the curve (AUC) scores exceeding 0.95 for predicting various stages of chronic kidney disease (CKD) and demonstrated a high correlation in predicting eGFR decline.

In the third study, an algorithm employing a 3D residual U-Net architecture integrated with histogram matching was developed for the detection and monitoring of

Multiple Sclerosis (MS) through Voxel-Guided Morphometry (VGM) maps. This model, designed to accurately highlight MS-related brain structure changes in MRI volumes, demonstrated its adaptability and generalizability across unseen datasets. Validated on diverse patient datasets, it achieved an average improvement of 4.2% in Mean Absolute Error (MAE) over the reference method, confirming its robustness and potential as a clinical tool for precise, efficient MS lesion dynamics analysis.

In the final study, a CNN-based framework was developed to automate acute Abdominal Aortic Dissection (AD) detection in CT scans. This approach, trained on a small internal dataset, was further validated using a large external set. The model demonstrated a high AUC and sensitivity in detecting AD, confirming its reliability and potential to transform emergency radiology through AI-assisted diagnosis.

This thesis underscores the potential of deep learning in overcoming the limitations posed by small datasets in medical image analysis, paving the way for more accessible and efficient AI-driven diagnostics and therapeutic planning.

Deep Learning basierte medizinische Bildanalyse mit kleinen Datensätzen

Das Aufkommen von Deep Learning (DL) hat die Möglichkeiten der medizinischen Bildanalyse erheblich verbessert und bietet eine Grundlage für die automatische und genaue Interpretation von Bilddaten. Diese Arbeit konzentriert sich auf die Herausforderungen und Lösungen im Zusammenhang mit der Verwendung von Deep Learning für die medizinische Bildanalyse in Szenarien, die durch kleine Datensätze eingeschränkt sind. DL-Algorithmen zeigen eine bemerkenswerte Leistung, wenn sie auf großen Datensätzen trainiert werden. Im medizinischen Bereich stellt die Zusammenstellung großer, annotierter Datensätze jedoch eine große Herausforderung dar. Die Beschaffung solcher Datensätze wird durch den Schutz der Privatsphäre, die Seltenheit medizinischer Zustände und den hohen Zeitaufwand für die Erfassung großer annotierter Datensätze erheblich erschwert. In dieser Arbeit werden diese Herausforderungen durch den Einsatz von DL-Techniken angegangen, die das Lernen aus begrenzten Daten maximieren sollen.

Diese Arbeit stellt einen Rahmen vor, der Aufmerksamkeitsmechanismen, geeignete Sampling-Strategien, modifizierte Verlustfunktionen und Optimierer einsetzt, um die Effektivität von Convolutional Neural Networks (CNNs) auf begrenzten Datensätzen zu verbessern. Darüber hinaus werden Vorverarbeitungstechniken wie Resampling und Histogramm-Matching eingesetzt, um Verschiebungen in der Datenverteilung abzumildern und die Generalisierbarkeit des Modells über verschiedene medizinische Bildgebungsquellen hinweg zu verbessern.

In der ersten Studie wurde eine Deep-Learning-Methode zur präzisen Segmentierung des Gesamtnierenvolumens (TKV) bei autosomal dominanter polyzystischer Nierenerkrankung (ADPKD) unter Verwendung von MRT-Daten entwickelt, um die Herausforderung der genauen Schätzung des TKV zu bewältigen, die für die Überwachung des Krankheitsverlaufs entscheidend ist. Diese Methodik beinhaltet Aufmerksamkeitsmechanismen, die Cosinus-Verlustfunktion und Sharpness-Aware Minimization (SAM) innerhalb einer U-Net-Architektur, um den Fokus auf relevante Merkmale zu verbessern, die Einschränkungen kleiner Datensätze zu überwinden und die Verallgemeinerbarkeit des Modells zu erhöhen. Bei der Validierung anhand von 100 MRT-Scans zeigte sich eine signifikante Verbesserung der Genauigkeit, wobei ein Dice-Ähnlichkeitskoeffizient von 0,918 erreicht wurde, und es wurde die Wirksamkeit von Ensemble-Modellen zur weiteren Verbesserung der Genauigkeit demonstriert.

Die nächste Studie war eine Erweiterung der ersten Studie, in der ein verallgemeinerbarer Algorithmus für die Segmentierung von Nieren entwickelt wurde. Durch die Einbeziehung von Nyul-Normalisierung, Resampling und Aufmerksamkeitsmechanismen zeigte dieses CNN-Framework eine hohe Verallgemeinerbarkeit und Genauigkeit bei unterschiedlichen Patientendatensätzen. Bei der Validierung an zwei separaten Kohorten wurden erhebliche Verbesserungen gegenüber dem Basismodell erzielt, was sein klinisches Potenzial für eine präzise TKV-Berechnung unterstreicht.

In der nächsten Studie wurde ein Ansatz für die automatische Prognose der Nierenfunktionsverschlechterung bei ADPKD-Patienten anhand von MRT-Daten entwickelt. Mit Hilfe einer Doppelmodellstrategie, die ein CNN zur Segmentierung des

Nierenvolumens mit Aufmerksamkeitsmechanismen und ein MLP zur Vorhersage des Krankheitsverlaufs integriert, kombiniert diese Methode Bild- und Biomarkermerkmale. Die Methode wurde an 135 Patienten validiert und erreichte eine prognostische Genauigkeit mit AUC-Werten von über 0,95 für die Vorhersage verschiedener Stadien der chronischen Nierenerkrankung (CKD) und wies eine hohe Korrelation bei der Vorhersage des Rückgangs der eGFR auf.

In der dritten Studie wurde ein Algorithmus entwickelt, der eine 3D-Residual-U-Net-Architektur mit integriertem Histogramm-Matching für die Erkennung und Überwachung von Multipler Sklerose (MS) durch Voxel-Guided Morphometry (VGM)-Karten verwendet. Dieses Modell wurde entwickelt, um MS-bedingte Hirnstrukturveränderungen in MRT-Volumina genau hervorzuheben, und hat seine Anpassungsfähigkeit und Verallgemeinerbarkeit für ungesehene Datensätze bewiesen. Bei der Validierung an verschiedenen Patientendatensätzen wurde eine durchschnittliche Verbesserung des mittleren absoluten Fehlers (Mean Absolute Error, MAE) von 4,2% gegenüber der Referenzmethode erzielt, was seine Robustheit und sein Potenzial als klinisches Werkzeug für eine präzise, effiziente Analyse der Dynamik von MS-Läsionen bestätigt.

In der abschließenden Studie wurde ein CNN-basiertes System zur automatischen Erkennung akuter abdominaler Aortendissektionen (AD) in CT-Scans entwickelt. Dieser Ansatz, der an einem kleinen internen Datensatz trainiert wurde, wurde anhand eines großen externen Datensatzes weiter validiert. Das Modell zeigte eine hohe AUC und Sensitivität bei der Erkennung von AD und bestätigte damit seine Zuverlässigkeit und sein Potenzial, die Notfallradiologie durch KI-gestützte Diagnose zu verändern.

Diese Arbeit unterstreicht das Potenzial von Deep Learning bei der Überwindung der Beschränkungen durch kleine Datensätze in der medizinischen Bildanalyse und ebnet den Weg für eine zugänglichere und effizientere KI-gestützte Diagnose und Therapieplanung.

Summary of publications included in cumulative dissertations

Name of the doctoral student: Anish Raj

Title of dissertation: Deep Learning based Medical Image Analysis using Small Datasets

Supervised by: Prof. Frank G. Zöllner

☐ I wish to submit a cumulative dissertation, and hereby ask the doctoral committee to examine whether the quantity and quality of the proposed publications are sufficient to meet the requirements for a cumulative dissertation.

☒ The doctoral committee has previously examined whether my publications are suitable for a cumulative dissertation, and this is a final overview of the publications contained in my cumulative thesis.

1. List of peer-reviewed publications included in the cumulative dissertation. For each publication, provide a complete list of authors, title, journal, journal impact factor, and whether the manuscript has been accepted for publication, is in revision after peer review, or has been submitted and is awaiting peer review. Shared first authorships should be clearly indicated. Please also indicate whether the publication is an original research report, a review, or another type of article.

Publication 1: Raj, A., Tollens, F., Hansen, L., Golla, A. K., Schad, L. R., Nörenberg, D., & Zöllner, F. G. (2022). Deep learning-based total kidney volume segmentation in autosomal dominant polycystic kidney disease using attention, cosine loss, and sharpness aware minimization. Diagnostics, 12(5), 1159, doi: 10.3390/diagnostics12051159.
Original research report, Published, Impact Factor: 3.6

Publication 2: Raj, A., Tollens, F., Caroli, A., Nörenberg, D., & Zöllner, F. G. (2023). Automated prognosis of renal function decline in ADPKD patients using deep learning. Zeitschrift für Medizinische Physik, Volume 34, Issue 2, 2024, Pages 330-342, doi: 10.1016/j.zemedi.2023.08.001.
Original research report, Published, Impact Factor: 2.0

Publication 3: Raj, A., Gass, A., Eisele, P., Dabringhaus, A., Kraemer, M., & Zöllner, F.G. (2024). A generalizable deep voxel-guided morphometry algorithm for the detection of subtle lesion dynamics in multiple sclerosis. Frontiers in Neuroscience, 18, 1326108, doi: 10.3389/fnins.2024.1326108.
Original research report, Published, Impact Factor: 4.3

Publication 4: Raj, A., Allababidi, A., Kayed, H., Gerken, A. LH., Müller, J., Schoenberg, S. O., Zöllner, F. G., & Rink, J. S. (2024). Streamlining acute Abdominal Aortic Dissection management - an AI based CT imaging workflow. Journal of Imaging Informatics in Medicine, doi: 10.1007/s10278-024-01164-0.
Original research report, Published, Impact Factor: 4.4

Publication 5: Raj, A., Hansen, L., Tollens, F., Nörenberg, D., Villa, G., Caroli, A., & Zöllner, F. G. (2024). Generalizable Kidney Segmentation for Total Volume Estimation. Proc. Bildverarbeitung für die Medizin 2024, Erlangen, Germany, pp.285–290, doi: 10.1007/978-3-658-44037-4_75.
Original research report, Published, Impact Factor: Does not exist (peer-reviewed conference proceedings)

2. Summary of the doctoral student's contribution to the work reported in each manuscript.

Work steps	Publication 1	Publication 2	Publication 3	Publication 4	Publication 5
Conception (%)	100	100	80	80	100

Literature search (%)	100	100	90	80	100
Ethics proposal (%)	0	0	0	0	0
Animal experimentation proposal (%)	-	-	-	-	-
Data collection (%)	0	20	30	20	0
Data analysis (%)	100	100	100	100	100
Interpretation of results (%)	100	100	70	90	100
Manuscript writing (%)	90	90	90	80	90
Revision (%)	100	100	100	80	100
Indicate which figures and tables resulted from your dissertation work.					

3. For cumulative dissertations, the doctoral student should be first author on at least two of the publications. In the case of joint first authorship or last authorship, please justify below why the publication should be considered equivalent to a single first authorship.

_____

_____

(4) I hereby certify that this is a true representation of the doctoral student's contribution to the publications listed.

Signature of the doctoral candidate

Signature of supervisor

Contents

List of Figures

List of Tables

List of Acronyms

AAA	Abdominal Aortic Aneurysm
AD	Aortic Dissection
ADPKD	Autosomal Dominant Polycystic Kidney Disease
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under the Curve
CBAM	Convolution Block Attention Module
CI	Confidence Interval
CKD	Chronic Kidney Disease
CNN	Convolutional Neural Network
CNS	Central Nervous System
CT	Computed Tomography
DL	Deep Learning
DSC	Dice Similarity Coefficient
eGFR	estimated Glomerular Filtration Rate
ELU	Exponential Linear Unit
FC	Fully Connected
FCNN	Fully Convolutional Neural Network
FN	False Negative
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
HtTKV	Height-adjusted Total Kidney Volume
HU	Hounsfield Unit
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multi-Layer-Perceptron
MRI	Magnetic Resonance Imaging
MS	Multiple Sclerosis
MSCT	Multi-slice Computed Tomography

MSE	Mean Squared Error
MSSD	Mean Symmetric Surface Distance
PReLU	Parametric Rectified Linear Unit
ReLU	Rectified Linear Unit
RF	Radio Frequency
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SAM	Sharpness Aware Minimization
SE	Squeeze and Excitation
SGD	Stochastic Gradient Descent
SSIM	Structural Similarity Index Measure
TKV	Total Kidney Volume
VGM	Voxel-Guided Morphometry

1. Introduction and Outline

1.1 Motivation

Since 2012, the rise of Deep Learning (*DL*), initiated by breakthroughs such as AlexNet [1], has transformed how we analyze digital images, especially in healthcare. The rapid adoption of emerging technologies such as Convolutional Neural Networks (*CNNs*) has demonstrated the profound impact these can have on the interpretation of medical images, including Magnetic Resonance Imaging (*MRI*) and Computed Tomography (*CT*) scans [2, 3]. The quantity of these scans has steadily increased, creating a growing demand for their expert analysis. This surge highlights the critical need for automation in the processing and assessment of medical images, a task that remains predominantly time-consuming due to the necessity of manual Region of Interest (*ROI*) delineation for volume-based analysis [4]. Automated algorithms can bring improvement in this area by offering not only a reduction in analysis time but also providing objective, reproducible measurements to assist physicians in their assessments.

However, the integration of DL into the healthcare sector is not without its challenges. Among these, the most significant is the scarcity of large, annotated datasets, compounded by privacy regulations, the rarity of certain medical conditions, and the logistical hurdles involved in compiling such datasets [5, 6]. Furthermore, data originating from different sources (sites), such as different MRI machines and protocols, introduce variability in data distributions, posing a substantial challenge for DL models. This variability often means that models trained on data from one source cannot be effectively generalized to data from unknown sources [7]. Consequently, the need for large, diverse datasets for the training of effective DL models [8, 9] becomes a bottleneck that hinders the development and deployment of automated solutions capable of handling the intricacies of medical image analysis with the desired accuracy and robustness.

This thesis aims to bridge this gap by employing DL methods that are optimized for performance on smaller datasets. By adopting approaches such as attention mechanisms, modified loss functions and optimizers in conjunction with preprocessing methods for data harmonization, this work presents a set of methods that aim to improve the generalizability and efficacy of medical image analysis algorithms. These strategies are thoroughly applied across a variety of medical imaging tasks such as segmentation, classification, and regression, demonstrating their versatility and potential to improve diagnostic accuracy, facilitate disease monitoring, and eventually contribute to personalized patient care.

By evaluating these methodologies across diverse medical imaging tasks, including the segmentation of Total Kidney Volume (*TKV*) in Autosomal Dominant Polycystic Kidney Disease (*ADPKD*) patients and the quantitative analysis of Multiple Sclerosis (*MS*) progression, this thesis not only highlights the practical applications of these DL techniques but also explores their generalizability to different diseases. This comprehensive evaluation underscores the potential of these techniques to become integral components of Machine Learning (*ML*) in healthcare, addressing important needs for efficient, reliable, and scalable solutions for medical image analysis.

1.2 Outline

The composition of this thesis is cumulative, covering three main areas within medical image processing, namely:

1. Segmentation,
2. Classification, and
3. Regression.

Chapter 3 through Chapter 7 each present a self-sufficient scientific study. Consequently, each such chapter contains an introduction, a detailed exposition of materials and methods, a presentation of results, followed by a discussion, and a conclusion. Additionally, an acknowledgment of contributions and funding support is mentioned at the conclusion of each chapter.

Chapter 2 is designed to provide a concise overview of medical imaging, with an emphasis on CT and MRI. It also introduces the fundamental principles of DL in image processing and an introduction to the diseases that were analyzed in this thesis.

In Chapter 3, the application of image segmentation is presented. This chapter details the training of various CNNs enhanced by attention mechanisms, the cosine loss function, and Sharpness Aware Minimization (*SAM*), aimed at segmenting kidney volumes in patients with ADPKD through T1-w MRI. The combination of the aforementioned methodologies facilitated achieving performance on par with that of human experts in the estimation of TKV.

Chapter 4 presents a small expansion upon the research presented in Chapter 3. This study details the utilization of two datasets to develop and validate a robust algorithm for TKV segmentation in patients with ADPKD. The method incorporates histogram matching to achieve data distribution harmonization, followed by the training of a CNN on one dataset. The resulting algorithm demonstrates the capacity to generalize effectively to the unseen dataset.

Chapter 5 introduces an application derived from the initial study presented in Chapter 3. This subsequent study details how kidney volumes, automatically segmented with the help of the first study, were combined with data from additional biomarkers. Through the integration of a CNN and a Multi-Layer-Perceptron (*MLP*), this approach aids in forecasting the progression of kidney conditions in patients with ADPKD after an eight-year period.

Chapter 6 introduces a generalizable DL algorithm capable of generating Voxel-Guided Morphometry (*VGM*) for the quantification of MS progression efficiently and effectively across multiple MRI datasets. This research aligns the intensity distributions from disparate MRI datasets and utilizes attention-based CNNs to surpass the existing baseline reference method in performance across all datasets. Demonstrating robustness, the algorithm offers potential for implementation in clinical environments to analyze MS progression and the effectiveness of treatments.

Chapter 7 presents another use case for image classification. Here, a 3D CNN was developed to identify abdominal Aortic Dissection (*AD*) using CT scans. Initially, the abdominal area is automatically isolated, followed by the extraction of the aortic

ROI. Subsequently, the CNN discriminates between AD and non-AD cases within the aortic ROI. The algorithm has been further validated on a substantial external dataset, demonstrating high sensitivity. This approach holds promise for integration into emergency clinical procedures, enabling automated detection of AD.

Chapter 8 provides a comprehensive summary of the thesis, including an in-depth review of the outcomes from the scientific studies detailed from Chapter 3 through Chapter 7.

Chapter 9 outlines potential avenues for future research and delves into both the significance and the limitations of the work that has been presented.

1.3 Citations of Previous Publications

Several chapters of this thesis have already been published. The citations for these chapters are:

Chapter 3: Raj, A., Tollens, F., Hansen, L., Golla, A. K., Schad, L. R., Nörenberg, D., & Zöllner, F. G. (2022). Deep learning-based total kidney volume segmentation in autosomal dominant polycystic kidney disease using attention, cosine loss, and sharpness aware minimization. Diagnostics, 12(5), 1159, doi: https://doi.org/10.3390/diagnostics12051159.

Chapter 4: Raj, A., Hansen, L., Tollens, F., Nörenberg, D., Villa, G., Caroli, A., & Zöllner, F. G. (2024). Generalizable Kidney Segmentation for Total Volume Estimation. Proc. Bildverarbeitung für die Medizin 2024, Erlangen, Germany, pp.285-290, doi: https://doi.org/10.1007/978-3-658-44037-4_75.

Chapter 5: Raj, A., Tollens, F., Caroli, A., Nörenberg, D., & Zöllner, F. G. (2023). Automated prognosis of renal function decline in ADPKD patients using deep learning. Zeitschrift für Medizinische Physik, Volume 34, Issue 2, 2024, Pages 330-342, doi: https://doi.org/10.1016/j.zemedi.2023.08.001.

Chapter 6: Raj, A., Gass, A., Eisele, P., Dabringhaus, A., Kraemer, M., & Zöllner, F.G. (2024). A generalizable deep voxel-guided morphometry algorithm for the detection of subtle lesion dynamics in multiple sclerosis. Frontiers in Neuroscience, 18, 1326108, doi: https://doi.org/10.3389/fnins.2024.1326108.

Chapter 7: Raj, A., Allababidi, A., Kayed, H., Gerken, A. LH., Müller, J., Schoenberg, S. O., Zöllner, F. G., & Rink, J. S. (2024). Streamlining acute Abdominal Aortic Dissection management - an AI based CT imaging workflow. Journal of Imaging Informatics in Medicine, doi: https://doi.org/10.1007/s10278-024-01164-0.

2. Background

2.1 Medical Imaging

The realm of medical imaging encompasses various techniques that utilize physical phenomena to capture detailed images of a patient's anatomy. These images are invaluable for diagnosis, therapeutic planning, and tracking disease progression over time [10, 11]. Due to the unique information each imaging modality offers about the tissues being examined, the integration of different imaging techniques is frequently employed. This strategy ensures a comprehensive evaluation of the patient's condition. In the sections that follow, Magnetic Resonance Imaging (*MRI*) and Computed Tomography (*CT*) are described as they are relevant to this thesis.

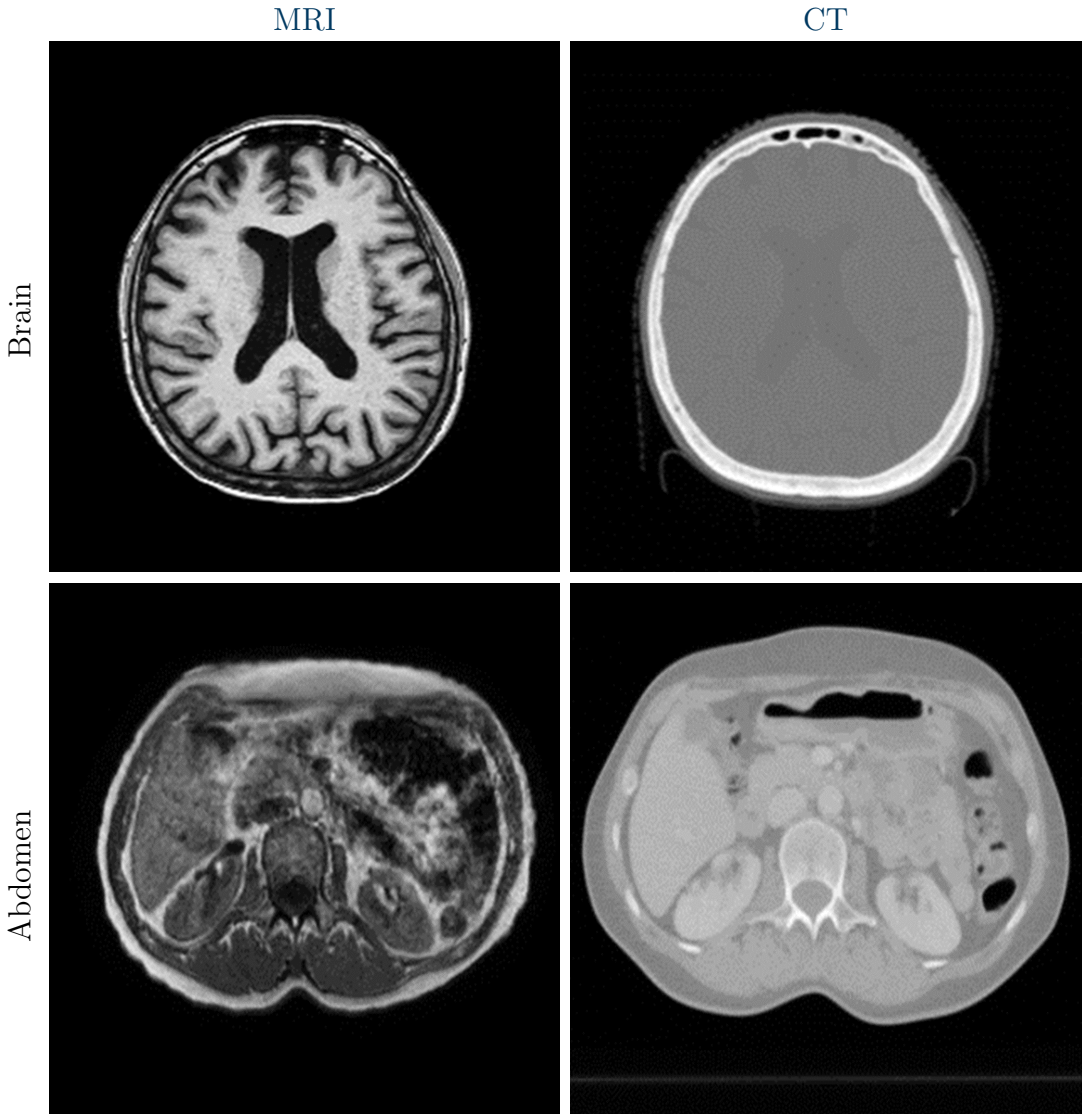


Figure 2.1: Example of MRI and CT scans of the brain and abdomen region.

2.1.1 Magnetic Resonance Imaging

MRI operates on the fundamental concept of nuclear magnetic resonance. This phenomenon involves the absorption and subsequent emission of electromagnetic radiation by atomic nuclei when they are subjected to an external magnetic field [12].

During an MRI procedure, a patient is positioned within a powerful magnetic field denoted as B_0. This causes the spins of the hydrogen protons in the body to align predominantly along the magnetic field, with the B_0 direction designated as the z-axis in the MRI coordinate system. The alignment of these protons generates a secondary magnetic field, also along the z-axis, which oscillates in sync with the proton spins, thereby inducing an electrical current in nearby coils. The rate at which these protons spin around the z-axis, known as the Larmor frequency ω, is influenced by the magnetic field strength B_0 and a specific constant called the gyromagnetic ratio γ (Equation 2.1).

$$\omega = -\gamma B_0 \tag{2.1}$$

A Radio Frequency (*RF*) pulse is applied to alter the axis of proton precession. This pulse must match the proton's Larmor frequency ω for energy absorption to occur. The extent to which the precession axis is altered is a function of the RF pulse's duration. This adjustment leads to a change in the orientation of the proton's magnetic field. Post RF pulse application, the system's magnetization returns to its initial state, with proton spins realigning with B_0 (known as relaxation). The realignment speed varies across different body tissues, characterized by two primary relaxation times: T_1 and T_2.

The flipping of the proton's precession axis away from the z-axis diminishes the z-axis magnetization (M_z). Over time, M_z gradually recovers as the spins realign with B_0, eventually restoring to its original state $M_{z,0}$ as $t \to \infty$. This process is described by the tissue-specific T_1 time constant:

$$M_z(t) = M_{z,0} - (M_{z,0} - M_z(0))\, e^{-\frac{t}{T_1}} \tag{2.2}$$

Conversely, when the precession axis is adjusted by 90°, it generates a magnetization within the x,y-plane, orthogonal to B_0. This transverse magnetization ($M_{x,y}$) fades as the spins reorient with the z-axis, approaching zero as $t \to \infty$, governed by the T_2 relaxation time:

$$M_{x,y}(t) = M_{x,y}(0)\, e^{-\frac{t}{T_2}} \tag{2.3}$$

Spatial differentiation of the origins of the electrical currents detected in the coils is achieved by modifying the magnetic field strength through continuous linear gradient fields G_x, G_y and G_z. This variation results in a position-dependent Larmor frequency (Equation 2.4) for each proton, enabling what is known as frequency encoding. Phase encoding is another technique utilized to map the scan volume, involving the temporary application of a gradient field between the RF pulse and signal acquisition, leading to spin dephasing and varied signal measurement.

$$\omega = -\gamma(B_0 + G_x x + G_y y + G_z z) \tag{2.4}$$

MRI's capability to produce images with varying contrasts depends on the sequence of radio frequency pulses and imaging parameters used. Although MRI primarily provides qualitative imagery, with image intensity values lacking standard measurement units, quantitative MRI methods are available but less frequently applied [13]. Given MRI's sensitivity to protons, water-rich tissues yield the strongest signals, whereas bone and other low water-content tissues are less visible or not at all. For instance, in Figure 2.1, a T1-w brain MRI shows high contrast among different brain tissues with the skull appearing as a dark outline, while an abdominal T1-w MRI reveals distinct contrasts among various soft tissues.

2.1.2 Computed Tomography

CT is an imaging technique that employs the attenuation properties of X-rays to generate detailed images of the internal structures of the human body [14]. Unlike traditional X-ray imaging, which can result in the overlay of images from different anatomical features, CT technology produces three-dimensional (3D) volumetric data, eliminating issues related to superimposition.

The core components of a CT scanner include an X-ray generator and a detection system, both mounted on a rotating gantry. This setup encircles a patient table that can move as required. The generation of X-rays occurs when high-speed electrons collide with a metal anode within the X-ray tube, slowing abruptly. The resultant X-ray photon energy is contingent on the electron's velocity, which is influenced by the applied acceleration voltage, typically ranging from 25kV to 150kV for medical diagnostics [15]. The intensity of the X-ray beam produced is regulated by the current flowing to the anode.

As X-rays traverse the body, their intensity diminishes due to attenuation, which occurs exponentially. This attenuation is described by the Beer-Lambert law:

$$I = I_0 e^{-\mu x}$$

where I is the transmitted intensity, I_0 is the initial intensity, μ is the linear attenuation coefficient, and x is the thickness of the material.

This attenuation is the result of various interactions within the body, including Rayleigh scattering, Compton scattering, and the photoelectric effect. The extent of attenuation is dependent on the type of material the X-rays pass through and their wavelength. In the range of energies used for diagnostic purposes, photoelectric absorption and Compton scattering are predominantly responsible for attenuation, with Rayleigh scattering playing a minimal role.

- **Rayleigh Scattering (Coherent Scattering)**: Rayleigh scattering involves the elastic scattering of X-rays by bound electrons without a change in energy (wavelength). This type of scattering is more significant at lower photon energies and contributes less to attenuation at the higher energies used in diagnostic X-rays.

- **Compton Scattering (Incoherent Scattering)**: In Compton scattering, X-ray photons are scattered by outer-shell electrons, resulting in a loss of energy and an increase in wavelength. The change in wavelength ($\Delta\lambda$) is given by the Compton wavelength shift equation:

$$\Delta\lambda = \lambda' - \lambda = \lambda_c(1 - \cos\theta)$$

  where λ is the incident wavelength, λ' is the scattered wavelength, $\lambda_c = \frac{h}{m_e c}$ is the Compton wavelength of the electron, and θ is the scattering angle. Compton scattering is significant at the photon energies typically used in medical imaging.
- **Photoelectric Effect (Photoelectric Absorption)**: In the photoelectric effect, X-ray photons are completely absorbed, resulting in the ejection of electrons from inner atomic shells. The probability of photoelectric absorption (τ) depends strongly on the photon energy (E) and the atomic number (Z) of the absorbing material, approximately following the relation:

$$\tau \propto \frac{Z^3}{E^3}$$

The overall attenuation coefficient (μ) is the sum of the contributions from these interactions:

$$\mu = \mu_{\text{Rayleigh}} + \mu_{\text{Compton}} + \mu_{\text{photoelectric}}$$

Upon exiting the body, X-rays are captured by the detection system, where a collimator precedes the detector array to eliminate scatter-induced noise by filtering out non-direct photons. The detector array consists of numerous detector elements, each comprising a scintillator to convert X-ray photons to visible light, and a photodetector to transform this light into electrical signals.

The architecture of the detector array varies with the scanner model. Multi-slice Computed Tomography (*MSCT*) scanners, for example, feature multiple rows of detectors arranged in a curved format, working in concert with a fan-shaped X-ray beam. The scanning process encompasses several rotational angles around the patient, with each angle providing a distinct radiographic projection. These projections are then used to compute the spatial distribution of attenuation coefficients across the scanned volume. For comprehensive coverage, the patient table moves incrementally between scans.

The reconstruction of CT images from collected projections is achieved through algorithms such as filtered back projection or iterative reconstruction techniques [16]. This process results in a voxel-based representation of the scanned volume, with each voxel reflecting the average attenuation coefficient of the contained subvolume, determined by slice thickness and spatial resolution.

Radiodensity within CT images is quantified using Hounsfield Units (*HUs*), a scale established by Hounsfield [14]. This scale linearly correlates a voxel's attenuation coefficient with those of water and air at standard conditions, facilitating uniform

image interpretation across different CT systems. The equation for calculating HUs is as follows:

$$\text{CT-Number}(\mu) = \frac{\mu - \mu_{\text{water}}}{\mu_{\text{water}} - \mu_{\text{air}}} \cdot 1000\,\text{HU} \tag{2.5}$$

The HU scale effectively differentiates between various tissue types, each occupying distinct ranges on the scale, as illustrated in Figure 2.2. This calibration allows for the consistent comparison of CT images, regardless of the scanner used. Examples of CT imaging, such as cranial and abdominal scans (Figure 2.1), demonstrate the contrast capabilities of CT, showcasing differences in tissue densities and compositions as represented on the HU scale.

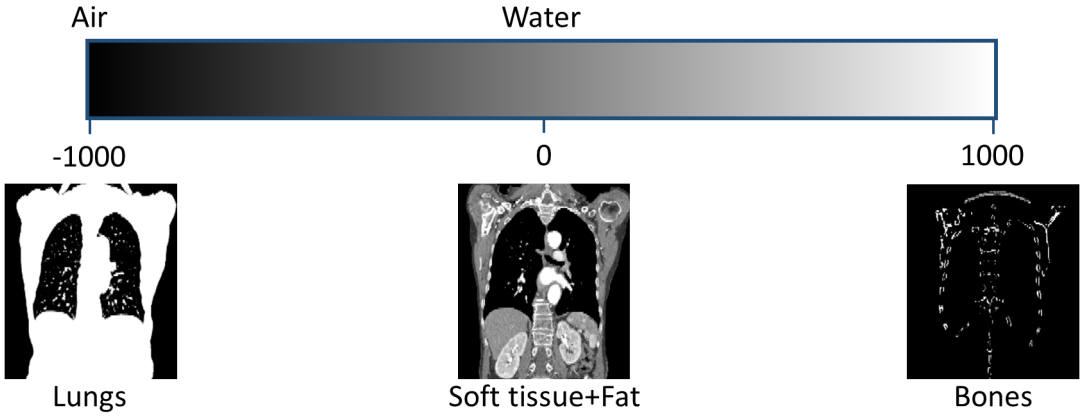


Figure 2.2: The HU scale illustrating the categorization of various tissue types.

2.2 Deep Learning

Deep learning, is a subset of machine learning that is inspired by the operational mechanisms of the human brain's neural networks [17, 18]. It aims at enabling machines to learn from data and make decisions without being explicitly programmed for specific tasks. This approach leverages Artificial Neural Networks (*ANNs*) to facilitate the automatic extraction and learning of complex patterns from large datasets [18]. Unlike traditional machine learning algorithms that necessitate manual feature selection, deep learning algorithms autonomously identify relevant features through a hierarchical learning process, progressively building complex representations.

The ANNs are composed of layers of interconnected nodes or neurons, each layer designed to perform specific transformations on its inputs. This architecture enables the modeling of complex functions through the depth of the network, which is characterized by the number of hidden layers.

Deep learning excels in tasks that involve large-scale and high-dimensional data, including but not limited to image and speech recognition, natural language processing, and predictive analytics. Through its capacity to learn from experience and

understand the world in terms of a hierarchy of concepts, deep learning can achieve a level of accuracy that often surpasses human capability in certain domains [19]. The effectiveness of deep learning models grows with the amount of available data, leveraging computational power and sophisticated algorithms to uncover insights within complex data structures [9]. In the following sections, we delve into the fundamental principles of deep learning, with a particular focus on Convolutional Neural Networks (*CNNs*), which are utilized within this thesis for image analysis.

2.2.1 Mathematical Model of Artificial Neural Networks

The mathematical foundation of ANNs is predicated on the concept of simulating the neural processing of the human brain to perform a wide array of tasks ranging from simple to highly complex [18]. An ANN is composed of nodes (neurons) arranged in layers: an input layer to receive the data, several hidden layers to process the data, and an output layer to produce the desired outcome.

Mathematically, the operation within a neuron involves the weighted sum of its inputs plus a bias term, which is then passed through an activation function to determine the neuron's output. For a given neuron, let $x_1, x_2, \ldots, x_n$ represent the inputs, with each input x_i associated with a weight w_i, and b denote the bias. The weighted sum, z, is given by:

$$z = \sum_{i=1}^{n} w_i x_i + b$$

The output of the neuron, y, is obtained by applying an activation function f to z:

$$y = f(z)$$

This process allows the network to model non-linear relationships. The choice of activation function is crucial for the network's ability to capture complex patterns in the data, with common examples including the sigmoid and Rectified Linear Unit (*ReLU*) functions.

In multi-layer networks, the output of one layer serves as the input to the next. This forward propagation of data through the network facilitates the hierarchical learning of features, where each layer captures increasingly abstract representations of the input data. The learning process in an ANN involves adjusting the weights and biases to minimize a loss function that measures the difference between the network's predictions and the actual target values. Through techniques such as backpropagation and optimization algorithms like gradient descent, the network iteratively updates its parameters to improve its performance on a given task.

This mathematical framework underpins the ability of deep learning models to learn from and make predictions on data, forming the basis for their application across a diverse range of fields and challenges in artificial intelligence.

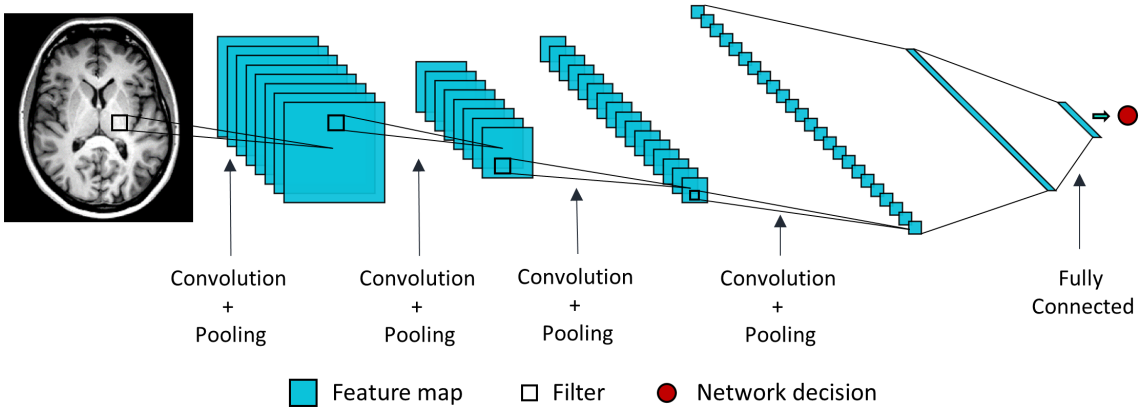


Figure 2.3: An illustration of a typical CNN architecture used for image classification tasks. The CNN consists of 4 convolutional layers, 4 pooling layers and 2 fully connected layers.

2.2.2 Convolutional Neural Networks

CNNs draw their architectural inspiration from the biological processes observed in the visual cortex of cats, as discovered by Hubel and Wiesel [20]. This seminal work identified that the visual cortex comprises small, specialized regions sensitive to specific areas of the visual field, with different cells responding to various shapes and orientations. This hierarchical and layered structure of visual processing in mammals laid the groundwork for the development of CNNs, aimed at extracting features from data in a similarly hierarchical manner [21, 22].

Designed to emulate this spatial organization, CNNs excel in handling data with inherent spatial relationships, particularly image and video data. Their architecture is structured to preserve spatial relationships across layers, ensuring that features extracted at each layer are relevant to localized regions of the input. This is achieved through a combination of different types of layers:

- **Convolution layers** perform feature extraction by applying filters that capture spatial patterns such as edges or textures.
- **Pooling layers** reduce dimensionality and computational complexity by summarizing the features in each region.
- **Activation functions** introduce non-linearity, enabling the network to learn complex patterns.
- **Fully connected layers** integrate learned features across the image into a final output, such as a class label.

CNNs are distinguished by their ability to learn hierarchical feature representations automatically, making them highly effective for tasks like image classification, regression, and beyond. Figure 2.3 illustrates a typical architecture of a CNN.

Subsequent sections (i.e., Section 2.2.2.1 to Section 2.2.2.4) will delve into the specifics of each layer type, offering a detailed description of their roles and mechanisms within the CNN framework.

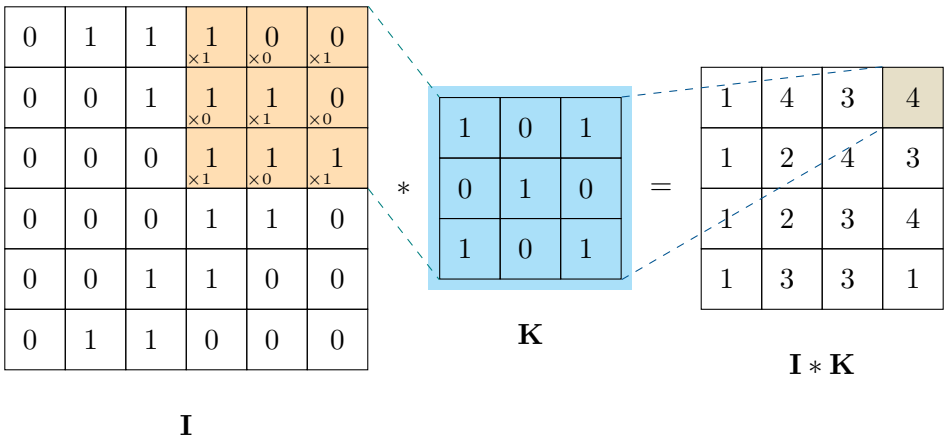


Figure 2.4: Illustration of the convolution operation showing the kernel **K** applied to the input matrix **I**, resulting in the convolution output $\mathbf{I} * \mathbf{K}$.

2.2.2.1 Convolution Layer

The convolutional layer is a fundamental building block of CNNs, designed to automatically and adaptively learn spatial hierarchies of features from input images. The convolutional layers employ a set of parameters organized into 3-dimensional structures known as filters or kernels, predominantly square in their spatial dimensions (e.g., 3×3 or 5×5), with the filter depth matching that of the layer upon which it is applied. A mathematical operation called convolution is at the heart of the convolution layer. It involves sliding a filter over the input image and computing the dot product of the filter weights with the local region it covers in the input. This process determines the spatial dimensions (i.e., height and width) of the subsequent layer's output.

Given an input or feature map in the q-th layer with dimensions $L_q \times B_q \times d_q$ (representing height, width, and depth respectively), and assuming a filter of size $F_q \times F_q \times d_q$, the spatial dimensions of the output (or the $(q+1)$-th layer) can be defined as:

$$L_{q+1} = L_q - F_q + 1,$$
$$B_{q+1} = B_q - F_q + 1.$$

This configuration is depicted in Figure 2.4, illustrating how an initial input matrix **I** with dimensions $6 \times 6 \times 1$ undergoes convolution with a kernel **K** sized $3 \times 3 \times 1$. This operation yields a convolution output $\mathbf{I} * \mathbf{K}$, with resultant dimensions $4 \times 4 \times 1$. The process demonstrates the application of a single kernel, leading to a feature map that highlights the transformation of the input matrix through the convolution operation. The depth of the output feature map, in this scenario being 1, indicates the number of unique kernels utilized.

The term "parameter footprint" refers to the extent of parameters involved in convolution, calculated as $(F_q^2 \times d_q \times (d_q + 1))$ for the q-th layer, directly influencing the model's capacity by dictating the number of learnable parameters. Filters are designed to identify specific spatial patterns within small image regions, necessitating a diverse set of filters for comprehensive feature extraction.

Mathematically, the convolution operation from the q-th to the $(q+1)$-th layer is formalized as:

$$h_{ijp}^{(q+1)} = \sum_{r=1}^{F_q} \sum_{s=1}^{F_q} \sum_{k=1}^{d_q} w_{rsk}^{(p,q)} h_{i+r-1,j+s-1,k}^{(q)}, \tag{2.6}$$

$$\forall i \in \{1..., L_q - F_q + 1\}$$
$$\forall j \in \{1..., B_q - F_q + 1\}$$
$$\forall p \in \{1..., d_{q+1}\}$$

where $h_{ijp}^{(q+1)}$ denotes the feature map at position (i, j) in the p-th feature map of the $(q+1)$-th layer, and $w_{rsk}^{(p,q)}$ represents the weight of the p-th filter at position (r, s, k) in the q-th layer. It should be highlighted that the initial layers predominantly undertake the task of identifying simpler geometric configurations, such as edges, whereas subsequent layers progressively synthesize more complex formations from these elementary shapes [23]. Moreover, convolution inherently demonstrates translation equivariance, ensuring that any displacement within the input image correspondingly shifts the feature map equivalently. A critical aspect to consider is the expansion of the receptive field from the q-th to $(q+1)$-th layer; this increment signifies that each subsequent feature can encompass a more extensive spatial area of the input layer, thereby capturing broader contextual information.

One challenge with convolution is the reduction in spatial dimensions from one layer to the next, which can lead to loss of information, particularly at the borders of the image. This is mitigated by padding the input with zeros around its borders, ensuring the output layer maintains the spatial dimensions of the input. However, sometimes it is not important to perform convolution at every spatial position and so it is advisable to reduce the level of granularity of convolution. The use of strides can adjust the granularity of the convolution operation, affecting both the receptive field and spatial dimensions of the output. Employing a stride of S_q alters the output dimensions to:

$$\text{Height:} \quad \frac{L_q - F_q}{S_q} + 1,$$
$$\text{Width:} \quad \frac{B_q - F_q}{S_q} + 1.$$

The primary effect of utilizing larger strides (> 1) is to swiftly expand the receptive field while simultaneously diminishing the spatial dimensions of the layer [22].

2.2.2.2 Pooling Layer

The pooling operation functions on small grid regions of size $P_q \times P_q$ in every layer and outputs a layer with the same depth. The most widely used pooling operation is max-pooling, where the maximum value is returned for every region of size $P_q \times P_q$ on the activation maps. Usually, a stride greater than 1 is used in pooling operation, and in these events the new dimensions are given by

$$Height : (L_q - P_q)/S_q + 1$$

$$Width : (B_q - P_q)/S_q + 1$$

The Figure 2.5 illustrates the max-pool operation. Another pooling operation is average pooling, where the average value is returned from the region of size $P_q \times P_q$ on the activation maps.

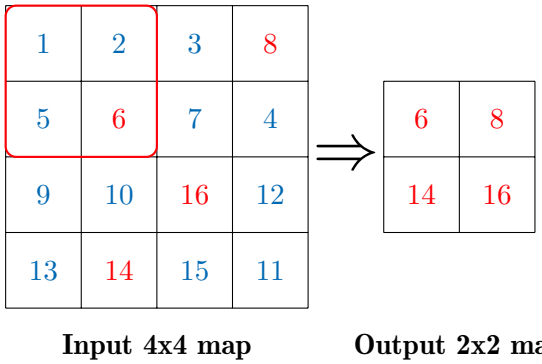


Figure 2.5: Illustration of the Max Pooling Operation. A 2x2 pooling window (highlighted in red) is applied to the input 4x4 feature map (stride 2), resulting in the 2x2 output map based on the maximum values in each window.

This operation greatly reduces the spatial dimensions of the activation maps [22]. Pooling layers are integral to CNNs due to their capacity to diminish the dimensionality of feature maps, thereby decreasing computational demands. Additionally, these layers enhance the network's resilience to minor translations and facilitate the extraction of more abstract features from the input image.

2.2.2.3 Activation Functions

Given that the majority of real-world problems exhibit non-linear characteristics, activation functions are employed within both intermediate and final layers of a CNN to enable the network to model non-linear behaviors. These functions are crucial for imparting non-linearity into the convolution operations (which are inherently linear), thereby enhancing the network's ability to learn complex patterns.

Some of the most commonly used activation functions within the intermediate layers of a CNN are ReLUs, Exponential Linear Units ($ELUs$), and Parametric Rectified Linear Units ($PReLUs$).

- **ReLU**

$$ReLU(x) = \max(0, x). \tag{2.7}$$

  This function outputs zero for any negative input and returns the input itself for any positive input.

- **ELU**

$$ELU(x) = \begin{cases} x & \text{if } x > 0, \\ \alpha(e^x - 1) & \text{if } x \leq 0. \end{cases} \tag{2.8}$$

  Here, α is a constant that controls the saturation level for the ELU function for negative inputs.

- **PReLU**

$$PReLU(x) = \begin{cases} x & \text{if } x > 0, \\ \alpha x & \text{if } x \leq 0. \end{cases} \tag{2.9}$$

  In PReLU, α is a learnable parameter, which allows the network to adapt this aspect of the activation function during the training process.

The final layer's activation function is chosen based on the prediction goal. For regression tasks, a linear function $f(x) = x$ is used, allowing for continuous output suitable for predicting numerical values.

For binary classification, the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ is preferred, mapping inputs to a probability between 0 and 1, ideal for distinguishing between two classes.

In multi-class classification scenarios, the softmax function is applied: $\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$, which converts logits into normalized probabilities, ensuring the model's outputs sum to 1 and are interpretable as class probabilities.

2.2.2.4 Fully Connected Layer

In CNNs, Fully Connected (*FC*) layers, or dense layers, are typically positioned at the end, serving as a crucial component for decision-making tasks like classification and regression. Each neuron in an FC layer is interconnected with all activations from the preceding layer, facilitating the integration and transformation of learned features into a format suitable for making predictions.

While convolutional layers are adept at feature extraction due to their local connectivity and shared weights, FC layers excel in aggregating these features into a global representation, thereby enabling the network to understand the broader context of the input data. This characteristic stems from their dense connectivity, which, although computationally intensive, significantly amplifies the network's learning capacity. It is common practice to employ multiple FC layers to enhance the model's computational power and to refine the abstraction levels of the features being learned.

However, this advantage comes at the cost of an increased parameter count, leading to a substantial parameter footprint for FC layers compared to their convolutional counterparts. This dense interconnectivity implies that FC layers are more prone to overfitting, especially when dealing with high-dimensional input data. As such, regularization techniques are often applied within these layers to mitigate overfitting.

2.2.2.5 CNN Architectures

CNNs have recently seen rapid evolution since their inception in the 1980s, leading to groundbreaking architectures tailored for diverse applications, including image classification, object detection, and semantic segmentation. This section delves into the pioneering architectures that have shaped the landscape of deep learning in image processing.

Influential CNN Architectures

- **LeNet-5:** Introduced by LeCun et al. [21], *LeNet-5* is considered the first CNN, designed for handwritten digit recognition. It laid the groundwork for CNNs with its structure of convolutional, pooling, and fully connected layers.
- **AlexNet:** The *AlexNet* architecture marked a revolution in image classification when it won the ImageNet challenge in 2012 [1]. Comprising five convolutional layers, three max-pooling layers, and three fully connected layers, AlexNet's success underscored the potential of CNNs, facilitated by the growing capabilities of Graphics Processing Units (*GPUs*) and the availability of large datasets [24].
- **VGG-16:** Building on the foundations laid by AlexNet, *VGG-16* further demonstrated the importance of network depth in achieving higher accuracy in image classification tasks [25]. Its architecture features sixteen convolutional and fully connected layers, emphasizing the use of small convolutional filters for deep networks.
- **ResNet:** The introduction of *ResNet* marked a significant advancement, enabling the training of substantially deeper networks through the use of residual blocks [26]. This innovation effectively addressed the vanishing gradient problem, paving the way for networks with depths of over a hundred layers.

Extension to Semantic Segmentation

Semantic segmentation architectures are designed to classify each pixel of an image into a predefined category, enabling the precise delineation of objects within images. Unlike image classification models that output a single prediction for the entire image, semantic segmentation models produce a pixel-wise map of classifications. A typical semantic segmentation architecture comprises two main components: an encoder and a decoder.

Encoder: The encoder part of a semantic segmentation network is responsible for capturing the contextual information within the image. It progressively reduces the spatial dimensions of the input image through a series of convolutional and pooling layers, extracting and condensing feature information. This process results in a compressed representation of the input, highlighting essential features while reducing data redundancy.

Decoder: The decoder's role is to reconstruct the feature information captured by the encoder back to the original image dimensions. It progressively increases the spatial resolution of the encoded features through upsampling or transposed convolution layers. The decoder utilizes the condensed features to produce a dense prediction map, where each pixel is assigned a class label. The architecture may also include skip connections from the encoder to the decoder, allowing the decoder to leverage both high-level semantic information and low-level details to improve the accuracy of the segmentation.

Following the overview of how semantic segmentation architectures operate, we delve into specific architectures adapted for this task:

- **Fully Convolutional Neural Networks (*FCNNs*):** FCNNs marked a paradigm shift by adapting CNNs for pixel-wise predictions, making end-to-end training

and inference on images of any size possible. This adaptation involves converting FC layers to convolutional ones, enabling the network to output spatial maps instead of class scores [27].

- **U-Net:** Tailored for biomedical image segmentation, U-Net features a symmetric architecture that excels in tasks with limited data. Its distinct U-shaped design, incorporating skip connections, ensures precise localization by combining high-level context with detailed spatial information, making it highly effective for medical imaging analysis [28].
- **V-Net and 3D U-Net:** These architectures extend the principles of U-Net to 3D image data, addressing the unique challenges of volumetric segmentation. By adapting U-Net's efficient encoding and decoding pathways for three-dimensional inputs, they enable detailed segmentation of 3D medical scans [29, 30].

The development of these architectures has significantly advanced the field of semantic segmentation, enabling detailed and accurate pixel-level understanding of images across various applications, particularly in medical imaging where such precision is crucial for diagnosis and therapy.

2.2.3 Network Training

The training of a neural network is a process that transforms raw data into predictive insights. This transformation is achieved through a series of steps, forming a pipeline that encompasses data preparation, model training, and iterative optimization to refine the model's predictions. Here is a detailed look at each stage in the pipeline:

1. **Data Collection and Preprocessing:** The foundation of any neural network model is data. The first step involves gathering a comprehensive dataset relevant to the task at hand. Once collected, the data undergoes preprocessing, which may include normalization, augmentation, and division into training, validation, and test sets. This preparation is crucial for ensuring that the network learns from a well-structured and representative dataset.
2. **Model Architecture Definition:** Before training begins, the architecture of the neural network must be defined. This includes selecting the type of network (e.g., CNN for image-related tasks) and designing the layer structure, including the number of layers, the number of neurons in each layer, activation functions, and connectivity patterns. This architectural blueprint dictates how the data will flow through the network and how complex patterns can be learned.
3. **Forward Pass:** Training commences with the forward pass, where input data is fed into the network. As the data propagates through each layer, the network applies weights, biases, and activation functions to compute the output.
4. **Loss Calculation:** The output from the forward pass is evaluated against the actual target values using a loss function, which quantifies the difference between the predicted and true values. This loss reflects the current performance of the model; the goal of training is to minimize this value.
5. **Backpropagation:** With the loss computed, backpropagation calculates the gradient of the loss function with respect to each weight in the network. This process uses the chain rule to propagate the error backward from the output layer to the input layer, determining how each weight contributed to the error.

6. **Optimization:** Following backpropagation, an optimization algorithm updates the model's weights and biases to minimize the loss. This step is iterative, with the model undergoing numerous epochs of training, progressively refining its parameters to reduce error.
7. **Regularization and Hyperparameter Tuning:** Throughout the training, regularization methods are applied to prevent overfitting, ensuring the model generalizes well to new data. Concurrently, hyperparameter tuning is conducted to find the optimal settings for the network's architecture and training process, such as the learning rate, batch size, and number of epochs.
8. **Evaluation and Iteration:** After training, the model is evaluated on a separate validation (and possibly a test) set to assess its performance on unseen data. Metrics such as accuracy, and F1 score provide insight into the model's effectiveness (for classification tasks). Based on these results, further adjustments to the model architecture, training process, or data preprocessing might be made in an iterative process to enhance performance.
9. **Deployment:** Once the model achieves satisfactory performance, it is deployed for making predictions on new data in real-world applications. This step often involves integrating the model into an existing production environment, where it can provide insights, make decisions, or automate tasks based on its learned patterns.

Detailed descriptions of key training concepts such as backpropagation, loss functions, optimization strategies, and regularization techniques immediately follow in the subsequent subsections (i.e., Section 2.2.3.1 - Section 2.2.3.4).

2.2.3.1 Backpropagation

Backpropagation stands as a cornerstone of neural network training, instrumental in refining the network's parameters to minimize the loss function. This process employs the chain rule to compute the gradients of the loss function relative to the network's weights, facilitating a systematic update mechanism that enhances model accuracy.

The operation of backpropagation unfolds through two distinct stages: the forward pass and the backward pass. In the forward pass, the network computes outputs and the local derivatives at various nodes, setting the stage for the backward pass. Here, gradients are calculated in reverse, from the output layer back to the input layer, enabling the precise adjustment of weights in accordance with the computed gradients. These adjustments are governed by an optimization algorithm, detailed further in the Section 2.2.3.3, which methodically updates the weights to optimize the network's performance.

The application of the chain rule in backpropagation is pivotal for calculating the gradients with respect to the weights across all layers. For a given sequence of hidden layers $h_1, h_2, \ldots, h_k$ culminating in an output O, and considering the weight connection from layer h_r to h_{r+1} as $w_{(h_r,h_{r+1})}$, the gradient of the loss function with respect to any weight in the network is determined through a multivariable chain rule. This accounts for the multitude of paths from h_1 to O, computing the gradient as follows:

$$\frac{\partial L}{\partial w_{(h_{r-1},h_r)}} = \frac{\partial L}{\partial O} \cdot \left[\sum_{[h_r,h_{r+1},\ldots,h_k,O]\in P} \frac{\partial O}{\partial h_k} \prod_{i=r}^{k-1} \frac{\partial h_{i+1}}{\partial h_i} \right] \frac{\partial h_r}{\partial w_{(h_{r-1},h_r)}}, \tag{2.10}$$

where P represents the ensemble of all paths from h_r to O. This nuanced approach to gradient computation underpins the model's learning process, ensuring that each weight is optimally adjusted to minimize the loss and thus, enhance the model's predictive capability. Further exploration of this topic is available in [31].

2.2.3.2 Loss Functions

Loss functions, also known as cost functions, play a critical role in quantifying the discrepancy between the actual outcomes (y) and the model's predictions ($\hat{y}$). The primary objective of optimizing a neural network is to minimize this discrepancy, thereby enhancing the model's prediction accuracy. A smaller loss function value indicates a model that is better aligned with the ground truth, making loss function minimization a fundamental aspect of model training. Among the various loss functions employed, the following are particularly noteworthy:

- **Mean Squared Error (*MSE*)** for regression quantifies the average squared discrepancy between actual values and predictions:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2.11}$$

  where n is the total number of observations. MSE is sensitive to outliers due to squaring the error.
- **Mean Absolute Error (*MAE*)** also for regression, measures the average absolute difference:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{2.12}$$

  offering robustness to outliers compared to MSE.
- **Cross-Entropy Loss**, essential for classification tasks, it measures the performance of a classification model whose output is a probability value between 0 and 1. It increases as the predicted probability diverges from the actual label. It is particularly useful for models outputting probabilities and is defined for a multi-class classification problem with m classes as:

$$\text{Cross-Entropy Loss} = -\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}\log(\hat{y_{ij}}), \tag{2.13}$$

- **Dice Loss**, particularly useful for segmentation tasks, focuses on measuring the overlap between the predicted segmentation and the ground truth. It is especially favored in medical image segmentation due to its effectiveness in handling class imbalance:

$$\text{Dice Loss} = 1 - \frac{2}{m}\sum_{j=1}^{m}\left(\frac{\sum_{i=1}^{n}(y_{ij}\cdot\hat{y_{ij}}) + \epsilon}{\sum_{i=1}^{n} y_{ij} + \sum_{i=1}^{n}\hat{y_{ij}} + \epsilon}\right), \tag{2.14}$$

  where ϵ is a small constant added for numerical stability and m is the number of classes.

Each of these loss functions has its unique characteristics and applications, chosen based on the specific requirements of the task at hand, be it regression, classification, or segmentation. The selection of an appropriate loss function is pivotal in guiding the learning algorithm towards optimal model performance.

2.2.3.3 Optimization

Optimization algorithms play a crucial role in machine learning by iteratively reducing the loss function to enhance the predictive accuracy of models. This section outlines two fundamental optimization strategies commonly employed in model training.

Gradient Descent Gradient Descent is a foundational optimization technique that iteratively adjusts model parameters to minimize the loss function. The core principle involves calculating the gradient of the loss function with respect to the model's parameters, denoted by θ, and updating these parameters in the opposite direction of the gradient. This process is akin to descending a hill by moving in the direction of the steepest slope, where the magnitude of each step is governed by a hyperparameter known as the learning rate (η). The equation for parameter update in gradient descent is given by:

$$\theta^{\text{updated}} = \theta - \eta \nabla_\theta \text{Loss}(\theta), \tag{2.15}$$

where $\nabla_\theta \text{Loss}(\theta)$ represents the gradient of the loss function. The choice of learning rate is critical; too small a rate results in slow convergence, while too large a rate can lead to divergence or overshooting the minimum.

For large datasets, computing the gradient across all data points becomes computationally intensive. To address this, **Stochastic Gradient Descent (*SGD*)** is employed, which approximates the gradient using a randomly selected subset of data at each iteration. This approach accelerates the optimization process, albeit with increased variance in the loss function trajectory towards the minimum.

Adam Optimizer The Adam Optimizer, an acronym for Adaptive Moment Estimation [32], integrates the advantages of two other extensions of stochastic gradient descent: Momentum and RMSProp [33]. It computes exponentially decaying averages of past gradients, m_t, akin to momentum, and squared gradients, v_t, similar to RMSProp, to adaptively adjust the learning rates for each parameter, reducing the necessity for meticulous hyperparameter tuning (η). This capability makes Adam especially suited for handling sparse gradients on noisy problems. Adam's update rules at time step t are given by:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \nabla_\theta \text{Loss}(\theta_t), \tag{2.16}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) (\nabla_\theta \text{Loss}(\theta_t))^2, \tag{2.17}$$

where β_1 and β_2 are the decay rates that control the moving averages of the gradient and its square, respectively, typically close to 1.

To correct for the initial bias towards zero, Adam adjusts both m_t and v_t using bias-corrected first- and second-moment estimates:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \tag{2.18}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}, \tag{2.19}$$

Finally, the parameter update rule is:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t, \tag{2.20}$$

where η is the learning rate, and ϵ is a small scalar added to improve numerical stability. This demonstrates how Adam adjusts the learning rate for each parameter dynamically, based on estimates of the first- and second-moments of the gradients, making it an effective optimizer for a wide range of deep learning tasks.

2.2.3.4 Regularization

Deep neural networks, characterized by their extensive parameter counts, are inherently prone to overfitting. Overfitting occurs when a model memorizes the training data to the extent that its generalization capability on unseen data is compromised. To mitigate this, regularization techniques are necessary in model training.

Early Stopping A prevalent indicator of overfitting is the divergence of training and validation losses: while the training loss continues to decrease, the validation loss begins to deteriorate. Early stopping addresses this by halting the training process as soon as the validation performance starts to decline, thereby preventing the model from overfitting.

l_1 and l_2 Regularization To further combat overfitting, l_1 and l_2 regularization techniques are integrated into the cost function during training. These techniques adjust the cost function by adding a penalty term, which is proportional to the size of the coefficients:

l_2 regularization, or Ridge regularization, discourages large weights through the penalty term $\alpha \sum_{i=1}^{n} \theta_i^2$, effectively keeping the model weights as small as possible. Here, α signifies the regularization strength. This method is known for its ability to handle multicollinearity, reduce model complexity, and enhance the model's interpretation by shrinking parameters.

Conversely, l_1 regularization, or Lasso regularization, adds a penalty equivalent to the absolute value of the magnitude of coefficients: $\alpha \sum_{i=1}^{n} |\theta_i|$. This form of regularization can lead to sparse models where certain weights can become zero, thus performing feature selection by eliminating non-informative features from the model.

Dropout Among the most effective regularization strategies in deep learning is dropout [34]. This technique involves randomly omitting a subset of neurons at each step of the training process with a probability p, known as the dropout rate. This randomness ensures that no single neuron becomes overly dependent on its input neurons, thus preventing co-adaptation and fostering a model that is robust to minor variations in input data. It's crucial to note that dropout is only applied during training, not during model evaluation or inference.

These regularization methods are fundamental in preventing overfitting, enabling neural networks to generalize better to unseen data. By incorporating such techniques, models can achieve higher accuracy and robustness, which is crucial for deploying reliable machine-learning solutions.

2.3 Medical Background

The medical imaging techniques developed in this thesis were applied and evaluated in three distinct diseases, namely Autosomal Dominant Polycystic Kidney Disease (*ADPKD*), Multiple Sclerosis (*MS*), and Aortic Dissection (*AD*). They are briefly introduced in the following:

2.3.1 ADPKD

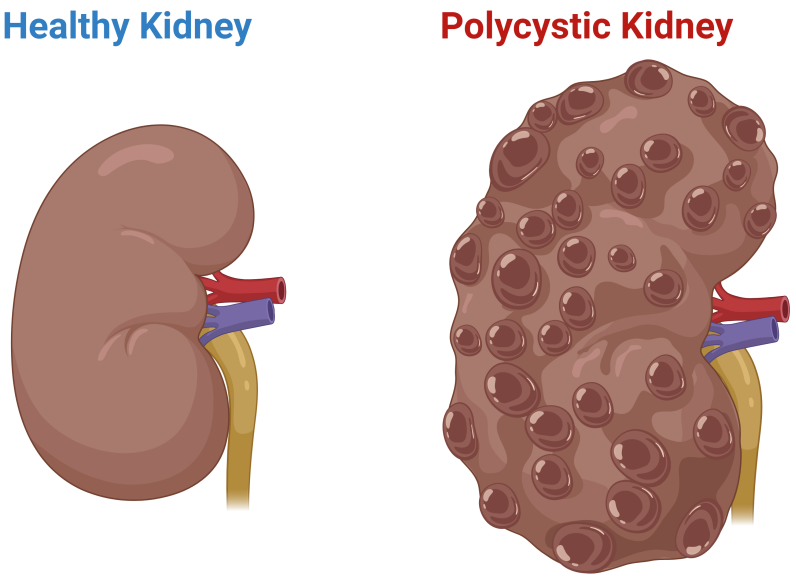


Figure 2.6: An illustration showing a kidney with cysts in comparison to a healthy kidney [35].

ADPKD is a genetic disorder primarily marked by the growth of numerous cysts in the kidneys. These cysts may lead to significant enlargement and dysfunction of the kidneys and can affect other organs including the liver, the pancreas, and the brain. The disease is prevalent (between one in 1000 and one in 2500 individuals) worldwide and is equally likely to affect individuals of any race or gender [36, 37].

ADPKD typically results from genetic mutations in the PKD1 and PKD2 genes, which are critical for producing the proteins polycystin 1 and polycystin 2. These

proteins play essential roles in the structural and functional integrity of kidney cells and other organ systems. Mutations disrupt these functions, initiating cyst development potentially as early as during fetal development, though symptoms may not manifest until later decades of life [36]. Furthermore, over time, these cysts expand, disrupting normal kidney morphology and function. An illustration of a cystic kidney is provided in Figure 2.6.

Individuals with ADPKD commonly experience hypertension, hematuria, and pain due to the expanding cysts. Diagnostic procedures primarily involve imaging techniques. Ultrasounds are typically used to detect larger cysts (diameter > 1 cm), while MRI or CT scans provide more detailed images necessary for identifying smaller cysts and assessing disease progression [36]. The disease progression varies, but it can severely impact life quality through complications like chronic kidney disease, necessitating significant medical interventions such as dialysis or kidney transplantation in severe cases. While there is currently no cure for ADPKD, treatment strategies are aimed at symptom management and decelerating disease progression. This includes the management of hypertension, pain relief, and the treatment of associated complications.

2.3.2 MS

MS is an autoimmune disorder of the Central Nervous System (*CNS*), characterized by inflammatory demyelination and axonal transection. This condition leads to severe neurological damage. MS typically manifests in young adults aged 20 to 30 years, significantly impacting physical function, cognition, quality of life, and employment capabilities [38].

Globally, the prevalence of MS varies widely, ranging from 5 to 300 per 100,000 individuals, with higher rates observed in regions farther from the equator. This geographical variation suggests a link between sunlight exposure, vitamin D levels, and disease prevalence. Women are affected nearly three times more often than men, indicating possible gender-specific genetic or hormonal factors influencing disease susceptibility [38].

MS involves the formation of lesions within the CNS, particularly in the white matter, although gray matter and cortical lesions are also prevalent. The disease process includes phases of active inflammation and myelin destruction (Figure 2.7) followed by periods where the disease may be less active or in remission. The involvement of both T cells and B cells in the disease development suggests a complex immune-mediated mechanism underlying MS [38]. Clinical presentation often includes unilateral optic neuritis, partial myelitis, and various brainstem syndromes, developing acutely or sub-acutely. The disease progresses in most patients from a relapsing-remitting course to a more steady decline in function, known as secondary progressive MS (SPMS) [38]. Meanwhile, the diagnosis of the disease relies on the 2017 McDonald Criteria, which requires demonstration of lesion dissemination in both space and time within the CNS, confirmed through clinical assessments and supported by MRI findings and cerebrospinal fluid analysis for oligoclonal bands. Furthermore, the management of MS includes a combination of disease-modifying therapies (DMTs), symptomatic treatments, and rehabilitative strategies. Since the

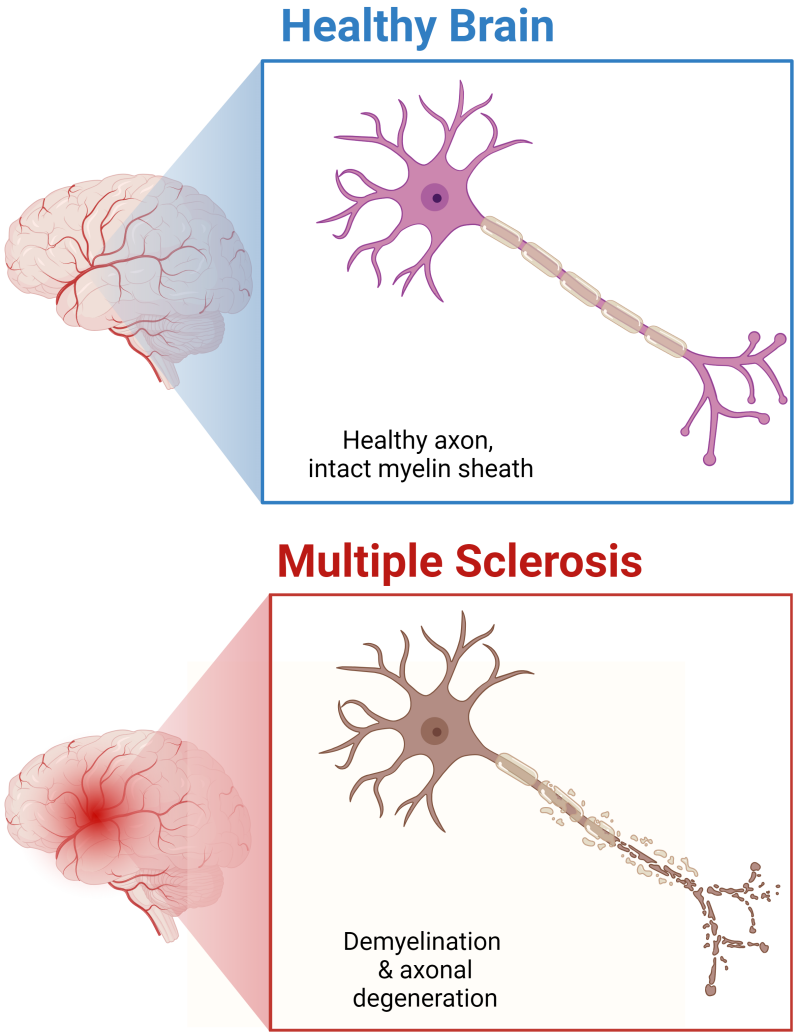


Figure 2.7: Depiction of a nerve cell affected by MS in comparison to a healthy nerve cell [35].

approval of the first DMT in 1993, multiple classes of DMTs have become available, each targeting different aspects of the immune system to reduce the frequency of relapses and slow disease progression.

2.3.3 AD

AD is a severe condition characterized by the tearing of the intima (inner layer) of the aorta, allowing blood to enter the media (middle layer) and create a false lumen (Figure 2.8). This process can cause the true and false lumens to separate, posing a significant risk of rupture and life-threatening complications. It primarily affects individuals aged 65–75, with an incidence rate that may be as high as 35 cases per 100,000 people annually in this age group [39]. Risk factors include hypertension, dyslipidemia, and certain genetic conditions like Marfan syndrome, which affect the connective tissues.

The disease process begins typically with an intimal tear, although in some cases, rupture of the vasa vasorum (small vessels supplying the aorta) may initiate the dissection. The progression can extend both antegrade and retrograde from the initial

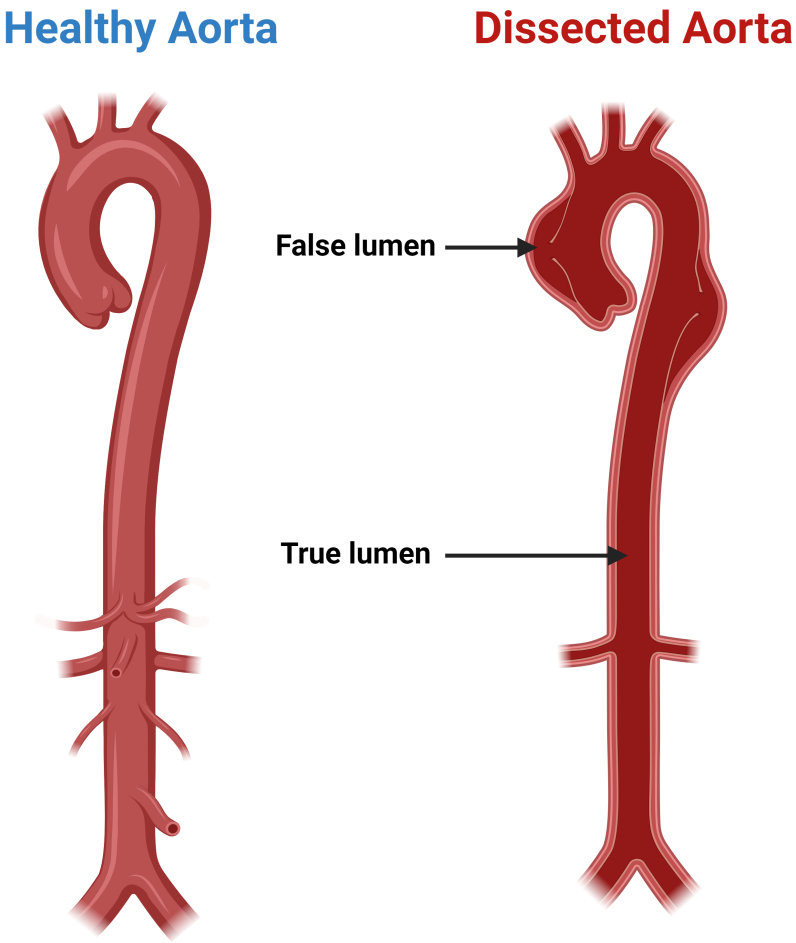


Figure 2.8: An illustration of the healthy and dissected aorta with a true and false lumen [35].

tear site, potentially involving major arterial branches and leading to varied clinical manifestations depending on the extent and location of the dissection. Symptoms of AD can include sudden severe chest or back pain, fainting, shortness of breath, and symptoms of stroke or other organ impairment due to disrupted blood flow [39].

Diagnosis is based on imaging studies such as CT scans, MRI, or ultrasound, which can identify the presence of a dissection and help differentiate the true from the false lumen. The Stanford classification, which is widely used, categorizes dissections into Type A (involving the ascending aorta) and Type B (restricted to the descending aorta), each with distinct management strategies. Immediate management depends on the type of dissection; Type A usually requires surgical intervention, while Type B might be managed with medication or endovascular techniques depending on the specifics of the case. Long-term management involves strict blood pressure control and surveillance imaging to monitor the progression or resolution of the dissection [39].

3. Deep Learning-Based Total Kidney Volume Segmentation in Autosomal Dominant Polycystic Kidney Disease Using Attention, Cosine Loss, and Sharpness Aware Minimization

Anish Raj [1], Fabian Tollens [2], Laura Hansen [1], Alena-Kathrin Golla [1], Lothar R. Schad [1], Dominik Nörenberg [2] and Frank. G. Zöllner [1]

[1]Computer Assisted Clinical Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany

[2]Department of Radiology and Nuclear Medicine, University Medical Centre Mannheim, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany

3.1 Abstract

Early detection of the Autosomal Dominant Polycystic Kidney Disease (*ADPKD*) is crucial as it is one of the most common causes of end-stage renal disease (ESRD) and kidney failure. The Total Kidney Volume (*TKV*) can be used as a biomarker to quantify disease progression. The TKV calculation requires accurate delineation of kidney volumes, which is usually performed manually by an expert physician. However, this is time-consuming and automated segmentation is warranted. Furthermore, the scarcity of large annotated datasets hinders the development of deep learning solutions. In this work, we address this problem by implementing three attention mechanisms into the U-Net to improve TKV estimation. Additionally, we implement a cosine loss function that works well on image classification tasks with small datasets. Lastly, we apply a technique called Sharpness Aware Minimization (*SAM*) that helps improve the generalizability of networks. Our results show significant improvements (p-value < 0.05) over the reference kidney segmentation U-Net. We show that the attention mechanisms and/or the cosine loss with SAM can achieve a Dice Similarity Coefficient (*DSC*) of 0.918, a Mean Symmetric Surface Distance (*MSSD*) of 1.20 mm with the mean TKV difference of -1.72%, and R^2 of 0.96 while using only 100 Magnetic Resonance Imaging (*MRI*) datasets for training and testing. Furthermore, we tested four ensembles and obtained improvements over the best individual network, achieving a DSC and MSSD of 0.922 and 1.09 mm, respectively.

3.2 Introduction

ADPKD is a hereditary disorder with a slow and gradual development of cysts in the kidneys. ADPKD leads to renal enlargement and eventually to end-stage

renal disease (ESRD) with renal failure [40, 41]. Daalgard et al. [42] reported that ESRD might occur within five years after detecting ADPKD. Hence, it is important to monitor ADPKD progression in patients. The TKV increases with ADPKD progression and, therefore, can be used as a risk predictor of disease development [43] and to quantify disease progression [44].

Consequently, the Food and Drug Administration (FDA) has accepted TKV as an important biomarker [45] for determining renal function in patients. Moreover, the TKV is the only MRI-established biomarker so far [46] and can be determined by segmenting the kidneys from the MRI volumes. Manual segmentation by an experienced physician is, however, time-consuming and prone to observer variability [47]. Alternatively, deep learning approaches are generally much faster and have recently achieved state-of-the-art results in medical imaging [48].

In the past, kidney segmentation has been approached via classical image processing techniques, such as algorithmic segmentation methods, model-based segmentation methods, and their combinations. An overview of related work is given in two reviews by Zöllner et al. [46, 47]. However, only a few machine learning (especially deep learning approaches) for kidney segmentation have recently been proposed.

Kline et al. [49] approached kidney segmentation of ADPKD patients with 2000 training and 400 test cases. They employed multi-observer artificial neural networks consisting of 11 Convolutional Neural Networks (*CNNs*) with variable depths and parameters. They post-processed the prediction map using the two largest connected components and then applied active contours and edge detection to finalize the segmentation. The resulting average DSC was 0.97. Based on this work, van Gastel et al. [50] used T2-weighted MR images from 440 ADPKD patients for training and extended the method to also include liver segmentation. They added inception blocks and residual connections to the network in [49], obtaining a DSC of 0.96. In another approach, Bevilacqua et al. [51] implemented R-CNN to first determine the Region of Interest (*ROI*) containing kidneys and then a semantic segmentation CNN to delineate the kidneys. Their approach reached a DSC of 0.88 with a dataset of 57 images from four patients. Mu et al. [52] employed a multi-resolution method using a modified V-Net [29] to segment ADPKD kidneys in 305 patients. The resulting DSC reached 0.95. Daniel et al. [53] developed an automated system using the U-Net [28] to segment kidneys and ultimately determine TKV for renal disease detection. They used T2-weighted MR images from 30 healthy and 30 Chronic Kidney Disease (*CKD*) patients. Their system achieved a DSC of 0.93 on a test dataset consisting of 10 patients.

The drawback of such deep learning approaches is that they require huge amounts of data to train the networks to achieve high (segmentation) accuracy. This problem is further escalated in the medical sector where image data are rare in most cases. The reason is that acquiring large medical datasets is hindered by the complexity and high cost of large-scale experiments or in the case of a rare disease, a limited number of patients. Additionally, a class imbalance between the background and the segmented object exists. To mitigate this problem, data augmentation techniques [54] or approaches to generate synthetic data are proposed [55, 56].

Our contribution in this work is to address the problem of a small dataset by introducing a cosine loss function, which, to the best of our knowledge, has not been implemented so far for medical image segmentation tasks. Moreover, we integrate SAM [57] with the loss function to improve model generalizability. We further investigate and incorporate three attention mechanisms [58–62] with CNNs so that the networks can focus on relevant image regions. As a final experiment, we explore four ensembles that consist of different attention networks and loss functions with SAM for automatic kidney volume segmentation in patients with ADPKD.

3.3 Materials and Methods

3.3.1 Image Data

The patient image data were obtained from the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK), National Institutes of Health, USA, and were recorded in the Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) study [41]. It contains T1- and T2-weighted MRI scans of patients with different stages of CKD. For this work, we retrieved 100 datasets from the NIDDK database. The male-to-female ratio is 50:50 with an average age of 30 ± 10 years. It includes patients from healthy (CKD stage 1) to ADPKD (CKD stage 3) cases. The number of cases belonging to CKD stages 1, 2, and 3 is 41, 41, and 18, respectively. We only focused on T1-weighted MRIs. Images were recorded with a matrix of 256×256 and 30–80 slices with an in-plane resolution of 1.41×1.41 (± 0.13) mm^2 and slice thickness of 3.06 (± 0.29) mm. Images were recorded in the coronal orientation. More details of the imaging protocol can be found in the original study in [41].

3.3.2 Image Annotation

For each dataset, left and right kidneys including cysts, were segmented manually as a reference standard. Two experienced physicians independently performed segmentation on coronal MR images using an in-house developed annotation tool based on MeVisLab SDK (MeVis Medical Solutions, Inc., Bremen, Germany), which also allowed for an analysis of the inter-user agreement of kidney segmentations. The mean inter-user agreement was found to be 0.91 ± 0.06 (Dice) with a coefficient of variation of 0.07.

3.3.3 Pre-Processing

We first normalized the images using (Equation 5.1),

$$\hat{I} = \frac{I - \mu(I)}{\sigma(I)}, \tag{3.1}$$

where $\mu(I)$ and $\sigma(I)$ are the mean and standard deviations, respectively, of the original image I. $\hat{I}$ is the normalized image. Furthermore, as a data augmentation technique, we used a constrained label sample mining approach where patches were extracted from MRI slices with patch center probability of 50:50 on the label:background pixel for each batch [63]. We trained the networks on patches of size 96×96 and 128×128. For testing, we used the whole image size of 256×256. The pre-processing was implemented using SimpleITK 1.2.4 [64].

3.3.4 Attention Module

Oktay et al. [58] proposed attention gates in the U-Net to guide the network in selecting relevant features and disregard irrelevant ones by using higher-level features as a guide to suppress trivial and noisy responses in the lower-level skip connections. Figure 3.1 illustrates this attention module. The gating signal g is the higher-

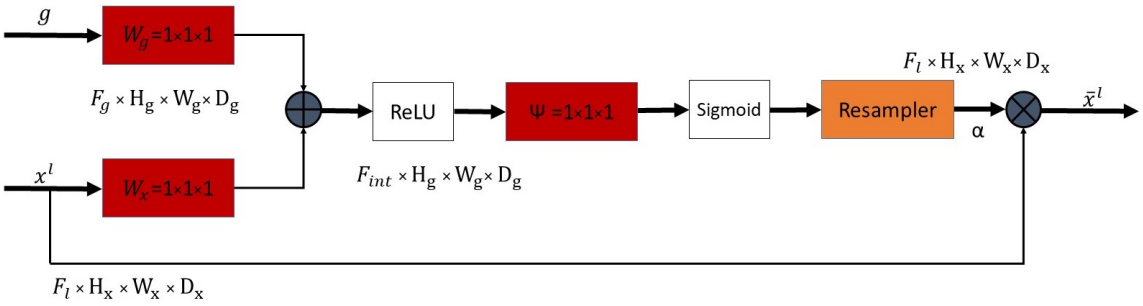


Figure 3.1: Attention module used in the Attention U-Net. It takes two inputs: g and x^l where g is the higher-level feature and spatially smaller than the previous layer feature x^l. g is used to guide lower-level features x^l to emphasize relevant image regions. This is achieved by calculating α coefficients that are element-wise multiplied with x^l to produce attention maps. Adapted from [58].

level feature and x^l is the corresponding previous layer skip connection. Since g is spatially smaller and has more feature maps than x^l, it is convoluted with a 1 $\times$ 1 filter to obtain the same number of features. Afterward, g is upsampled to have the same spatial dimensions as x^l. Then, both g and x^l are concatenated followed by Rectified Linear Unit (*ReLU*), 1 $\times$ 1 convolution, sigmoid activation, and eventually an upsampling layer resulting in the attention coefficients α. The attention coefficients are then multiplied with the skip connection to obtain the final attention feature maps.

3.3.4.1 Convolutional Block Attention Module

The Convolution Block Attention Module (*CBAM*) calculates and combines both spatial and channel attention into one network [60]. Briefly, a 1D channel attention map and thereafter a 2D spatial attention map are computed. The spatial attention map is generated by performing max and average-pooling operations along the channel dimension of the input feature map. The pooled feature maps are concatenated and fed forward through a convolution layer to yield a 2D spatial attention map. There are two versions of channel attention: 1) squeeze and excitation attention proposed by Hu et al. [59] and 2) channel attention, as outlined in [60]:

3.3.4.2 Squeeze and Excitation Attention

This attention mechanism focuses on channel relationships in a network. Briefly, the attention module performs two operations sequentially to form a Squeeze and Excitation (*SE*) block. First, a squeeze operation on the input layer is executed, which is then followed by an excitation operation. The squeeze operator does a global average-pooling of spatial information into a channel descriptor of size 1 $\times$ 1

× C, where C is the number of channels in the input layer. This output vector is then processed by an excitation operation that captures full channel-wise dependencies [59]. The following equation describes the exact excitation operation,

$$E_x = \phi_{sig}(W_2\phi_{ReLU}(W_1 z)), \tag{3.2}$$

where z is the output from the squeeze operation, ϕ_{ReLU} the ReLU activation, W_1 & W_2 are two Fully Connected (*FC*) layers, and ϕ_{sig} is the sigmoid activation. The final output is obtained by multiplication of E_x and the input layer. The squeeze and excitation operation is depicted in Figure 3.2.

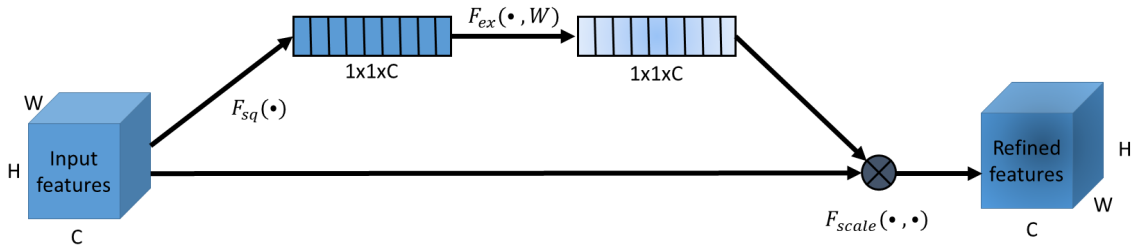


Figure 3.2: Squeeze and excitation module. The module first squeezes the input features to a vector with the shape 1xC (C: number of channels in the input features). This is followed by the excitation operation where a multi-layer perceptron processes the input vector. The final output feature map is calculated by multiplication between the output vector and the input feature map. Adapted with permission from Ref. [59]. 2018, IEEE

.

3.3.4.3 Channel Attention in CBAM

The CBAM's channel attention [60] is calculated by squeezing the spatial dimension of the input feature map using average and max-pooling and then feeding them forward through a small Multi-Layer-Perceptron (*MLP*).

The channel and spatial attention modules used in CBAM are illustrated in Figure 3.3a. The complete CBAM block is shown in Figure 3.3b. In our experiments, we incorporate SE and CBAM blocks in the Attention U-Net [58] encoder to implement the SE U-Net and CBAM U-Net, respectively.

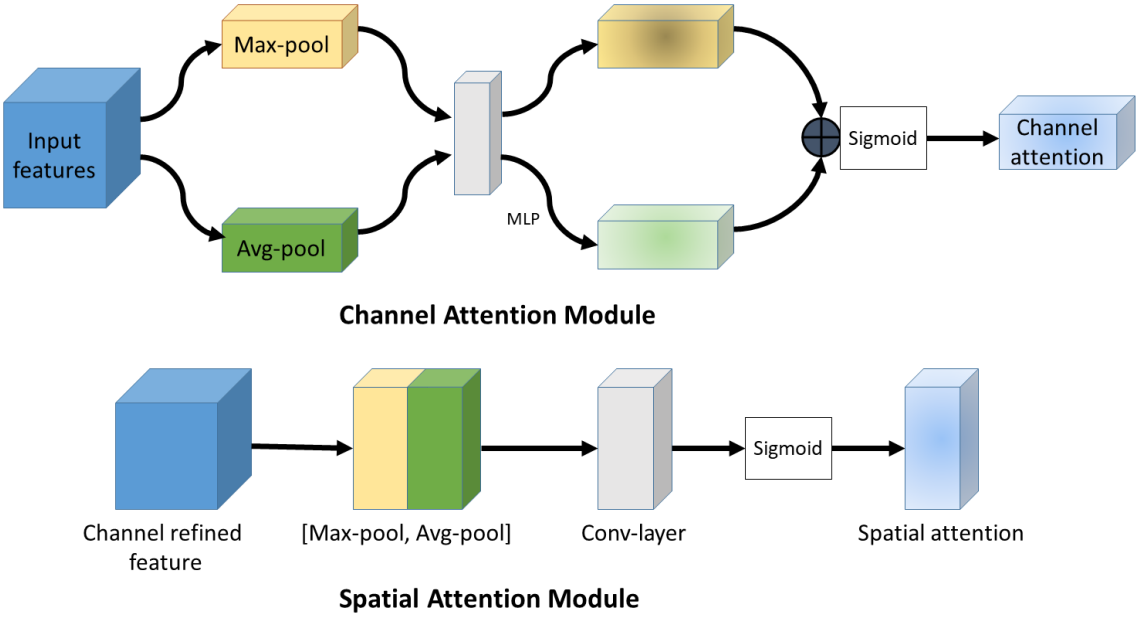
3.3.5 Cosine Loss

The cosine loss has been shown to improve the image classification accuracy for small datasets [65]. Hence, we adapt this loss function for our kidney segmentation task. The loss is given by,

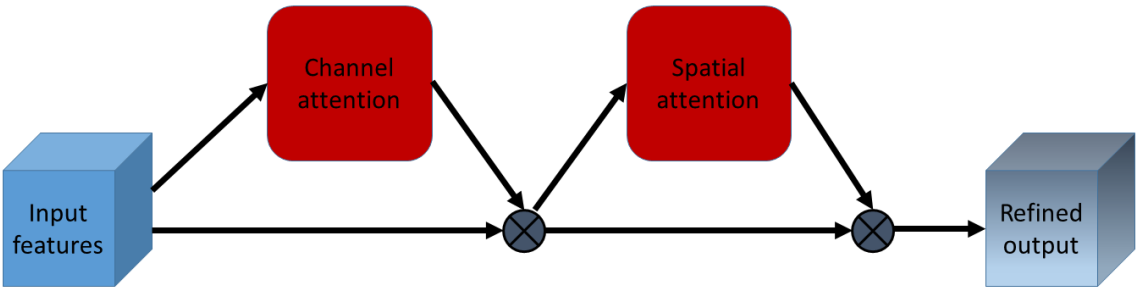
$$S(\hat{Y}, Y) = \frac{\langle \hat{Y}, Y \rangle}{\|\hat{Y}\|_2 \cdot \|Y\|_2}, \tag{3.3}$$

$$\mathcal{L}_{COS}(\hat{Y}, Y) = 1 - S(\hat{Y}, Y), \tag{3.4}$$

where S and $\mathcal{L}_{COS}$ are the cosine similarity and cosine loss, respectively, between the prediction $\hat{Y}$ and the ground truth Y.



(a) Channel and spatial attention blocks in CBAM



(b) CBAM

Figure 3.3: The scheme of the convolutional block attention module. (**a**) The channel attention module implements max and average pooling along the channel dimension. It produces a refined 1D channel attention vector. Meanwhile, the spatial attention module applies max and average pooling along spatial dimensions to produce a 2D spatial attention map. (**b**) The CBAM combines the channel and spatial attention maps (red-colored squares) detailed in (**a**) and applies them to the input feature map sequentially to produce an output feature map, which is refined along both the channel and spatial dimensions. Adapted with permission from Ref. [60]. 2018, Springer Nature

.

3.3.6 Sharpness Aware Minimization

Foret et al. [57] introduced a technique called SAM that helps improve the generalizability of neural networks. Briefly, the method searches for a neighborhood of parameters with homogeneous low loss values, signifying a wide loss curve at the minimum, thereby, reducing the loss value and sharpness of the loss curve. A wide minimum suggests that the parameters in the neighborhood will generally yield consistently better predictions compared to a minimum with a sharp curve.

3.3.7 Networks

All networks implemented in this work are based on the 2D U-Net architecture by Ronneberger et al. [28] extended with residual connections. The baseline 2D U-Net architecture is illustrated in Figure 3.4a. For the baseline experiments, the 2D U-Net is used with DSC loss (cf. (Equation 3.5)) and a combination of DSC and cross-entropy (CE) loss.

$$\mathcal{L}_{DSC}\left(\hat{Y}, Y\right) = 1 - \frac{2 \cdot \sum_{c=1}^{3} \sum_{i=1}^{N} \hat{y}_{i,c} \cdot y_{i,c}}{\sum_{c=1}^{3} \sum_{i=1}^{N} \hat{y}_{i,c} + \sum_{c=1}^{3} \sum_{i=1}^{N} y_{i,c}} \tag{3.5}$$

The cross-entropy loss is given by,

$$\mathcal{L}_{CE}\left(\hat{Y}, Y\right) = -\frac{\sum_{c=1}^{3} \sum_{i=1}^{N} y_{i,c}\, log(\hat{y}_{i,c})}{3 \cdot N} \tag{3.6}$$

where $\hat{y}_{i,c}$ and $y_{i,c}$ correspond to the individual voxel probabilities and label, respectively, with c and N being the number of classes and voxels in a batch, respectively. (Equation 3.7) depicts the combination of $\mathcal{L}_{CE}$ and $\mathcal{L}_{DSC}$ where $\mathcal{L}_{CE}$ is weighted with $\lambda = 10$ [66].

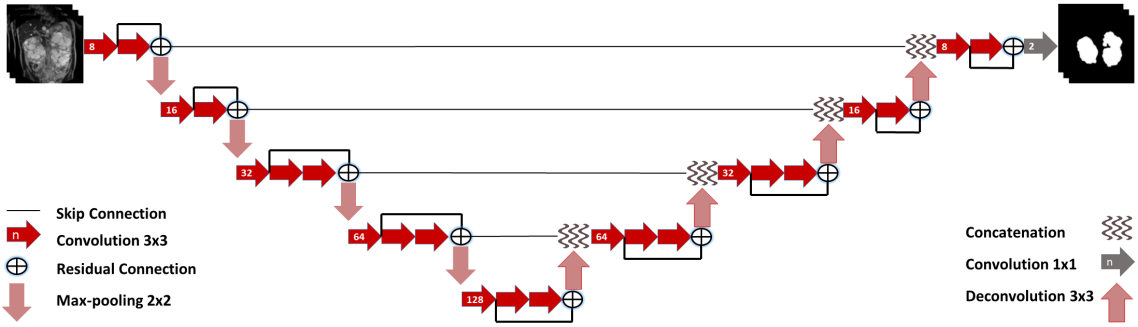
$$\mathcal{L}_{CE+DSC}\left(\hat{Y}, Y\right) = \mathcal{L}_{DSC}\left(\hat{Y}, Y\right) + \lambda\, \mathcal{L}_{CE}\left(\hat{Y}, Y\right) \tag{3.7}$$

The Attention U-Net as introduced in Section 3.3.4, is depicted in Figure 3.4b. The other variants that involve SE modules/CBAMs in the Attention U-Net encoder part are shown in Figure 3.4c. In preliminary experiments, we found that modifying baseline U-Net with only CBAMs or SE modules performed worse or similar to the baseline. Hence, for any further experiments, we combined them only with the Attention U-Net. All described networks were implemented using TensorFlow 2.0 and Python 3.7.

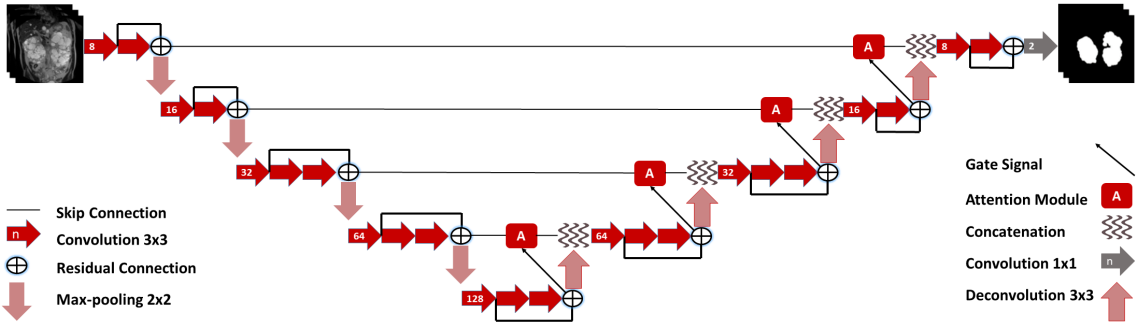
3.3.8 Training

We trained the networks on T1-weighted MR images and implemented a batch size of 16 and 8 for patch sizes 96 and 128, respectively, using the Adam optimizer [32] and a learning rate of 10^{-3}. We used exponential linear units (elus) [67] as activation functions with batch normalization, L2-regularization (10^{-7}) and dropout with probability of 0.01. Furthermore, we performed 5-fold cross-validation with a split of 70:10:20 patient image volumes in train:validation:test sets. We selected 160 samples per patient MRI volume during training.

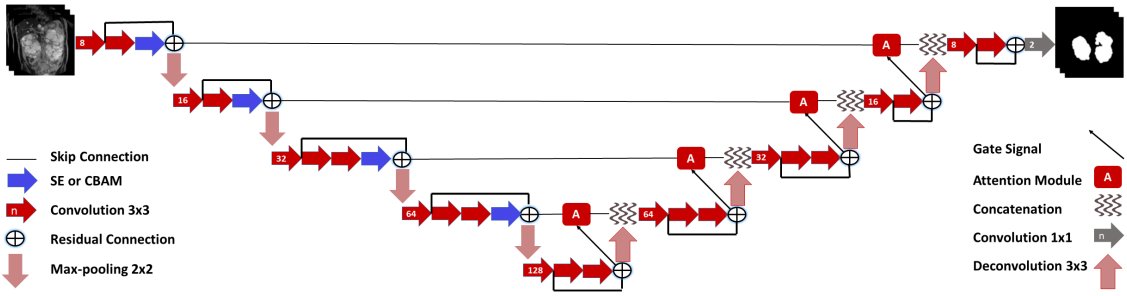
Each network was trained for at least 20 epochs. After that, the training stopped as soon as the difference in segmentation accuracy of each kidney in the validation data were less than 10^{-4} over the last ten epochs. We then selected the network's weights with the highest average accuracy on the validation data from these last ten epochs.



(a) Baseline U-Net



(b) Attention U-Net



(c) SE or CBAM Attention U-Net

Figure 3.4: The network architectures for kidney segmentation: (**a**) baseline U-Net without any attention modules, (**b**) U-Net with attention modules as described by Oktay et al. [58] and (**c**) U-Net combining SE or CBAM [59, 60] with attention modules from [58].

3.3.9 Ensembles

We created two ensembles from our proposed networks. The first one consists of four networks (SE U-Net, CBAM U-Net, Attention U-Net, and U-Net) trained using cosine loss with SAM ($\mathcal{L}_{COS} + \mathcal{SAM}$). The second ensemble consists of seven networks that include the four networks from the first ensemble plus three networks (SE U-Net, CBAM U-Net, and Attention U-Net) trained using cross-entropy + DSC loss with SAM ($\mathcal{L}_{\mathcal{CE}+\mathcal{DSC}} + \mathcal{SAM}$). We selected these networks to test all attention networks and loss functions. We employed two methods for calculating the ensemble result: 1) simultaneous truth and performance level estimation (STAPLE) [68] and 2) majority voting.

3.3.10 Evaluation

To compare the proposed models, we used the DSC, the MSSD, and the TKV as evaluation metrics. Firstly, we used the DSC to assess the overlap between the ground truth Y and the segmentation $\hat{Y}$,

$$DSC\left(\hat{Y}, Y\right) = \frac{2\left|\hat{Y} \cap Y\right|}{|\hat{Y}| + |Y|} \tag{3.8}$$

Secondly, we employed the MSSD (in mm) that is more perceptive to alignment and shape:

$$MSSD\left(\hat{Y}, Y\right) = \frac{\sum_{\hat{y}\in\hat{Y}} \min d(\hat{y}, Y) + \sum_{y\in Y} \min d(y, \hat{Y})}{|\hat{Y}| + |Y|} \tag{3.9}$$

Finally, we calculated the TKV (in mL) by multiplying the number of voxels belonging to the segmented kidneys by their voxel volume (mm^3) divided by 1000 to convert the result to mL.

We compared the TKV of the manual and the obtained segmented kidneys of our networks using scatter plots and the coefficient of determination (R^2).

Furthermore, we used a paired t-test to check for the significance between the results from the baseline and our implemented methods. Here, the null hypothesis is that the baseline network configuration is better than the developed methods for the given evaluation metric. It is rejected at $p < 0.05$.

3.4 Results

The quantitative results for all experiments are displayed in Table 3.1 where the DSC and MSSD values are averaged over both the kidneys for better comprehension. For the baseline U-Net (Figure 3.4a), the DSC loss ($\mathcal{L}_{DSC}$) performs worse for all metrics and patch sizes than the combination of cross-entropy and DSC loss ($\mathcal{L}_{CE+DSC}$). Furthermore, all proposed methods perform better than the baseline. Post-processing employing the largest connected components further improves the DSC and MSSD. In most cases, improvements are significant (see Table 3.2). The best results among individual networks were obtained using the U-Net with $\mathcal{L}_{COS} + SAM$ and patch size of 128. Here, an average DSC of 0.918 ± 0.044 and an MSSD of 1.199 ± 1.525 mm were achieved. The segmentation quality was further improved using ensembles. The DSC and MSSD of 0.922 ± 0.047 and 1.094 ± 1.376 mm, respectively, were achieved using the ensemble of seven models with the majority voting scheme. Supplement Table 11.1 displays results for the left and right kidneys.

Examples of obtained segmentation after post-processing for each network are displayed in Figure 3.5. The top row depicts a case with a DSC of 0.96 (stage: 1, female, age: 34), where the two networks (U-Net ($\mathcal{L}_{\mathcal{COS}}$) & Attention U-Net ($\mathcal{L}_{\mathcal{CE}+\mathcal{DSC}}$)) surpass the baseline only by a DSC of ≈ 0.002. The lower row depicts the obtained segmentation from a case with a high load of cysts that are distributed not only in the kidneys but all over the abdomen (stage: 2, female, age: 32). We observe that

Architecture	Loss	DSC ↑ 96 × 96	MSSD (mm) ↓ 96 × 96	DSC ↑ 128 × 128	MSSD (mm) ↓ 128 × 128
Baseline U-Net	$\mathcal{L}_{DSC}$	0.789 ± 0.109	8.803 ± 5.947	0.855 ± 0.079	5.435 ± 5.257
	$\mathcal{L}_{CE+DSC}$	0.865 ± 0.075	4.229 ± 4.094	0.888 ± 0.067	3.297 ± 3.951
SE U-Net	$\mathcal{L}_{CE+DSC}$	**0.889 ± 0.060**	**3.199 ± 3.941**	0.892 ± 0.060	2.805 ± 3.087
	$\mathcal{L}_{COS}$	**0.879 ± 0.080**	3.566 ± 4.164	0.892 ± 0.057	**2.654 ± 2.843**
	$\mathcal{L}_{CE+DSC} + SAM$	**0.895 ± 0.060**	**2.357 ± 2.790**	**0.903 ± 0.052**	**2.248 ± 2.719**
	$\mathcal{L}_{COS} + SAM$	**0.902 ± 0.055**	**2.228 ± 3.045**	**0.899 ± 0.058**	**2.450 ± 3.272**
CBAM U-Net	$\mathcal{L}_{CE+DSC}$	**0.878 ± 0.080**	3.517 ± 4.327	**0.899 ± 0.056**	**2.490 ± 3.013**
	$\mathcal{L}_{COS}$	**0.894 ± 0.056**	**2.636 ± 2.648**	**0.898 ± 0.057**	**2.332 ± 2.568**
	$\mathcal{L}_{CE+DSC} + SAM$	**0.880 ± 0.064**	3.689 ± 4.120	**0.903 ± 0.060**	**2.549 ± 4.997**
	$\mathcal{L}_{COS} + SAM$	**0.885 ± 0.073**	3.520 ± 5.274	**0.902 ± 0.056**	**2.090 ± 2.683**
Attention U-Net	$\mathcal{L}_{CE+DSC}$	**0.882 ± 0.069**	**3.331 ± 3.681**	**0.898 ± 0.057**	3.158 ± 4.191
	$\mathcal{L}_{COS}$	**0.892 ± 0.060**	**3.382 ± 4.790**	**0.901 ± 0.061**	**2.427 ± 3.250**
	$\mathcal{L}_{CE+DSC} + SAM$	**0.886 ± 0.068**	**3.144 ± 4.266**	**0.903 ± 0.054**	**2.461 ± 2.717**
	$\mathcal{L}_{COS} + SAM$	**0.896 ± 0.060**	**2.621 ± 3.270**	**0.907 ± 0.057**	**2.338 ± 3.987**
U-Net	$\mathcal{L}_{COS}$	**0.885 ± 0.069**	**3.007 ± 3.317**	**0.902 ± 0.061**	**2.228 ± 2.856**
	$\mathcal{L}_{COS} + SAM$	**0.899 ± 0.056**	**2.754 ± 3.541**	**0.909 ± 0.049**	**2.417 ± 3.542**
Ensemble-4-STAPLE	$\mathcal{L}_{COS} + SAM$	**0.904 ± 0.058**	**2.479 ± 3.621**	***0.913 ± 0.052***	***1.967 ± 2.841***
Ensemble-7-STAPLE	$\mathcal{L}_{CE+DSC} + SAM + \mathcal{L}_{COS} + SAM$	**0.903 ± 0.059**	**2.472 ± 3.575**	***0.916 ± 0.052***	***1.732 ± 2.467***
Ensemble-4-VOTING	$\mathcal{L}_{COS} + SAM$	***0.910 ± 0.051***	***1.886 ± 2.615***	**0.914 ± 0.049**	***1.506 ± 2.018***
Ensemble-7-VOTING	$\mathcal{L}_{CE+DSC} + SAM + \mathcal{L}_{COS} + SAM$	***0.910 ± 0.051***	***1.934 ± 2.690***	***0.918 ± 0.048***	***1.484 ± 2.083***

Table 3.1: The DSC and MSSD (in mm) values for various networks with loss functions as DSC ($\mathcal{L}_{DSC}$), cross-entropy+DSC ($\mathcal{L}_{CE+DSC}$), and cosine loss ($\mathcal{L}_{COS}$). The experiments were performed for two patch sizes: 96 and 128, with the numbers in bold implying significant difference (p-value < 0.05) between the baselines and the corresponding network configuration. The underlined numbers signify the best in the respective category excluding ensemble results. As can be seen, adding attention to the U-Net can improve the results significantly (Attention and CBAM U-Nets). Furthermore, cosine loss alone (U-Net) or with Attention and CBAM U-Nets provides better DSC than the corresponding cross-entropy+DSC loss networks. Finally, the ensembles outperform the best model in each category. The ensembles in italics imply significantly better results (p-value < 0.05) than the corresponding best performing model.

every network has difficulties obtaining a DSC > 0.80. Nonetheless, the proposed networks outperform the baseline U-Net by up to 18% (U-Net ($\mathcal{L}_{COS} + \mathcal{SAM}$)).

Investigating the cases with DSC ≤ 0.80, we observe that our proposed networks can reduce the number of such cases by half: 13 for the baseline U-Net ($\mathcal{L}_{CE+DSC}$) versus 5 for the other networks. It is also worth noting that these 5 cases were among the 13 cases with low DSC of the baseline U-Net. Moreover, the ensemble with 7 networks reduces this number to only 3 cases.

3.4.1 Attention Mechanisms

The networks SE U-Net, CBAM U-Net, and the Attention U-Net with $\mathcal{L}_{CE+DSC}$ as loss function all outperform the baseline results across all metrics (Table 3.1). For CBAM and Attention U-Net with $\mathcal{L}_{CE+DSC}$, the results are significantly better (p-value < 0.05) than the baseline U-Net ($\mathcal{L}_{CE+DSC}$) for 3/4 metrics (both the DSC scores and one MSSD value). In this configuration, the SE U-Net outperforms the other two for patch size of 96, however, for patch size of 128, CBAM and Attention U-Net perform better in terms of the DSC (0.899 and 0.898, respectively).

Architecture	Loss	DSC ↑ 96 × 96	MSSD (mm) ↓ 96 × 96	DSC ↑ 128 × 128	MSSD (mm) ↓ 128 × 128
Baseline U-Net	$\mathcal{L}_{CE+DSC}$	0.880 ± 0.071	2.382 ± 2.950	0.902 ± 0.600	1.530 ± 2.542
SE U-Net	$\mathcal{L}_{CE+DSC}$	**0.897 ± 0.058**	**1.687 ± 1.970**	0.899 ± 0.055	1.630 ± 2.054
	$\mathcal{L}_{COS}$	0.884 ± 0.088	2.002 ± 3.000	0.899 ± 0.051	1.579 ± 1.791
	$\mathcal{L}_{CE+DSC} + SAM$	**0.894 ± 0.070**	**1.580 ± 1.941**	0.904 ± 0.057	1.290 ± 1.225
	$\mathcal{L}_{COS} + SAM$	**0.902 ± 0.060**	**1.438 ± 1.600**	0.900 ± 0.079	1.593 ± 2.912
CBAM U-Net	$\mathcal{L}_{CE+DSC}$	**0.890 ± 0.070**	**1.937 ± 2.620**	0.906 ± 0.052	1.395 ± 1.480
	$\mathcal{L}_{COS}$	**0.901 ± 0.057**	**1.367 ± 1.158**	0.903 ± 0.058	1.504 ± 1.863
	$\mathcal{L}_{CE+DSC} + SAM$	**0.892 ± 0.059**	**1.947 ± 2.228**	**0.910 ± 0.054**	1.294 ± 1.644
	$\mathcal{L}_{COS} + SAM$	**0.896 ± 0.065**	**1.926 ± 2.717**	**0.908 ± 0.054**	1.346 ± 1.705
Attention U-Net	$\mathcal{L}_{CE+DSC}$	0.891 ± 0.076	**1.794 ± 2.046**	**0.910 ± 0.042**	1.371 ± 1.297
	$\mathcal{L}_{COS}$	**0.904 ± 0.056**	**1.588 ± 1.912**	**0.910 ± 0.052**	1.377 ± 1.671
	$\mathcal{L}_{CE+DSC} + SAM$	**0.895 ± 0.065**	**1.922 ± 2.757**	**0.911 ± 0.051**	1.398 ± 1.852
	$\mathcal{L}_{COS} + SAM$	**0.905 ± 0.056**	**1.470 ± 1.713**	**0.913 ± 0.051**	1.312 ± 1.708
U-Net	$\mathcal{L}_{COS}$	**0.896 ± 0.074**	**1.812 ± 2.643**	**0.909 ± 0.057**	1.358 ± 1.970
	$\mathcal{L}_{COS} + SAM$	**0.908 ± 0.054**	**1.480 ± 1.885**	**0.918 ± 0.044**	**1.199 ± 1.525**
Ensemble-4-STAPLE	$\mathcal{L}_{COS} + SAM$	**0.909 ± 0.055**	**1.560 ± 2.087**	**0.919 ± 0.048**	**1.204 ± 1.508**
Ensemble-7-STAPLE	$\mathcal{L}_{CE+DSC} + SAM + \mathcal{L}_{COS} + SAM$	**0.909 ± 0.054**	**1.534 ± 1.881**	**0.921 ± 0.048**	**1.174 ± 1.528**
Ensemble-4-VOTING	$\mathcal{L}_{COS} + SAM$	***0.914 ± 0.053***	***1.204 ± 1.367***	**0.917 ± 0.048**	**1.125 ± 1.340**
Ensemble-7-VOTING	$\mathcal{L}_{CE+DSC} + SAM + \mathcal{L}_{COS} + SAM$	***0.914 ± 0.052***	***1.299 ± 1.620***	***0.922 ± 0.047***	**1.094 ± 1.376**

Table 3.2: Results after post-processing with the largest-connected components. The two evaluation metrics are the DSC and the MSSD with the bold values being significantly better (p-value < 0.05) than the corresponding baseline network. The underlined values represent the best outcomes in their respective category excluding ensemble results. Meanwhile, the highlighted values in yellow imply a significantly better (p-value < 0.05) score than the results from the corresponding networks without post-processing from Table 3.1. The ensembles in italics signify significantly better results (p-value < 0.05) than the corresponding best performing model.

Within the test set, five cases had a DSC ≤ 0.8. The average DSC of these five cases in baseline U-Net ($\mathcal{L}_{CE+DSC}$) is 0.638 ± 0.081, while for the Attention U-Net ($\mathcal{L}_{CE+DSC}$), this score rises by 6% to 0.704 ± 0.072.

3.4.2 Cosine Loss

Meanwhile, for the cosine loss ($\mathcal{L}_{COS}$), we again observe that the U-Net, the Attention U-Net, and the CBAM U-Net outperform the baselines significantly (p-value < 0.05) over all the patch sizes and metrics. We also find that U-Net and CBAM U-Net with $\mathcal{L}_{COS}$ provide the best DSC and MSSD values for patch sizes 128 and 96, respectively, among the networks without SAM. Furthermore, the Attention U-Net with $\mathcal{L}_{COS}$ surpasses its corresponding network with $\mathcal{L}_{CE+DSC}$ over both the DSC and one MSSD value (patch: 128).

For the Attention U-Net ($\mathcal{L}_{COS}$) an improvement of 3% in DSC is recorded (from DSC of 0.647 to 0.677) as compared to the baseline U-Net for the five cases with DSC ≤ 0.8.

3.4.3 SAM

The application of SAM on top of the networks yields the best performance overall for the individual networks. The SE U-Net with $\mathcal{L}_{COS} + SAM$ provides the best DSC (0.902) and MSSD (2.228 mm) values for patch size 96. Furthermore, the U-Net with $\mathcal{L}_{COS} + SAM$ yields the best DSC of 0.909 with patch size 128 among all

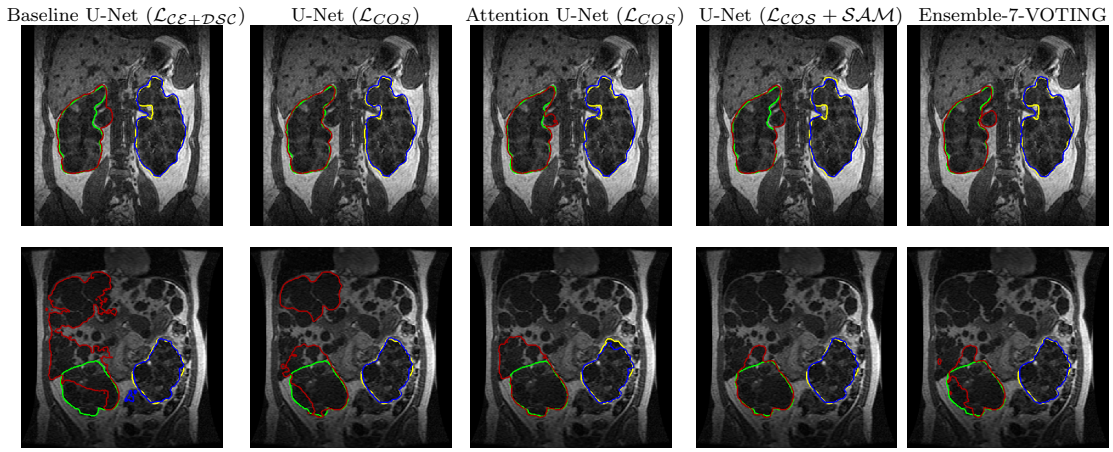


Figure 3.5: Qualitative results after post-processing (128 x 128): random slices of the best (top row) and the worst case segmentation (bottom row) from the baseline U-Net ($\mathcal{L}_{\mathcal{CE}+\mathcal{DSC}}$) compared to its corresponding U-Net ($\mathcal{L}_{\mathcal{COS}}$ & $\mathcal{L}_{\mathcal{COS}}+\mathcal{SAM}$), Attention U-Net ($\mathcal{L}_{\mathcal{CE}+\mathcal{DSC}}$), and Ensemble-7-VOTING predictions. The best case has stage 1 CKD, while, the worst case has stage 2 CKD. The ground truth segmentation is colored in green and yellow, while the network segmentations are colored in red and blue. Top row DSC (left to right): Baseline U-Net = 0.963, U-Net ($\mathcal{L}_{\mathcal{COS}}$) = 0.965, Attention U-Net ($\mathcal{L}_{\mathcal{CE}+\mathcal{DSC}}$) = 0.965, U-Net ($\mathcal{L}_{\mathcal{COS}}+\mathcal{SAM}$) = 0.962 and Ensemble-7-VOTING = 0.969. Bottom row DSC: Baseline U-Net = 0.606, U-Net ($\mathcal{L}_{\mathcal{COS}}$) = 0.713, Attention U-Net ($\mathcal{L}_{\mathcal{CE}+\mathcal{DSC}}$) = 0.747, U-Net ($\mathcal{L}_{\mathcal{COS}}+\mathcal{SAM}$) = 0.784 and Ensemble-7-VOTING = 0.780. There is no significant difference between the baseline and our modified U-Nets in the case of the best segmentation. However, there is a maximum of 18% improvement in the DSC for the worst case. The worst case has cysts all over the abdomen region, which makes model prediction difficult, nonetheless, the attention mechanisms, cosine loss, SAM, and ensembles help improve the segmentation and can be useful in locating cysts in other regions as well.

individual networks. Moreover, the overall best MSSD value of 2.09 mm is provided by the CBAM U-Net with $\mathcal{L}_{COS}+SAM$. We further observe that in six out of eight cases, the $\mathcal{L}_{COS}+SAM$ performs better than the corresponding $\mathcal{L}_{CE+DSC}+SAM$ when the respective DSC and MSSD values are compared.

For cases with DSC ≤ 0.8, the Attention U-Net with $\mathcal{L}_{COS}+SAM$ produces similar segmentation results as described before. Again, the network surpasses the baseline U-Net.

3.4.4 Ensemble

We observe that every ensemble model achieves a higher DSC than any individual network for the same patch size. The same is observed for MSSD values except for the ensembles with STAPLE and the patch size of 96. The highest results are obtained using ensemble with seven models and majority voting (DSC: 0.918 ± 0.048 & MSSD: 1.484 ± 2.083 mm, see Table 3.1). The DSC and MSSD after post-processing reach 0.922 ± 0.047 and 1.094 ± 1.376 mm, respectively. We also observe that some ensemble results are significantly better (p-value < 0.05) than the corresponding

best individual network (marked in italic in Table 3.1 and Table 3.2). Moreover, the ensembles with majority voting outperform those using STAPLE. No significant differences in performance between the ensembles with four and seven models were observed ($p > 0.05$).

3.4.5 Evaluation of Total Kidney Volume

Table 3.3 displays the R^2 values of manual segmented TKVs (ground truth) versus calculated TKVs from the selected networks' segmentations. We observe that there is a high correlation between ground truth and segmentation for smaller volumes, while for larger volumes over- or under segmentation occurs. The R^2 for all networks is greater than 0.91, supporting the visual analysis. Furthermore, all the networks except one (U-Net ($\mathcal{L}_{COS}$)) outperform the baseline demonstrated by a higher R^2 with less deviation in the mean TKV difference (%) values. This conforms to increased segmentation accuracy in DSC and MSSD (cf. Table 3.2). The highest R^2 value of 0.9626 is achieved by the Attention U-Net ($\mathcal{L}_{CE+DSC}$). It outperforms the baseline U-Net by 5%. Furthermore, the mean TKV difference for the baseline is $-4.43 \pm 18.9\%$. In comparison, the mean TKV difference for the Attention U-Net ($\mathcal{L}_{CE+DSC}$) and the U-Net ($\mathcal{L}_{COS} + SAM$) are $-2.04 \pm 12.2\%$ and $-1.72 \pm 12.5\%$, respectively. The mean TKV differences for the ensembles of 7 networks and majority voting or STAPLE are $-0.65 \pm 13.76\%$ and $-4.00 \pm 14.82\%$, respectively. Meanwhile, the mean TKV differences for four network ensembles with majority voting or STAPLE are $2.34 \pm 13.36\%$ and $-3.11 \pm 14.63\%$, respectively.

Architecture	Loss	$R^2 \uparrow$	Mean TKV Difference (%) $\downarrow$
Baseline U-Net	$\mathcal{L}_{CE+DSC}$	0.915	-4.43 ± 18.90
Attention U-Net	$\mathcal{L}_{CE+DSC}$	0.962	-2.04 ± 12.20
	$\mathcal{L}_{COS}$	0.946	-1.94 ± 16.14
	$\mathcal{L}_{COS} + SAM$	0.951	-1.63 ± 15.73
U-Net	$\mathcal{L}_{COS}$	0.914	0.16 ± 16.79
	$\mathcal{L}_{COS} + SAM$	0.958	-1.72 ± 12.50
Ensemble-4-STAPLE	$\mathcal{L}_{COS} + SAM$	0.953	-3.11 ± 14.63
Ensemble-7-STAPLE	$\mathcal{L}_{CE+DSC} + SAM + \mathcal{L}_{COS} + SAM$	0.957	-4.00 ± 14.82
Ensemble-7-VOTING	$\mathcal{L}_{CE+DSC} + SAM + \mathcal{L}_{COS} + SAM$	0.952	-0.65 ± 13.76

Table 3.3: The R^2 and mean TKV difference (%) of selected network configurations (post-processed, patch size: 128) for ground truth TKV v/s predicted TKV (ml). The baseline linear fit has R^2 value of 0.915. The U-Net with $\mathcal{L}_{COS}$ and $\mathcal{L}_{COS} + \mathcal{SAM}$ have R^2 values of 0.914 and 0.958, respectively. The Attention U-Net with $\mathcal{L}_{COS}$ and $\mathcal{L}_{COS} + \mathcal{SAM}$ have R^2 values of 0.946 and 0.951, respectively. Among the Ensembles, the highest R^2 is 0.957. Meanwhile, the Attention U-Net with $\mathcal{L}_{CE} + \mathcal{DSC}$ achieves the overall highest R^2 value of 0.962.

3.5 Discussion

In this work, we investigated the impact of various attention modules, cosine loss, and SAM for improving kidney segmentation in ADPKD from T1-weighted MR

images to estimate TKV while mitigating the problem of limited data. Thereby, our goal was to achieve high segmentation accuracy while only using a limited number of annotated data. Compared to other approaches reported in the literature [49, 50], we achieved similar results but using only a fraction of the data employed by others. We further conducted experiments with ensembles of our networks yielding more improvements. In the following, we first discuss the impact of individual networks and then our results combining these into ensembles.

3.5.1 Individual Networks

We found that all attention networks with $\mathcal{L}_{COS}$ and $\mathcal{L}_{CE+DSC}$ outperformed the baseline networks across all metrics (Table 3.1). Combining the baseline U-Net with $\mathcal{L}_{COS}$ and SAM yielded the best result among the individual networks with DSC scores of 0.918 ± 0.04 and 0.908 ± 0.05 for patch sizes 128 and 96, respectively. Additionally, the MSSD was minimal for this combination. Generally, we observed that patch size can play an important role. Here, most of the networks that have been trained with a patch size of 128 significantly outperformed the networks trained with a patch size of 96. For example, the U-Net ($\mathcal{L}_{COS} + \mathcal{SAM}$) with a patch size of 128 significantly outperformed the same U-Net with a patch size of 96 ($p = 0.0008$ for DSC, $p = 0.039$ for MSSD).

The impact of the Attention U-Net is more prominent for challenging cases (i.e., segmentations with DSC < 0.80). The number of such cases for the baseline U-Net ($\mathcal{L}_{CE+DSC}$) was 13; however, this number is reduced to 5 in the case of the proposed networks. These five cases were common to both networks, indicating the failure of the U-Net architecture to accurately segment such difficult samples. These samples contain cysts in various regions of the abdomen and less-defined kidney boundaries and shapes.

Even though these cases were difficult to segment by all networks, we still see an improvement of up to 6% in the DSC compared to the baseline U-Net from DSC $= 0.638$ to the Attention U-Net ($\mathcal{L}_{CE+DSC}$) with DSC $= 0.704$. A reason might be that the U-Net ($\mathcal{L}_{CE+DSC}$) is unable to focus on relevant image information, e.g., kidney boundaries with low contrast for such difficult cases. However, the attention mechanism [58] uses higher-level features as a guide to help lower-level features to emphasize such regions in the image. The Attention, CBAM, and SE U-Nets were significantly better (p-value < 0.05) than the baselines, suggesting that the three attention mechanisms can be integrated to improve performance.

Besides the attention mechanisms, we found that simply training the U-Net with cosine loss significantly improved the performance (p-value < 0.05) over every metric (see Table 3.1). Every voxel segmented by the U-Net with $\mathcal{L}_{COS}$ lies on a unit hypersphere as they are l2-normalized. This way, the wrongly segmented voxels are heavily penalized by the cosine loss leading to the adaption of the weights during training. Results by Payer et al. support our results, though their loss function [69] is different to the one described in [65], which is implemented here. Nevertheless, our results show that the implemented loss function is also suitable for the medical image segmentation task at hand. Similar results using the cosine distance function in k-means clustering of DCE-MRI for kidney segmentation have been reported [70],

outperforming standard cluster similarity metrics. In another work [71], the authors used cosine similarity and attained state-of-the-art semantic segmentation results on the following datasets: ADE20K [72], Cityscapes [73], and COCO-Stuff [74]. In conclusion, the cosine loss function can be valuable in renal MRI segmentation.

The combination of SAM with our network architectures further boosts the performance. This indicates the usefulness of SAM in improving the generalizability of the models. However, a drawback is that it takes about twice the amount of time to train such a network compared to the same network without SAM. The reason is the need to calculate gradients twice in each iteration as it first calculates gradients for the weights and then for the neighborhood parameters. Nonetheless, this limitation renders minor with respect to the steadily increasing computational power available.

Furthermore, we notice that, on average, cosine loss (without SAM) brings about 0.65% improvement in DSC over the corresponding baseline as compared to 0.53% when only using SAM. This shows that the cosine loss is more important for segmentation accuracy than SAM. However, we also find that combining cosine loss with SAM yields an average DSC improvement of 1.35%. Hence, the combination of the two is vital for attaining best segmentation accuracy.

In comparison to other deep learning-based kidney segmentation approaches, our methods perform similarly. The approaches by Kline et al. [49], van Gastel et al. [50], and Mu et al. [52] report higher DSC (up to 0.97). Consequently, their mean TKV difference is smaller than our proposed method. However, these studies use larger datasets (up to 2400). Nonetheless, in our work, the aim was to explore and combine techniques that could deal with limited data. In this respect, using only a fraction of the data (100 cases), we still achieved comparable results. Increasing the number of datasets for our method might further improve the results. However, we explicitly aimed at investigating techniques mitigating limited data. Therefore, such a test is beyond the aim of this study.

Daniel et al. [53] attained slightly better DSC (0.93) than our methods using a plain U-Net on T2 images. They also used a small dataset for training and testing. However, the major difference is that they applied their approach to healthy volunteers and patients with chronic kidney disease. Neither of these data contained cysts that altered the appearance of the kidneys in the MRI drastically (see Figure 3.5). Furthermore, they trained their network on T2-weighted images while we conducted our study on T1-weighted images, which could also be a reason for the different performance. Finally, we outperformed the approach in [51] with a margin of 4%; however, they used data only from four patients.

3.5.2 Ensembles

Creating ensembles of our proposed networks further improved segmentation accuracy compared to the best performing individual network. The networks combined in an ensemble are known to reduce the variance component of the prediction error and, therefore, smooth out the predictions [75, 76]. The ensemble with seven networks and majority voting achieved the highest DSC and MSSD of 0.922 ± 0.047 and 1.094 ± 1.376 mm, respectively (Table 3.2). No significant differences in performances of ensembles with four and seven networks could be observed (p-value $>$ 0.05, Table 3.1).

Our ensembles cannot outperform the ones presented by Kline et al. [49]. However, the key differences lie in the data size and the number of networks employed. Kline et al. used a 24-fold larger dataset and an ensemble of 11 networks. Considering this, we believe our ensembles performed similarly.

3.5.3 Limitations

Nonetheless, our system has some limitations. For instance, there are some patients with heterogeneous distributions of cysts all over the abdomen. This makes it difficult to distinguish kidneys from other organs. In such cases, our models over-segment the kidneys by delineating parts of other abdominal organs that also contain cysts (e.g., liver). However, even though under- and over segmentation occur, the differences in the obtained versus manual segmented TKVs is 5% on average throughout all models (cf. Table 3.3). Furthermore, we have not yet demonstrated generalization in terms of applications/transfer to other domains. We are currently looking into exploiting publicly available data e.g., KiTS19 [77]. Initial results look promising, but in-depth evaluation is pending.

3.6 Conclusion

In this paper, we proposed approaches to overcome the problem of limited data for training convolutional neural networks for segmenting the TKV in kidneys with ADPKD. We demonstrated that combining the cosine loss function and SAM could achieve high segmentation accuracy while only using 100 datasets. Furthermore, TKV was obtained at high accuracy compared to manual segmentation (surpassing the inter-user agreement). Our study shows that fast and automated segmentation and TKV estimation is possible, allowing for clinical translation in the future.

Subsequently, we plan to transfer our methods to the available T2-weighted MR images and investigate a combination of both MR contrasts to fully exploit the benefit of the complementary image information.

Acknowledgments

The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) was conducted by the CRISP Investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The data and samples from the CRISP study reported here were supplied by the NIDDK Central Repositories. This manuscript was not prepared in collaboration with investigators of the CRISP study and does not necessarily reflect the opinions or views of the CRISP study, the NIDDK Central Repositories, or the NIDDK. We are thankful to the NIDDK for providing us with the patient data from the CRISP study. We gratefully acknowledge the support of the NVIDIA Corporation with the donation of an NVIDIA Titan Xp used for this research.

Conflict of Interest Statement

The authors declare no conflict of interest.

Funding

This research project is part of the Research Campus M^2OLIE and funded by the German Federal Ministry of Education and Research (BMBF) within the Framework "Forschungscampus: public–private partnership for Innovations" under the funding code 13GW0388A. It is funded under the funding code 01KU2102, under the frame of ERA PerMed. For the publication fee we acknowledge financial support by Deutsche Forschungsgemeinschaft within the funding programme "Open Access Publikationskosten" as well as by Heidelberg University.

4. Generalizable Kidney Segmentation for Total Volume Estimation

Anish Raj [1,2], Laura Hansen [1], Fabian Tollens [3], Dominik Nörenberg [3], Giulia Villa [4], Anna Caroli [4], and Frank G. Zöllner [1,2]

[1]Computer Assisted Clinical Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany
[2]Mannheim Institute for Intelligent Systems in Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany
[3]Department of Radiology and Nuclear Medicine, University Medical Centre Mannheim, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany
[4]Bioengineering Department, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Ranica (BG), Italy

4.1 Abstract

We introduce a deep learning approach for automated kidney segmentation in Autosomal Dominant Polycystic Kidney Disease (*ADPKD*). Our method combines Nyul normalization, resampling, and attention mechanisms to create a generalizable network. We evaluated our approach on two distinct datasets and found that our proposed model outperforms the baseline method with an average improvement of 9.45 % in Dice and 79.90 % in mean surface symmetric distance scores across both the datasets, demonstrating its potential for robust and accurate total kidney volume calculation from T1-w Magnetic Resonance Imaging (*MRI*) images in ADPKD patients.

4.2 Introduction

ADPKD is a common cause of chronic kidney disease, ultimately leading to end-stage renal disease and kidney failure in most cases [41]. The Total Kidney Volume (*TKV*) can be used to monitor ADPKD progression and is also recognized as a relevant prognostic marker [43, 78, 79]. TKV calculation requires accurate delineation of kidney volumes which is usually performed manually by an expert and is time-consuming. Therefore, automated segmentation is warranted [46, 47]. However, deep learning approaches should generalize well to unseen external datasets to become clinical applicable. Hence, in this work, we develop an approach by combining Nyul normalization [80], resampling, and attention mechanisms to create a generalizable neural network for ADPKD kidney segmentation.

4.3 Materials and Methods

4.3.1 Patient Data

The patient T1-w MRI data were obtained from two different sources. The first dataset was obtained from the National Institute of Diabetes and Digestive and

Kidney Disease (NIDDK), National Institutes of Health, USA, and was collected in the Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) study [41]. The second dataset was collected in the context of the EuroCYST Initiative (ClinicalTrials.gov Identifier: NCT02187432). From here on, we call the first and second Datasets A and B, respectively. Both MRI datasets were acquired in coronal orientation from ADPKD patients with Chronic Kidney Disease (*CKD*) stages 1 to 3. Tab. Table 4.1 depicts the characteristics of each dataset.

Characteristic	**Dataset A**	**Dataset B**
Patient count	93	41
Male:female ratio	50:50	46:54
Average age (years)	30 ± 10	44 ± 11
Image size	256 × 256 × [30-80]	512 × 512 × [40-66]
In-plane resolution (mm^2)	1.41 × 1.41 (±0.13)	0.73 × 0.73 (±0.03)
Slice thickness (mm)	3.06 ± 0.29	4 ± 0.00
TKV (ml)	1311 ± 977	1582 ± 970

Table 4.1: Descriptive statistics from both ADPKD datasets.

4.3.2 Image Annotation

Dataset A was annotated by two physicians with experience of 1 and 3 years, respectively [81]. The mean inter-user agreement for Dataset A was 0.91±0.06 (Dice) with a coefficient of variation of 0.07. Dataset B was annotated by a single expert with 4 years of experience in manual tracing of renal ADPKD.

4.3.3 Image Pre-Processing

We first resampled both datasets to a uniform resolution of 1.4×1.4×3.0 mm^3. We further resized Dataset B's image size to 256×256 voxels to keep it in line with Dataset A. Furthermore, to create similar histogram distribution between the two datasets, we trained Nyul normalizer [80] on Dataset A. Then we transformed Dataset B using the trained normalizer to align the image intensity distributions of both datasets. Next, we normalized intensities using volume-wise z-score normalization. These steps bring the distributions of both the datasets closer and help in network generalizability.

4.3.4 Network Architecture

We implemented two U-Net [28] based architectures for kidney segmentation. The first is nnUNet [82], which we used as a baseline model. The second is based on a combination of CBAM [60] and Attention U-Net [58]. The attention mechanisms help in extracting relevant image regions from spatial and channel dimensions of the feature maps. The CBAM and Attention modules were combined to form CBAM-Attention U-Net as described in [81]. Fig. Figure 4.1 depicts the network architecture of CBAM-Attention U-Net.

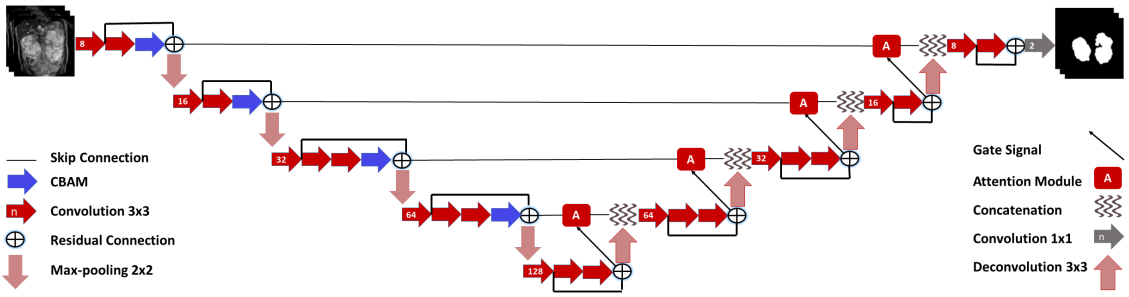


Figure 4.1: Convolution Block Attention Module (*CBAM*)-Attention U-Net with Convolution Block Attention Modules in encoder path and attention gates in decoder path of a U-Net architecture.

4.3.5 Training

We trained both baseline (nnUNet) and CBAM-Attention U-Net for 100 epochs with 5-fold cross-validation on Dataset A and then make prediction on Dataset B and vice versa for testing model generalizability on unseen datasets as a function of training size. The nnUNet was trained with its standard settings [82]. The CBAM-Attention U-Net was trained with a patch size of 128×128 with a combination of cross-entropy and dice loss and the image preprocessing as described in sub-section Section 4.3.3. The training:validation:test split for Dataset A for each fold consisted of 73:2:18 patients. For Dataset B the split was 31:2:8, respectively. Further training details can be found in [81].

4.3.6 Evaluation

We employed two metrics to compare the network predictions to the ground-truth; the Dice Similarity Coefficient (*DSC*) score for assessing the overlap and the Mean Symmetric Surface Distance (*MSSD*) (in mm) which is more perceptive to alignment and shape.

4.4 Results

Histograms of pre- and post-Nyul normalization are shown in Fig. Figure 4.2. The quantitative segmentation results are given in Tab. Table 4.2.

Network	Training dataset	DSC ↑		MSSD (mm) ↓		Volume difference (%)	
		Dataset A	Dataset B	Dataset A	Dataset B	Dataset A	Dataset B
Baseline	A	*0.914±0.057*	0.896±0.098	2.89±4.39	8.64±14.21	1.95±11.77	-0.41±5.21
	B	0.521±0.255	0.885±0.122	17.00±15.01	8.95±14.76	40.72±36.70	*-0.24±5.66*
CBAM	A	0.910±0.054	0.898±0.052	*1.21±1.45*	1.64±1.38	*1.06±13.76*	-0.35±11.85
U-net	B	0.800±0.160	*0.915±0.044*	3.33±6.90	*1.35±1.30*	18.77±22.73	2.86±7.98

Table 4.2: Quantitative results comparing the baseline nnUNet to our CBAM-Attention U-Net. All the scores are obtained by combining results from each test fold in 5-fold cross validation so that whole dataset is covered in testing. The best results are highlighted in italics. The up and down arrows indicate that higher and lower values denote better performance.

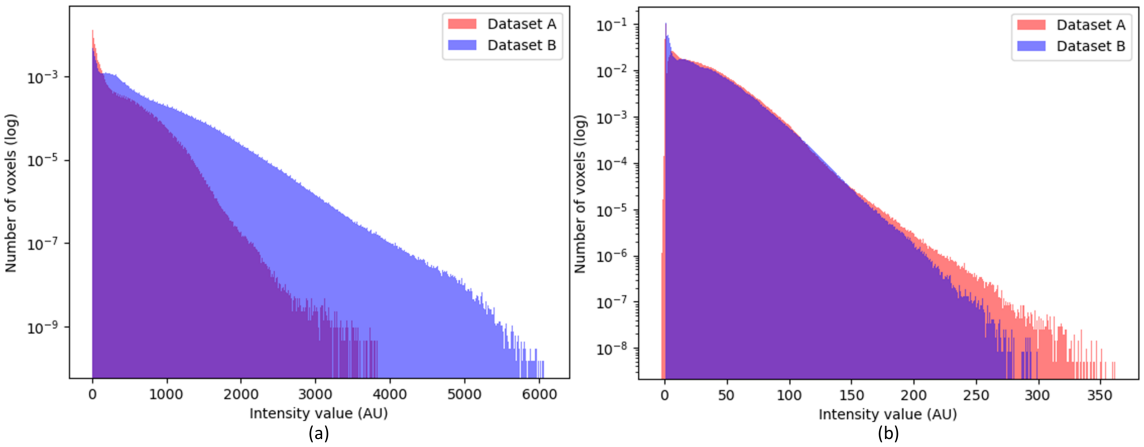


Figure 4.2: Histograms of datasets before (a) and after Nyul normalization (b), with Wasserstein distances of 227.5 and 2.4, respectively.

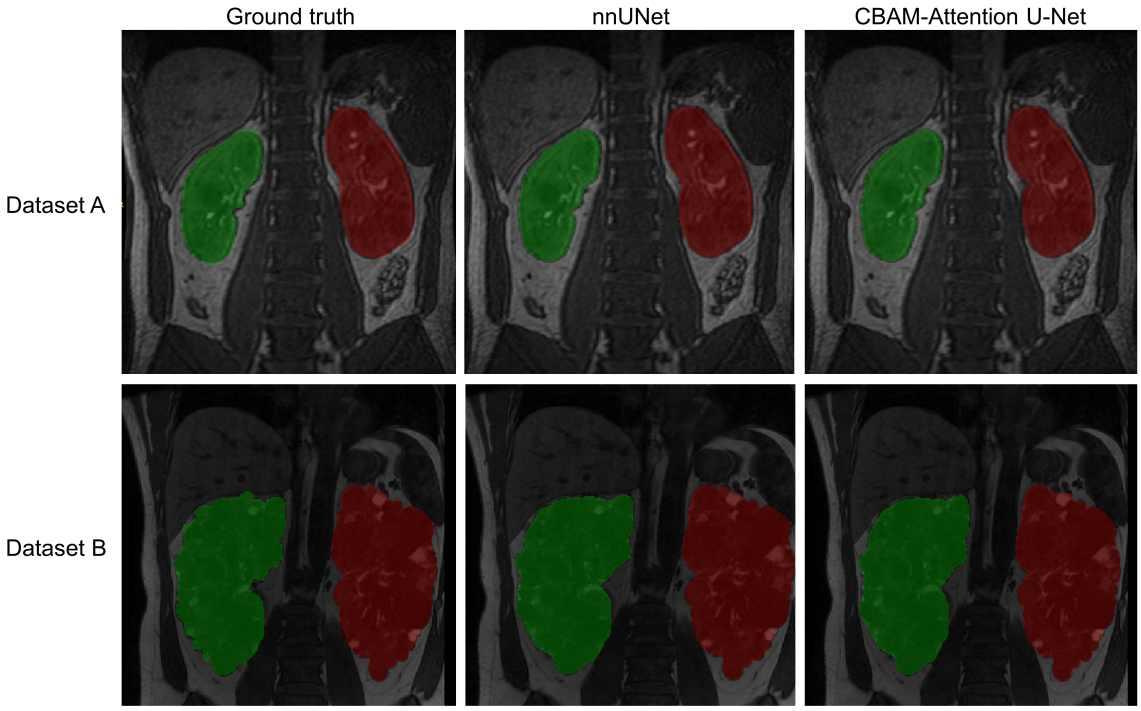


Figure 4.3: Qualitative results with the best predictions from both networks trained on Dataset A and B, respectively. For Dataset A DSCs are 0.966 and 0.963 for nnUNet and CBAM-Attention U-Net and for Dataset B 0.931 and 0.959, respectively.

For Dataset A, CBAM-Attention U-Net attains the best MSSD (1.21±1.45 mm) and volume difference (1.06±13.76%), while the baseline attains the best DSC (0.914±0.057) when trained on Dataset A. For Dataset B, the best DSC (0.915±0.044) and MSSD (1.35±1.30 mm) are obtained by the CBAM-Attention U-Net, however, the baseline provides the best volume difference of -0.24±5.66 % (here, both trained on Dataset B). Moreover, the CBAM-Attention U-Net attains a DSC of 0.898±0.052 on Dataset B when trained exclusively on Dataset A. Finally, qualitative results from the best and worst predictions are visualized in Figs. Figure 4.3 and Figure 4.4.

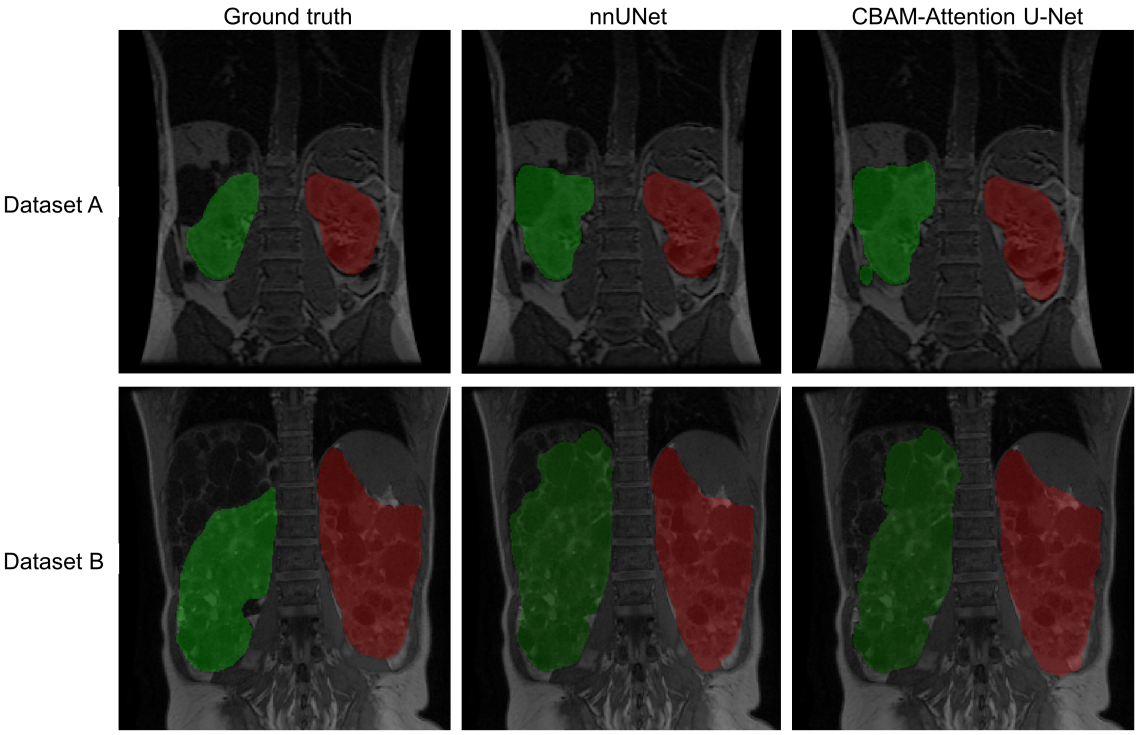


Figure 4.4: Qualitative results with the worst predictions from both networks trained on Dataset A and B, respectively. The bad performance of networks is due to the presence of cysts in the liver.

4.5 Discussion

This study aimed at creating a generalizable kidney segmentation algorithm for T1-w MRI images of patients with ADPKD. Similarity of data distribution is important to create a robust algorithm that can generalize to data from multiple sources. To achieve this, we made intensity distributions of the datasets similar by using re-sampling and Nyul normalization and then training CBAM-Attention U-Net. The CBAM-Attention U-Net outperformed nnUNet [82] (4/6 metric values), a common medical segmentation baseline. It exhibited generalization when trained on Dataset A, performing well on both sets (DSC $\approx$ 0.90). However, training on Dataset B (41 patients) showed reduced generalizability on Dataset A, indicating a likely dependence on training dataset size. Compared to other works [49, 83] (2000 and 400 cases, respectively), our approach attains lower DSC (v/s >0.97). However, they used a significantly larger training data than our study. To explore the impact of training dataset size on generalizability, additional external datasets would be necessary. In conclusion, we demonstrated that with enough data, we can create a generalizable ADPKD kidney segmentation algorithm. This approach can be helpful in reliably and automatically calculating TKV for ADPKD patients.

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

Acknowledgments

The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) was conducted by the CRISP Investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The data from the CRISP study reported here were supplied by the NIDDK Central Repository. This manuscript was not prepared in collaboration with investigators of the CRISP study and does not necessarily reflect the opinions or views of the CRISP study, the NIDDK Central Repository, or the NIDDK. Data collection in the context of the Euro-CYST Initiative was funded by ERA - EDTA. This study was supported in part by the German Federal Ministry of Education and Research (BMBF) under the funding code 01KU2102, and the Italian Ministry of Health, under the frame of ERA PerMed (ERAPERMED2020-326 - RESPECT). We also acknowledge a grant from the Italian Association for Polycystic Kidney (Associazione Italiana Rene Policistico - AIRP).

5. Automated Prognosis of Renal Function Decline in ADPKD Patients using Deep Learning

Anish Raj [1,2], Fabian Tollens [3], Anna Caroli [4], Dominik Nörenberg [3] and Frank G. Zöllner [1,2]

[1]Computer Assisted Clinical Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany
[2]Mannheim Institute for Intelligent Systems in Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany
[3]Department of Radiology and Nuclear Medicine, University Medical Centre Mannheim, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany
[4]Bioengineering Department, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Ranica (BG), Italy

5.1 Abstract

An accurate prognosis of renal function decline in Autosomal Dominant Polycystic Kidney Disease (*ADPKD*) is crucial for early intervention. Current biomarkers used are Height-adjusted Total Kidney Volume (*HtTKV*), estimated Glomerular Filtration Rate (*eGFR*), and patient age. However, manually measuring kidney volume is time-consuming and subject to observer variability. Additionally, incorporating automatically generated features from kidney Magnetic Resonance Imaging (*MRI*) images, along with conventional biomarkers, can enhance prognostic improvement. To address these issues, we developed two deep-learning algorithms. Firstly, an automated kidney volume segmentation model accurately calculates HtTKV. Secondly, we utilize segmented kidney volumes, predicted HtTKV, age, and baseline eGFR to predict Chronic Kidney Disease (*CKD*) stages >=3A, >=3B, and a 30% decline in eGFR after 8 years from the baseline visit. Our approach combines a Convolutional Neural Network (*CNN*) and a Multi-Layer-Perceptron (*MLP*). Our study included 135 subjects and the Area Under the Curve (*AUC*) scores obtained were 0.96, 0.96, and 0.95 for CKD stages >=3A, >=3B, and a 30% decline in eGFR, respectively. Furthermore, our algorithm achieved a Pearson correlation coefficient of 0.81 between predicted and measured eGFR decline. We extended our approach to predict distinct CKD stages after eight years with an AUC of 0.97. The proposed approach has the potential to enhance monitoring and facilitate prognosis in ADPKD patients, even in the early disease stages.

5.2 Introduction

ADPKD, due to the growth of cysts and therefore, degeneration of renal parenchyma leads to end-stage renal disease (ESRD) and renal failure [84]. It affects up to 12

million people worldwide. ADPKD accounts for up to 10% of patients with ESRD [84]. Given these numbers, early treatment of the disease when the renal parenchyma is still preserved is warranted. However, disease progression monitoring is difficult since kidney function may remain normal for several decades and is therefore not informative in the earliest stages of the disease. Clinical, genetic, environmental, epigenetic, and radiologic factors have been studied as predictors of progression to kidney failure in ADPKD [85, 86]. On the other hand, kidney volume has been shown to be a promising marker for disease progression [40, 41]. It is also recognized by the Food and Drug Administration (FDA) as a candidate for a prognostic biomarker for ADPKD progression [45] and used within recent clinical phase 3 studies (TEMPO 3/4 trial, primary outcome measure; the annual rate of change in Total Kidney Volume (*TKV*) over time) [37, 87].

MRI is recognized as an important tool to monitor the progression of ADPKD, and the TKV or the HtTKV has been shown to predict renal function decline in ADPKD patients [43, 85].

Thereby, complex interactions of different prognostic factors based on clinical, genetic, environmental, and radiological information determine the number of kidney cysts and their growth rates, which affect the TKV [85] and ultimately, renal function decline [88]. Based on this, prognostic models like the Mayo imaging classification tool [43] have been developed to stratify ADPKD patients into classes and predict disease progression. A multiple linear regression model is generally employed for this task. However, [89] have shown that using the conventional biomarkers (HtTKV, age, and eGFR) may not be sufficient for accurate predictions. They reported that for predicting eGFR decline after 8 years, a Pearson correlation coefficient 'r' of only 0.51 could be achieved in classifying whether a patient will reach CKD stage 3A or not (eGFR $<$ 60 ml/min/1.73 m^2), reach 3B or not (eGFR $<$ 45 ml/min/1.73 m^2), and reach a 30% decline in eGFR or not. They improved this prediction model by further incorporating texture features (repeating patterns of local image intensity variations that provide information regarding the spatial arrangement of image intensities) from T2-weighted MRI images. The texture features are extracted from manually segmented kidneys and consist of energy, entropy, and correlation values. By adding texture features to conventional biomarkers, the resulting Pearson's r reaches -0.70. However, this approach, requiring manual segmentation of the kidneys, is time-consuming and observer-dependent. Moreover, automated kidney segmentation approaches are steadily proposed and are emerging into different applications in renal MRI [46, 47]. Considering this, we created a fully automated system that can segment the kidneys automatically and can make an accurate prognosis. More in detail, we first develop a deep learning model that segments kidneys from T2-weighted MRI, enabling calculation of HtTKV. We then extract image features from the segmented kidneys and combine them with the conventional biomarkers (HTKV, age, eGFR) to predict renal function decline after 8 years, using data from 135 ADPKD patients with normal kidney function at baseline. Specifically, we classify each patient into a distinct CKD stage (CKD stages 1, 2, 3A, 3B, and 4) and predict percent change in the eGFR.

5.3 Materials and Methods

5.3.1 Patient Data

The patient data were acquired from the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK), National Institutes of Health, USA and were recorded in the Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) study [41]. The dataset consists of 241 patients in total. However, for our work, we selected patients who underwent T2-weighted MR imaging and had eGFR values > 70 ml/min/1.73 m^2 at the baseline visit as depicted in Figure 5.1. These criteria were selected in accordance with [43, 89]. The resulting dataset contains information from 135 patients. Table 5.1 lists the number of patients' CKD stages at baseline and after 8 years alongside demographic data, eGFR, and kidney volumes.

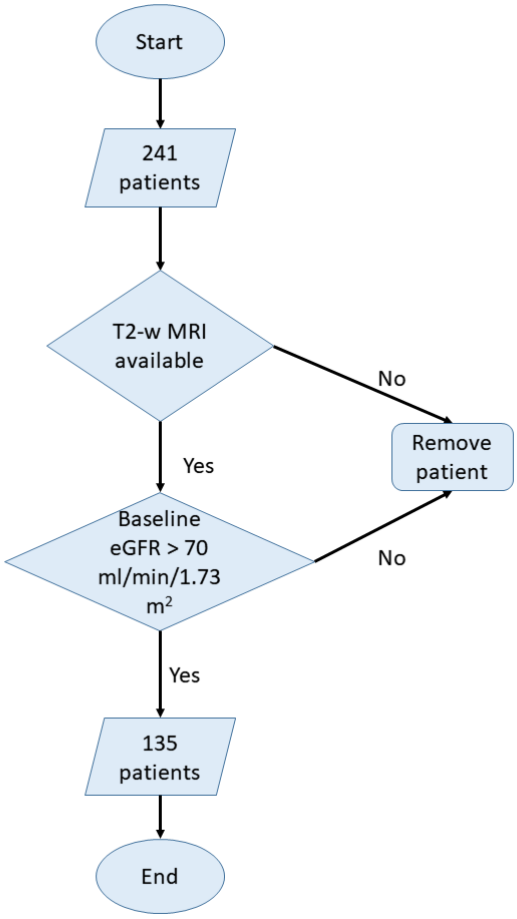


Figure 5.1: Flow chart for patient selection based on criteria from [43, 89].

To allow for a comparison to the SOTA method [89], the patients were regrouped into two groups based on the presence or absence of each of the following conditions at 8 years after baseline: reaching CKD stage 3A (eGFR < 60 ml/min/1.73 m^2), reaching CKD stage 3B (eGFR < 45 ml/min/1.73 m^2), and 30% decline in eGFR. As depicted in Table 5.2, our data resembles a similar class distribution as in the work of [89].

	Baseline	8-year Follow-up
TKV (ml)	974±585	1531±1135
HtTKV (ml/m)	559±322	880±615
eGFR (ml/min/1.73 m^2)	96±25	76±28
CKD stage 1	66	46
CKD stage 2	69	44
CKD stage 3A	-	24
CKD stage 3B	-	14
CKD stage 4	-	7
> 30% eGFR decline	-	48
Gender (M/F)	58/77	-
Age (years)	32±9	-

Table 5.1: Demographic, clinical and radiological data of the 135 ADPKD patients included in the study. Data are shown as mean ± SD or Number (%)

Criteria	Current study	SOTA method
Reached CKD stage 3A	0.333	0.360
Reached CKD stage 3B	0.155	0.180
30% eGFR decline	0.355	0.385

Table 5.2: Ratio of the positive number of samples to the total number of samples in our dataset as compared to the one in state-of-the-art (SOTA) approach from [89]. In all three criteria, we have more imbalance as compared to the work from [89].

5.3.2 Pre-Processing

We first normalized the renal MRIs, predicted HtTKV (ml/m) (obtained as per Section 5.3.4.1), age (years), and eGFR (ml/min/1.73 m^2) at the baseline visit using the Z-score normalization (Equation 5.1),

$$\hat{I} = \frac{I - \mu(I)}{\sigma(I)}, \tag{5.1}$$

where I and $\hat{I}$ are the original and normalized image/feature, respectively. The T2-weighted volumes were recorded with a size of 256 x 256 pixels and 12-30 slices [41]. However, the network for prognosis requires uniformly sized image volumes as inputs. Hence, we center-cropped and/or padded (as required) the image volumes to the size of 224 x 224 x 16 voxel. We ensure that complete kidney volumes were present within this reshaped volume.

5.3.3 Image Annotation

The annotations are used from a previous study on kidney segmentation by [81]. In that study, two physicians independently performed annotation on coronal MR images. The physicians had an experience of one and three years, respectively.

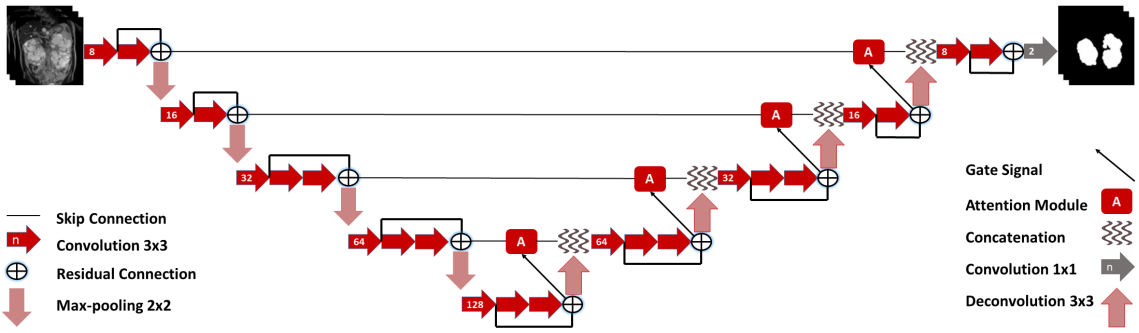


Figure 5.2: Segmentation network: 2D Attention U-Net with attention modules. The network is U-Net based [28] with attention gates that help the network focus on relevant image regions, e.g., kidneys. The network is used to segment kidneys from patient MRI volumes (baseline visit). While training, the input to the network is a patch size of 128 x 128. During inference, each slice of size 256 x 256 is used as input. The number of filters used in each block is written under the corresponding convolution block. The segmented kidneys are then used to calculate the HtTKV of the patient. Adapted from [58, 81].

Furthermore, they were overseen by an expert abdominal radiologist with nine years of experience. The annotations have the mean inter-user agreement of 0.91 ± 0.06 (Dice) and a coefficient of variation of 0.07.

5.3.4 Deep Learning Models

We implement two deep learning models for a fully automatic prognosis. We first used a segmentation network called attention U-Net (Figure 5.2) [58] to extract kidneys from MRI volumes. Thereafter, we use the segmented kidney volumes as input to our prognosis network (Figure 5.3) to make the final prognosis. We implement this two-step process so that we can first derive HtTKV automatically and then use the output from the first network and feed it to the second network for the classification/regression task.

5.3.4.1 Segmentation Network

The first network is based on previous work utilizing the attention U-Net for renal segmentation in ADPKD and TKV estimation from T1-weighted MRI [81]. Briefly, this network is composed of a U-shaped encoder-decoder architecture [28] and modified with attention gates [58], which help focus the network on relevant image regions. This is achieved by using higher-level features as a guide to suppress noise and irrelevant features in the lower-level features. The network architecture is illustrated in Figure 5.2.

The network was re-trained on 100 T2-weighted volumes of our dataset for which manually segmented kidneys were available. Kidney annotations were done similarly as described in [81]. To obtain kidney segmentations for the remaining patients without manual annotation, the trained network was applied. All automatically segmented kidneys from the network are then used to calculate the HtTKV. The HtTKV can be calculated by multiplying the number of segmented kidney voxels

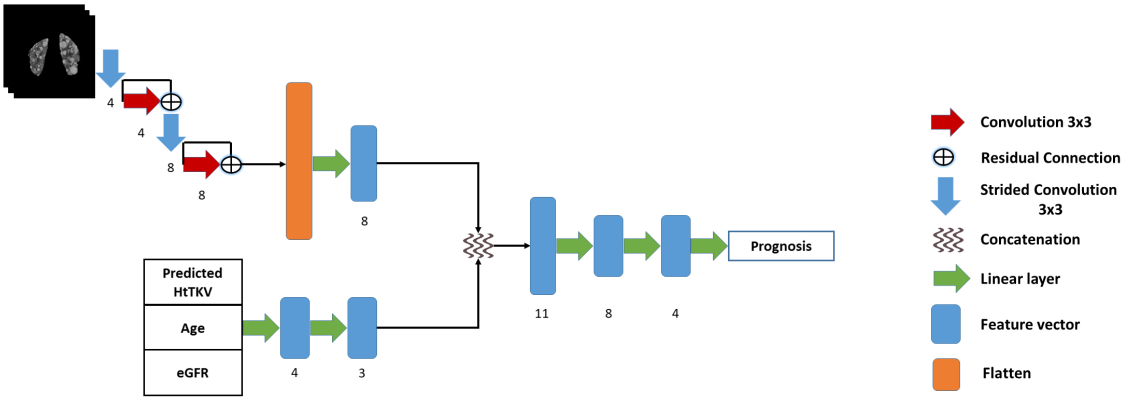


Figure 5.3: Our proposed prognosis network architecture for classification/regression prognosis. The first input is patient MRI volume (baseline visit) of size 224 x 224 x 16 that contains only segmented kidneys obtained from the segmentation network. The convolution layers are used to extract features from the volumes. The number of filters in each feature vector is written under the corresponding vector. The second input consists of a vector of predicted HtTKV, age, and eGFR at the baseline visit. Here, an MLP extracts features. Then, the features from both inputs are concatenated and used to make the final prognosis. Every convolution, strided convolution, and linear layer is followed by instance normalization and PReLU activation [19]. The strided convolutions have a stride of 4. There is no final activation for the regression task, but for classification, we use sigmoid.

by the voxel volume (in mm^3) and then dividing it by 1000 and the height of the patient (in m) to get the volume in ml/m.

5.3.4.2 Prognosis Network

Our proposed neural network architecture for predicting the final prognosis comprises two components: a CNN and a MLP. The CNN is employed to extract comprehensive image features from kidney MRI volumes. This CNN component consists of convolutional and strided convolutional layers. The strided convolutions have a stride of 4, resulting in downsampling the spatial dimensions of the image/features by a factor of 4 during each operation. The decision to utilize a shallow CNN is motivated by our limited dataset. Deeper networks typically require larger datasets to achieve robust performance. By employing a shallow network with a stride of 4, we can aggressively downsample the spatial information from the kidneys and extract features without resorting to a deeper network.

The MLP component comprises linear layers that generate one-dimensional feature vectors. We opt for linear layers as they are well-suited for processing and extracting information from structured or tabular data, such as biomarker values. The CNN component takes image volumes as input, while the MLP component accepts a vector composed of predicted HtTKV, age, and eGFR from the baseline visit. The features extracted from the CNN and MLP are concatenated, and a final MLP block is employed to make a prognosis. The concatenation of features from the CNN and MLP enables the fusion of information derived from kidney MRI volumes and conventional biomarkers.

To enhance the network's performance, we incorporate instance normalization layers [90] and Parametric Rectified Linear Unit (*PReLU*) activation functions [19] after each convolutional, strided convolutional, and linear layer (excluding the last linear layer). We opt for instance normalization instead of batch normalization due to the utilization of a small batch size resulting from our limited dataset. Batch normalization can introduce noise when applied to small batch sizes, thus, to mitigate this issue, we implement instance normalization in our network.

Depending upon the task type, the last layer is followed by the softmax activation in the case of the classification task or no activation in the case of the regression task. The complete network architecture is depicted in Figure 5.3.

The prognosis after 8 years is divided into two tasks:

1. Multi-class classification: here, we classify each patient under a distinct CKD stage category after 8 years (CKD stages 1, 2, 3A, 3B, and 4).
2. Regression: percent change in the eGFR.

As illustrated in Figure 5.3, we feed the segmented kidneys to a small CNN. Simultaneously, we use the predicted HtTKV, age, and eGFR at the baseline visit as input to a shallow MLP. Finally, we combine the features from kidney volumes and the MLP inputs and run them through a final MLP to make our prognosis.

For a proof-of-principle and to compare our approach to the SOTA method, we also considered the following task:

- Binary classifications: whether a patient will reach CKD stage 3A or not (eGFR $<$ 60 ml/min/1.73 m^2), reach 3B or not (eGFR $<$ 45 ml/min/1.73 m^2), and reach a 30% decline in eGFR or not.

Therefore, the last layer of the network is followed by a sigmoid function for each binary classification task.

5.3.5 Network Implementations & Training

5.3.5.1 Segmentation Network

The segmentation network for kidney segmentation is trained with a batch size of 16 and a patch size of 128. We used the adam optimizer [32] with a learning rate of 0.001. The loss function is cosine loss [81]. As the activation function, we employ Exponential Linear Units (*ELUs*) [67] with batch-normalization, dropout (probability=0.01) and L2 normalization (10^{-7}). Furthermore, we perform a 5-fold cross-validation [91] (as depicted in Figure 5.4) and train for a minimum of 20 epochs. The train:validation:test split was 70:10:20 subjects in each fold without shuffling. We employ early stopping to avoid overfitting and select the network with the highest dice score (validation data) for testing.

5.3.5.2 Prognosis Network

The prognosis network was trained for 30 epochs with 5-fold stratified cross-validation. The train:validation:test split was 95:13:27 subjects in each fold without shuffling. The batch size and learning rate were 8 and 0.001, respectively. For classification

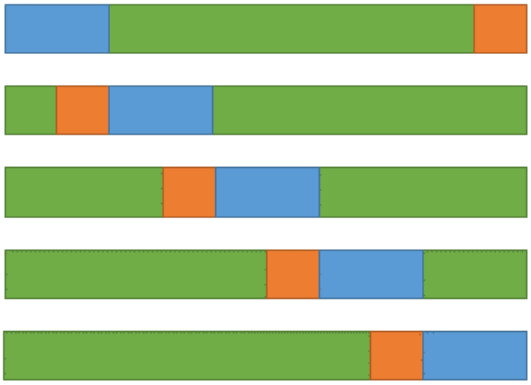


Figure 5.4: 5-fold cross-validation scheme for training networks [91] The networks are trained 5 times with different training, validation and test sets in each fold. Here, the training, validation and test sets are represented by green, orange and blue bars, respectively. Cross-validation is useful when there is limited dataset, and it helps evaluate the algorithms over a complete dataset.

tasks, we used weighted cross-entropy (CE) loss function and area under (AUC) the receiver operating characteristic curve (Receiver Operating Characteristic (*ROC*)) to select the best network. However, for the task of classifying distinct CKD stages, we use the f1-score to select the best network. We further use a weighted random sampling strategy to deal with the class imbalance in the classification of distinct CKD stages. Here, classes with less number of samples are weighted higher and accordingly sampled more often with replacement during training. Meanwhile, for regression, we implement Mean Squared Error (*MSE*) loss and use the Mean Absolute Error (*MAE*) score to choose the best network [92].

5.3.6 Evaluation

5.3.6.1 Segmentation Evaluation

We evaluate the performance of the segmentation network using the Dice Similarity Coefficient (*DSC*) score and Mean Symmetric Surface Distance (*MSSD*). First, we used the DSC to assess the overlap between the ground truth Y and the segmentation $\hat{Y}$,

$$DSC\left(\hat{Y}, Y\right) = \frac{2\left|\hat{Y} \cap Y\right|}{\left|\hat{Y}\right| + \left|Y\right|} \tag{5.2}$$

Then, we implemented the MSSD (in mm) that is more perceptive to alignment and shape:

$$MSSD\left(\hat{Y}, Y\right) = \frac{\sum_{\hat{y}\in\hat{Y}} \min d(\hat{y}, Y) + \sum_{y\in Y} \min d(y, \hat{Y})}{\left|\hat{Y}\right| + \left|Y\right|} \tag{5.3}$$

5.3.6.2 Prognosis Evaluation

To evaluate and compare the results, we used the weighted f1-score (since we have a high class imbalance for CKD stage 4) and the AUC of the ROC for classification outputs. The weighted f1-score is defined as,

$$Weighted\, f1 = \sum_{i=1}^{n=5} Support_i \times f1_i, \tag{5.4}$$

Where Support$_i$ is the support proportion of each class i, given by,

$$Support_i = \frac{\text{Number of instances in class } i}{\text{Total number of instances}}, \tag{5.5}$$

and precision, recall, and f1-score are given by

$$Precision = \frac{TP}{TP + FP}, \tag{5.6}$$

$$Recall = \frac{TP}{TP + FN}, \tag{5.7}$$

$$f1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{5.8}$$

The TP, FP, and FN are true positives, false positives, and false negatives, respectively. In addition, we used Pearson's correlation coefficient and Bland-Altman plots for evaluating regression results. Furthermore, for comparison, we employed the Mayo imaging classification tool [43, 93] to predict eGFR values at 8 years after baseline, using the baseline HtTKV.

Differences between the Mayo classification tool and our prediction network for eGFR was assessed by a two-sided t-test. Thereby, we compared the error in eGFR between the predicted values and the ground truth at 8 years. The null hypothesis that both are equal is rejected at $p < 0.05$.

5.4 Results

5.4.1 Kidney Segmentation

The segmentation network is able to segment kidney volumes at high accuracy. After post-processing with the largest connected components, the DSC and MSSD attain the values of 0.909± 0.069 and 0.721 ± 1.484 mm, respectively.

Figure 5.7 shows three examples of kidney segmentation from the proposed segmentation network.

We then applied the trained segmentation network to the remaining 35 datasets for which no ground truth annotation were available. Visual inspection of the obtained segmentations also showed good accuracy (cf. an example case in Figure 5.7, last row).

Inspecting the predicted HtTKV versus ground truth HtTKV a Pearson's correlation coefficient of 0.98 could be obtained (see Figure 5.5). In Figure 5.5, the regression line has a coefficient of regression r^2 value of 0.96. Furthermore, the mean percent difference between the predicted and the ground truth HtTKV is 13.47 ± 13.70 %. Also, the Bland-Altman plot (see Figure 5.6) supports this finding depicting a bias of 66 ml/m.

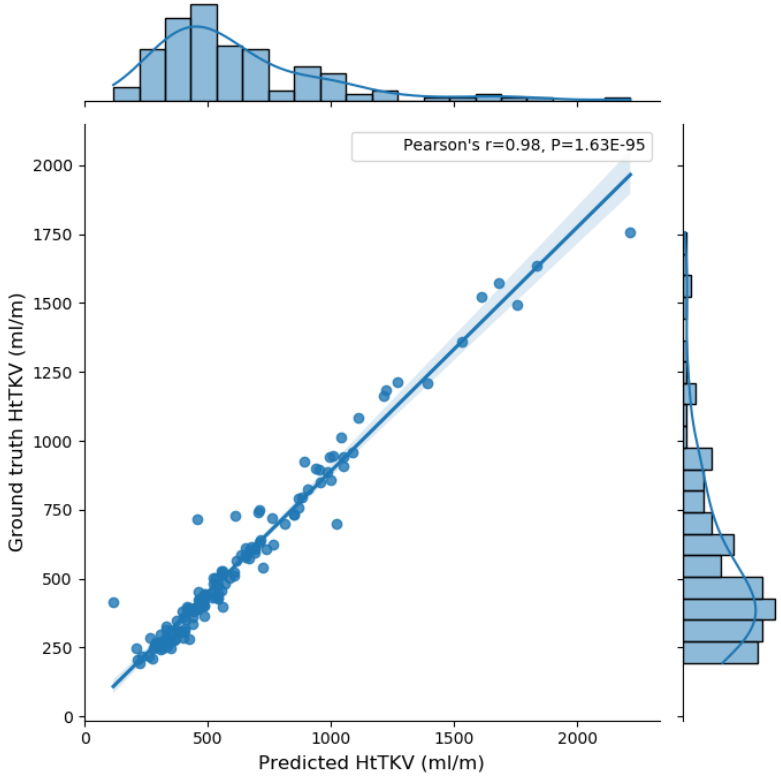


Figure 5.5: Baseline predicted HtTKV against the ground truth HtTKV. Here, the data represents all 135 patients. The TKV ground truth values for all patients are also provided by the NIDDK CRISP study [41]. We used these ground truth TKV values to calculate ground truth HtTKV. The predicted HtTKV is obtained from the segmentation network by segmenting kidneys from T2-weighted MRI volumes. Here, we make sure that the predicted data was not utilized during the training of the 5-fold cross-validation model, as shown in Figure 5.4. The Pearson correlation coefficient is 0.98.

Fold no.	**DSC ↑**	**MSSD (mm) ↓**
1	0.918 ± 0.030	0.367 ± 0.123
2	0.908 ± 0.047	0.845 ± 1.209
3	0.940 ± 0.018	0.332 ± 0.155
4	0.906 ± 0.091	0.688 ± 1.130
5	0.874 ± 0.107	1.410 ± 2.870
Average	0.909 ± 0.069	0.721 ± 1.484

Table 5.3: Segmentation results for each of the 5 folds after post-processing using largest connected components.

5.4.2 Prognosis Network

Our proposed approach was tested two-fold. First, as a proof-of-concept, we predicted whether a patient reaches CKD stage 3A, 3B, or a 30% decline in eGFR and compare the results with the SOTA method. Furthermore, we predicted eGFR

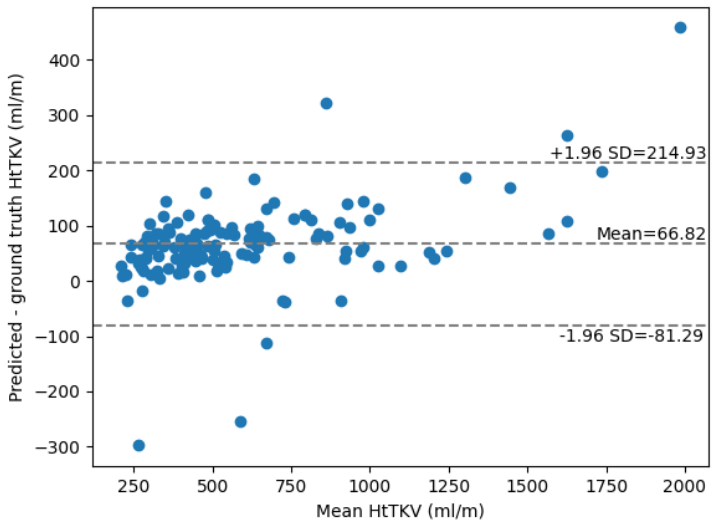


Figure 5.6: Bland-Altman plot comparing model predicted HtTKV (ml/m) values to the ground truth values.

values after 8 years for our patient data using the Mayo imaging classification tool [43, 93]. Second, we predicted distinct classes of CKD after 8 years from baseline.

Table 5.4 shows the results of our approach compared to the SOTA method.

Criteria	Precision	Recall	AUC	SOTA AUC
Reached CKD stage 3A	0.951	0.866	**0.965**	0.940
Reached CKD stage 3B	0.900	0.857	0.957	**0.960**
30% eGFR decline	0.846	0.916	**0.952**	0.850

Table 5.4: Classification results of reaching CKD stage 3A, reaching CKD stage 3B and having a 30% eGFR decline after 8 years. Abbreviations:SOTA = state-of-the-art

The AUC for reaching CKD stage 3B is over 0.950 and on par with the corresponding value from the state-of-the-art (SOTA) method (0.960) [89]. However, the AUC of reaching CKD stage 3A is 0.965 and is higher than that of the SOTA method (0.940). Further, the 30% eGFR decline predictions reach an AUC of 0.952 and clearly exceed the results of the SOTA method (0.850).

Moreover, we observe that the precision and recall for each criterion is about 90%, indicating good performance of the classifiers (Table 5.4).

Figure 5.8 depicts the predicted versus ground truth eGFR percent change after 8 years. Here, Pearson's correlation coefficient of 0.81 is attained (SOTA method = -0.700 [89]). The mean difference between predicted and ground truth eGFR percent change was found to be 1.12 ± 15.58 %. Plotting the respective Bland-Altman plot further supports this observation (see Figure 5.9). Most of the data is distributed within the 1.96 standard deviation range with a bias of only 1.12 percent change in eGFR, showing a small overestimation.

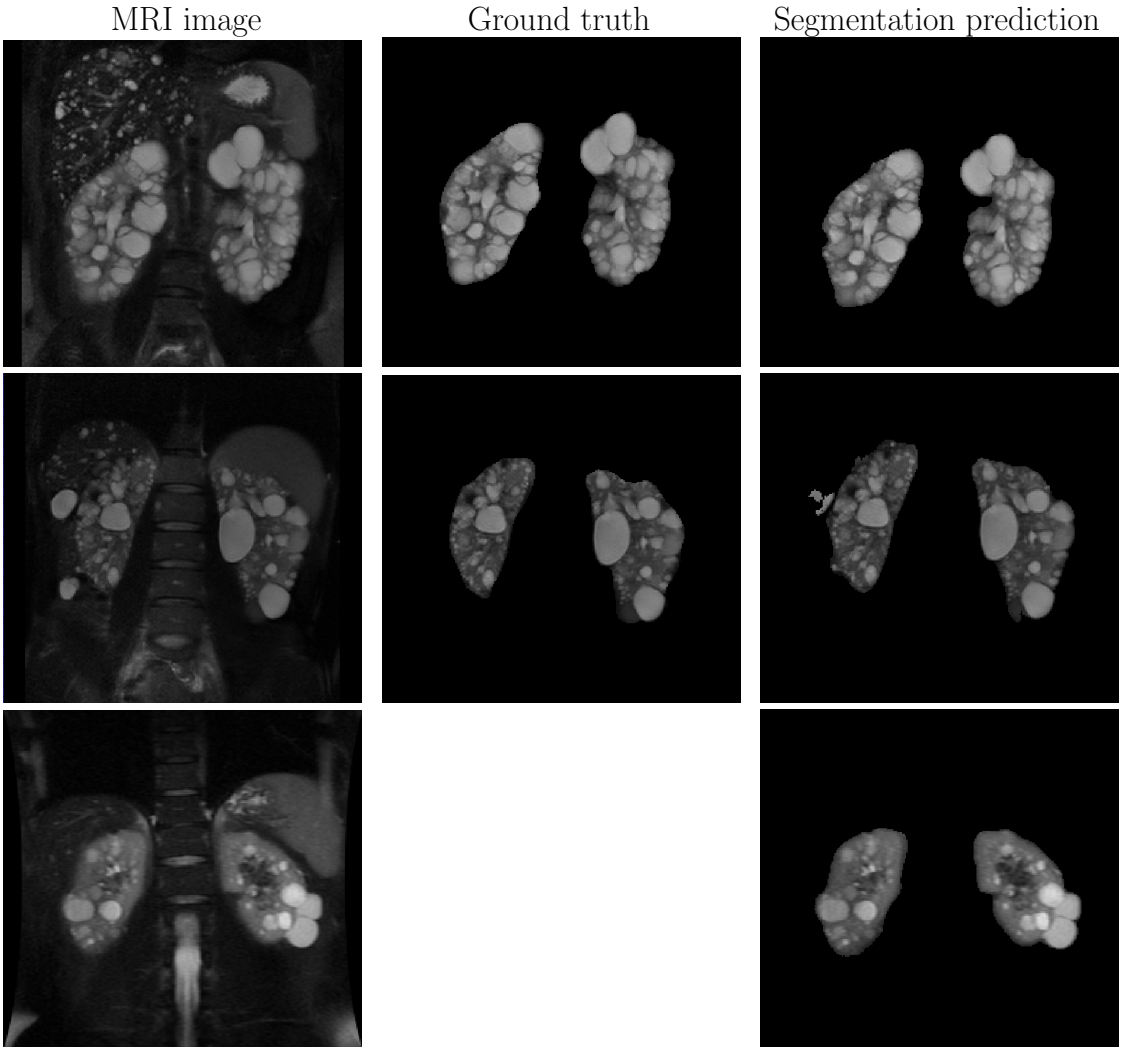


Figure 5.7: Kidney volume segmentation results. Left: T2-weighted MRI images, center: ground truth segmentations, and right: corresponding automatically segmented kidneys. Three examples show the segmented kidneys obtained from the segmentation network. The first two rows correspond to the cases that had ground truth annotation available, i.e. examples from one of the test set of 100 patients. The last row corresponds to a patient without ground truth segmentation available.

We found that Pearson's r for Mayo predicted eGFR was 0.64. In comparison, our prognosis network's predicted values had an r-value of 0.86 as shown in Figure 5.10. Furthermore, the corresponding Bland-Altman plots show that the Mayo imaging classification tool underestimates the absolute predicted eGFR (bias of -1.76 ml/min/1.73 m^2) while our method slightly overestimates the eGFR (bias of 1.18 ml/min/1.73 m^2). Moreover, our approach has a smaller range of 1.96 standard deviations (-26.87 to 29.22 ml/min/1.73 m^2) compared to the Mayo imaging classification tool (-44.88 to 41.37 ml/min/1.73 m^2).

The 8-year eGFR error obtained by our method was not significantly different than that obtained by Mayo classification (two-sided t-test, median difference [interquartile range]: 0.785 [9.95, -5.74] vs 1.85 [12.46, -15.22], $p = 0.0548$), though the p-value

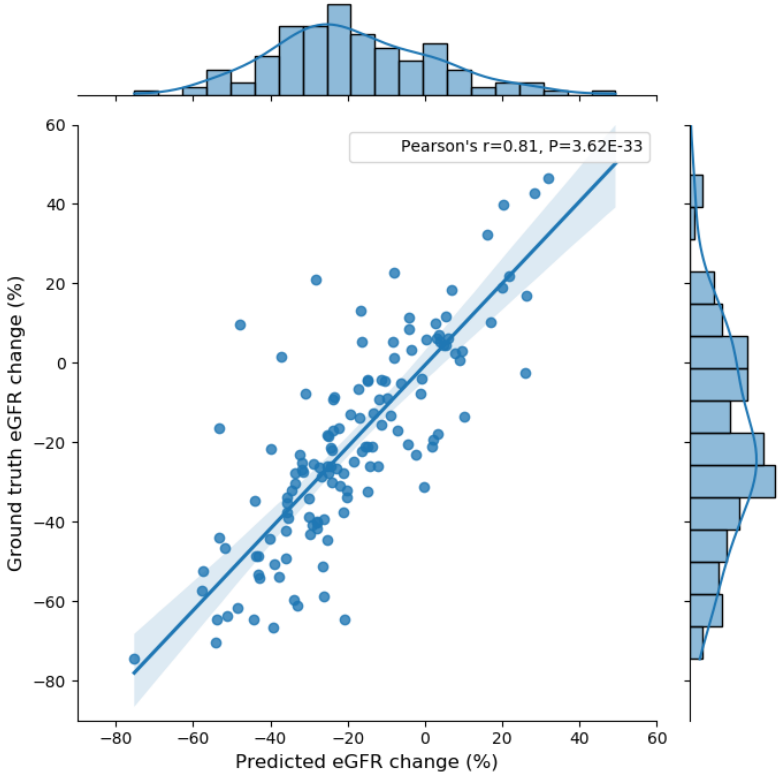


Figure 5.8: Correlation plot for predicted eGFR percent change v/s ground truth eGFR percent change after 8 years. The Pearson's correlation coefficient is 0.81.

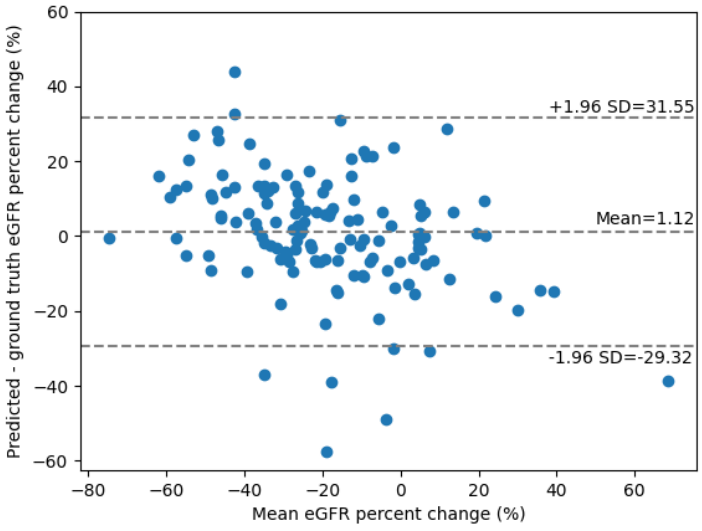


Figure 5.9: Bland-Altman plot comparing model predicted eGFR percent change (%) values to the ground truth values.

is very close to 0.05 and the difference between the two tools is also clearly shown by Bland Altman plots (Figure 5.10).

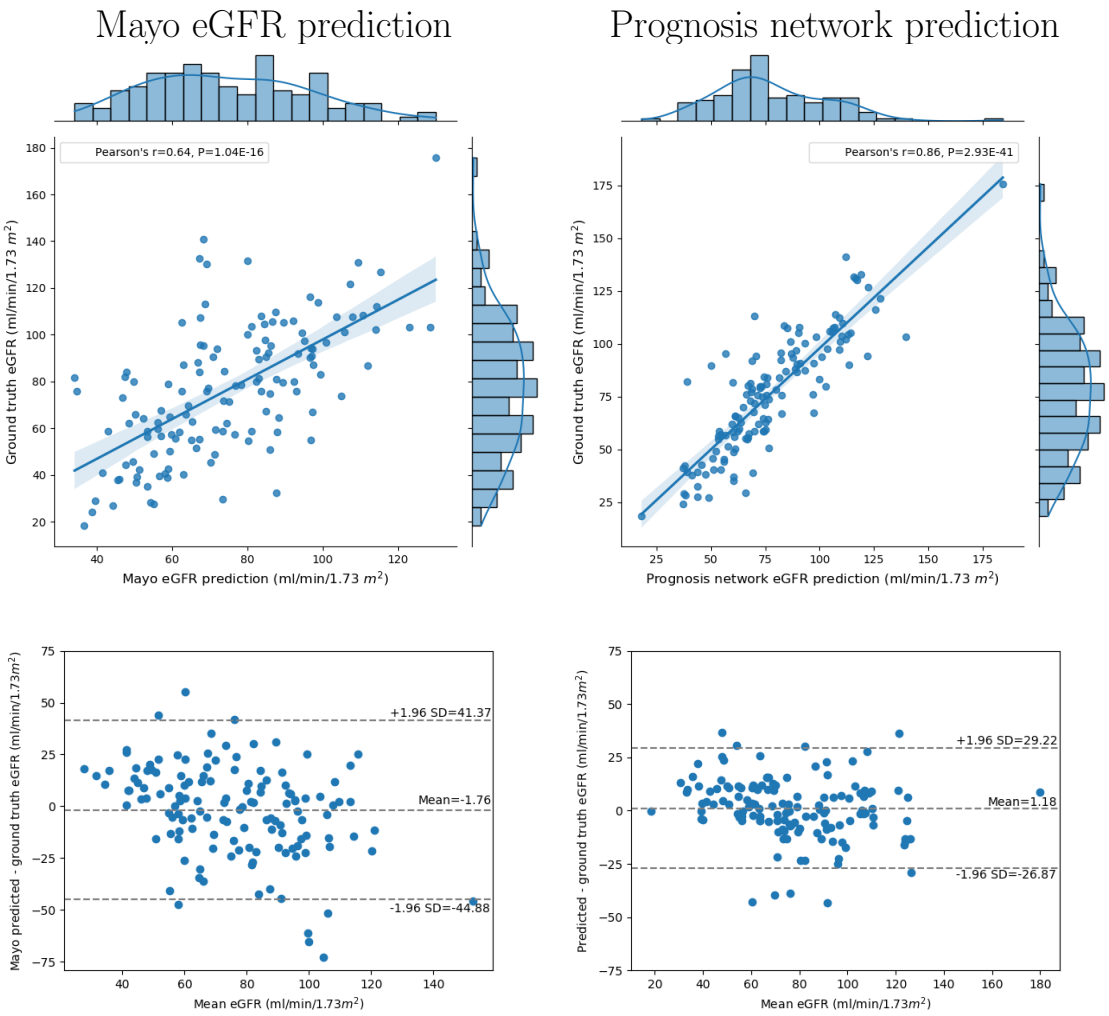


Figure 5.10: Correlation plots showing predicted versus ground truth eGFR values after 8 years in the first row. The left plot is obtained from Mayo classification tool (Pearsons's r=0.64) [43, 93] and the right plot is from our described prognosis network (Pearson's r=0.86). The second row depicts the corresponding Bland-Altman plots.

Finally, our model was also trained to predict each CKD stage distinctly for each patient after 8 years. Here, we reach a weighted f1-score of 0.851 (accuracy: 0.851) with an AUC of 0.972. The corresponding confusion matrix is shown in Figure 5.11. As can be seen, most of the predictions are on the diagonal of the confusion matrix. Furthermore, the overall accuracy increases to 0.955 when we factor in misclassified samples from adjacent stages in the confusion matrix.

5.5 Discussion

In this work, we combined T2-weighted MRIs of the kidneys with the established biomarkers (patient age, eGFR, and predicted HtTKV at the baseline visit) to predict renal function decline. The main advantage of our approach is that our model could classify patients into different CKD stages. This classification might better support the diagnosis of patients with ADPKD since it allows for precisely predicting the change in CKD class and therefore, the decline in renal function over time

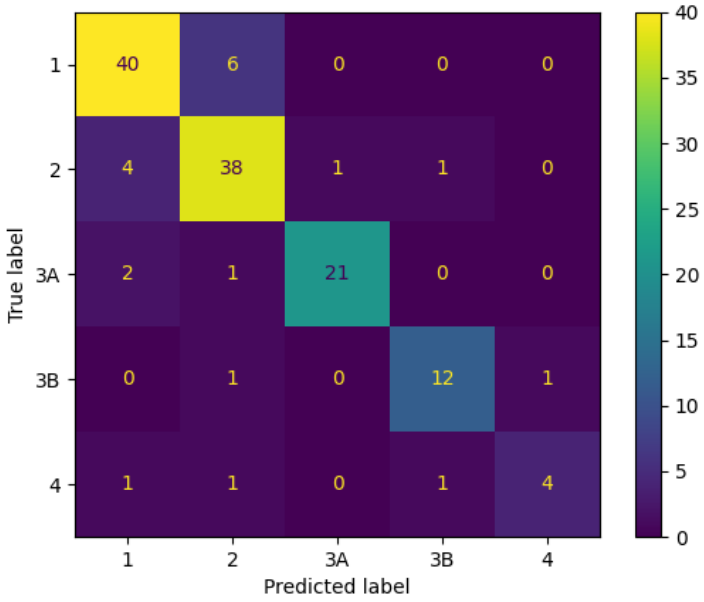


Figure 5.11: Confusion matrix depicting the predictions for distinct CKD stages after eight years. Weighted f1-score and AUC are 0.85 and 0.97, respectively.

rather than just predicting if a person will reach a certain CKD stage or not as previously reported in literature [89].

Nevertheless, our model could also demonstrate that it performs at par or better compared to the state-of-the-art method [89] for classification of reaching CKD stage 3A, 3B, and 30% eGFR decline. Furthermore, we also show that an eGFR percent decline could be predicted at a higher rate compared to the Mayo Image Classification tool.

5.5.1 Kidney Segmentation

From a visual inspection, it can be seen that the kidney segmentation is accurate. However, when inspecting the predicted HtTKV vs ground truth HtTKV plot (Figure 5.5), we observed that the network predicts slightly higher values, suggesting that it over-segments the kidneys. The mean percent difference of 13.47 ± 13.70 % between predicted and ground truth HtTKV confirms slight over-segmentation. A reason for this variability could be the slice thickness of 9.0 mm. Nonetheless, we still attained accurate prognosis performance with over-segmented kidneys as inputs to our model. Arguably, over-segmentation is better than under-segmentation since it is more likely to include the kidneys completely. Our segmentation network was originally established on T1 weighted MRIs and there we already showed good performance with respect to previously published studies [81]. In that study, we also noted that some samples contain cysts in various regions of the abdomen and less-defined kidney boundaries and shapes that might also lower the segmentation accuracy here. For comparison of T2 weighted segmentation, [50] achieved r^2 value of 0.99 between the predicted and ground truth TKV segmentation. In comparison,

our approach attains a lower r^2 of 0.96, however, it is worth noting that [50] employed four times more data than our method. Furthermore, [52] trained a V-Net [29] like architecture on 305 ADPKD patients. Their study attained a DSC of 0.96, which is higher than our DSC of 0.91. Their DSC score is better than ours since they employ 3 times more data than our study. Nonetheless, the DSC score of our study matches with the inter-user agreement of manual annotation (i.e. 0.91). In another work, [53] achieved a slightly higher DSC of 0.93 using a similar data size as our study. They employed T2-weighted images with the U-Net. The difference in performance could be due to the subjects in that study consisting of healthy and chronic kidney disease patients only. Those subjects did not have cysts present in the kidney region and hence, segmentation performance would be better in such cases.

5.5.2 Prognosis

The main objective of our proposed work was to automate ADPKD prognosis and obtain accurate results compared to the work of [89]. Often, we observe that tackling class imbalance is difficult for deep learning models[66, 94, 95], however, in this study, we achieved accurate results even though the number of positive samples was considerably less than negative samples, as seen in Table 5.1 and Table 5.2 (e.g. 23 positive samples v/s 112 negative samples for reaching CKD stage 3B). For the classification tasks (reaching CKD stage 3A, reaching CKD stage 3B, and 30% eGFR decline), our models obtain AUC > 0.950. Our method performs at par in the case of reaching CKD stage 3B classification when compared to the corresponding state-of-the-art results from [89]. Here, we achieve an AUC of 0.957. In comparison, [89] achieve an AUC score of 0.960. It is worth noting that our dataset is slightly more imbalanced than the one used in [89] as depicted in Table 5.2. Considering this imbalance, our approach is still robust enough in attaining accurate results for every criterion. Note-worthily, our method outperforms the state-of-the-art approach [89] in the cases of reaching CKD stage 3A, 30% eGFR decline, and eGFR percent change. For reaching CKD stage 3A and 30% eGFR decline, our AUCs of 0.965 and 0.952, respectively are higher than that of the state-of-the-art approach [89] (AUCs: 0.940 and 0.850, respectively).

Moreover, our model could be used to classify patients into the five CKD stages (1, 2, 3A, 3B, and 4) after eight years. Our network achieves an AUC and weighted f1-score of 0.972 and 0.851, respectively. We also find that about 95.5% of the predictions lie on the diagonal or the adjacent CKD stages in the confusion matrix (Figure 5.11). However, the network misclassifies three cases in the CKD stage 4 category. This occurs due to the class imbalance in the CKD stage 4 category as it consists of only seven cases. Increasing the dataset by including more samples for this class might improve the model's performance.

The regression task of eGFR percent change reaches a Pearson's r value of 0.810 as compared to -0.700 of the state-of-the-art method [89]. The features in [89] are negatively correlated to the eGFR percent change, hence, the negative sign. However, our method has a positive correlation since we directly compare the predicted eGFR percent change with the ground truth eGFR percent change (Figure 5.8). Furthermore, the residual standard deviation of our approach was also smaller (15.584 %

vs 17.900 %). It is clear that the networks are more accurate in predicting eGFR change (classification or regression) after 8 years. Although most of the predicted segmentations have higher HtTKV than ground truth HtTKV, there are 8 cases where there is under-segmentation. This suggests that in this case, the predicted segmentations do not cover the kidneys completely. The under-segmentation is on average -20.25 % lesser than the ground truth HtTKV for these cases. Nonetheless, the average absolute difference in eGFR percent change for these eight cases is 9.20%, while the average absolute difference for all 135 patients is 11.50%. This suggests that even though some segmentations do not cover kidneys completely, the other baseline inputs (age, baseline HtTKV, and eGFR) help the prognosis network to predict robust eGFR percent change values. Compared to the Mayo imaging classification tool [43], our algorithm reached a higher correlation of the predicted eGFR to the ground truth eGFR (r=0.64 vs. r=0.86, Figure 5.10). However, the 8-year eGFR error obtained by our method was not significantly different than that obtained by Mayo classification (p=0.0548). Nevertheless, in the light of the correlation and the Bland Altman plot, we believe that including more data might also show significance between the two approaches. Eventually, the main goal was to show the feasibility of automating common approaches to predict disease progression in ADPKD patients. Til now, our tool is on par with the state-of-the-art (Mayo Tool) but more automated.

A few other works exist that predict renal function decline but are not directly comparable to our approach because of differences in datasets and renal function decline definitions. [96] compared various machine learning algorithms on a dataset of 2166 subjects. They defined renal function decline as eGFR decline of more than 3 ml/min/1.73 m^2/year or follow-up eGFR $<$ 60 ml/min/1.73 m^2 (i.e., CKD stage 3A to 5). The algorithms were trained using 24 predictive variables, e.g., age and gender. Their best-performing algorithm was gradient boosting, reaching an AUC of 0.914 on the test data. [97] explored a prediction model to predict CKD stage 3A or positive proteinuria in a cohort of 348 subjects. Their prediction model combined a genetic risk score model with a non-genetic risk score model by incorporating genetic and non-genetic factors for prediction. This combined model achieved an AUC value of 0.894.

There exist a few approaches that seek to classify distinct CKD stages automatically. However, all approaches do not incorporate imaging data nor do a long-term (larger than 18 months, [98]) prediction of renal function decline. [99] created a random forest model that predicts 5 CKD stages (from 1 to 5). They achieved an f1-score of 0.778 with a dataset of 1718 samples. [100] employed a dataset of 400 instances and predicted 6 CKD stages (stages 3A and 3B separately). They obtained an overall accuracy of 0.855 using the J48 algorithm, a decision tree based approach. Finally, [101] trained a network of probabilistic neural networks on a dataset of 361 patients, achieving an overall accuracy of 0.967. Despite the difference in the underlying data like medical records (such as age, blood pressure, hypertension, hemoglobin, etc.) our integrated approach incorporating imaging information and clinical information (age, eGFR) could achieve a similar or higher classification accuracy while employing only about 3-fold fewer data and allowing for accurate kidney segmentation and TKV estimation simultaneously.

In the future, we plan to use T1-weighted kidney volumes and combine them with their T2-weighted counterparts to extract better features and attempt to improve performance. Moreover, so far we only predicted kidney function change for 8 years in the future. Translating our method to other time ranges will be investigated in the future. Furthermore, we plan to do multi-task learning by combining the training of segmentation and prognosis networks.

In conclusion, we have presented an automated approach to predict disease progression in ADPKD, in terms of eGFR decline and CKD stage change by integrating imaging information and clinical data. Simultaneously, renal segmentations are also obtained and used in further diagnostic tasks. Our approach might improve monitoring and support the prognosis of ADPKD patients from the earliest disease stages.

Conflict of Interest Statement

All the authors declare no competing interests.

Acknowledgments

The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) was conducted by the CRISP Investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The data from the CRISP study reported here were supplied by the NIDDK Central Repository. This manuscript was not prepared in collaboration with investigators of the CRISP study and does not necessarily reflect the opinions or views of the CRISP study, the NIDDK Central Repository, or the NIDDK. We are thankful to the NIDDK for providing us with the patient data from the CRISP study. We gratefully acknowledge the support of the NVIDIA Corporation with the donation of an NVIDIA Titan Xp used for this research. This research project is part of the Research Campus M2OLIE and funded by the German Federal Ministry of Education and Research (BMBF) within the Framework "Forschungscampus: public-private partnership for Innovations" under the funding code 13GW0388A. This study was supported in part by the German Federal Ministry of Education and Research (BMBF) under the funding code 01KU2102, and the Italian Ministry of Health, under the frame of ERA PerMed (ERAPERMED2020-326 - RESPECT). Dr. Caroli acknowledges a grant from the Italian Association for Polycystic Kidney (Associazione Italiana Rene Policistico - AIRP). The authors wish to thank Dr Norberto Perico, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Bergamo, Italy, for his valuable comments and suggestions.

6. A Generalizable Deep Voxel-Guided Morphometry Algorithm for the Detection of Subtle Lesion Dynamics in Multiple Sclerosis

Anish Raj [1,2], Achim Gass [3,4], Philipp Eisele [3,4], Andreas Dabringhaus [5], Matthias Kraemer [5,6] and Frank G. Zöllner [1,2]

[1]Computer Assisted Clinical Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany
[2]Mannheim Institute for Intelligent Systems in Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany
[3]Department of Neurology, University Medical Centre Mannheim, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany
[4]Mannheim Center for Translational Neurosciences, Heidelberg University, Mannheim, Baden Württemberg, Germany
[5]VGMorph GmbH, Mülheim an der Ruhr, Nordrhein-Westfalen, Germany
[6]NeuroCentrum, Grevenbroich, Nordrhein-Westfalen, Germany

6.1 Abstract

Multiple Sclerosis (*MS*) is a chronic neurological disorder characterized by the progressive loss of myelin and axonal structures in the central nervous system. Accurate detection and monitoring of MS-related changes in brain structures are crucial for disease management and treatment evaluation. We propose a deep learning algorithm for creating Voxel-Guided Morphometry (*VGM*) maps from longitudinal Magnetic Resonance Imaging (*MRI*) brain volumes for analyzing MS disease activity. Our approach focuses on developing a generalizable model that can effectively be applied to unseen datasets. Longitudinal MS patient high-resolution 3D T1-weighted follow-up imaging from 3 different MRI systems were analyzed. We employed a 3D residual U-Net architecture with attention mechanisms. The U-Net serves as the backbone, enabling spatial feature extraction from MRI volumes. Attention mechanisms are integrated to enhance the model's ability to capture relevant information and highlight salient regions. Furthermore, we incorporate image normalization by histogram matching and resampling techniques to improve the networks' ability to generalize to unseen datasets from different MRI systems across imaging centers. This ensures robust performance across diverse data sources. Numerous experiments were conducted using a dataset of 71 longitudinal MRI brain volumes of MS patients. Our approach demonstrated a significant improvement of 4.3% in Mean Absolute Error (*MAE*) against the state-of-the-art (SOTA) method. Furthermore, the algorithm's generalizability was evaluated on two unseen datasets (n=116) with an average improvement of 4.2% in MAE over the SOTA approach. Results confirm that the proposed approach is fast and robust and has the potential for broader clinical applicability.

6.2 Introduction

MS is a chronic neurological disorder characterized by progressive loss of myelin and axonal structures in the central nervous system [102]. Serial MRI examinations of MS patients represent an important part of the diagnostic and monitoring workout of MS patients, including therapeutic decisions [102–104]. While the appearances of new and contrast-enhanced MS lesions are mostly related to clinical relapses, smoldering chronic active lesions, which are often not detectable in routine MRI scans, represent chronic inflammation and tissue destruction and may correlate with slow and chronic disease progression [105]. Accurate detection and monitoring of MS-related changes in brain structures are important background information for clinical management [105]. Evaluating subtle alterations across multiple examinations has become feasible to characterize disease evolution over time [105]. This includes fine analysis of white matter lesions, enlargement of the cerebrospinal fluid (CSF) compartment, and grey matter atrophy [106].

Traditionally, the assessment of MS disease activity has primarily relied on the detection of new lesions [104]. Recently there has been an increasing interest in the detection of lesion activity including even subtle changes like smoldering lesions. There is a growing need for automated methods capable of generating complete maps quantifying structural brain tissue changes. Such methods are VGM [107] and deep VGM [108], where a neural network approximates a high dimensional deformation field for detecting changes in MS lesions in longitudinal MRI scans. The deep VGM approach by [108] is fast, however, we intended to improve its robustness making it more applicable to a clinical setting.It is vital to develop a robust deep-VGM approach that is independent of the MRI system used. We aimed to develop a model that can effectively generalize to unseen datasets, allowing for fast, robust, and reliable monitoring of subtle MS disease activity.

In summary, this paper investigates a generalizable deep learning approach for VGM map generation, such that it provides a generalizable tool for fast, accurate, and automated analysis of subtle MS disease activity.

6.3 Materials and Methods

6.3.1 Patient data

In this retrospective study, we analyzed two datasets of patients with MS from two different centers, following the 2010 diagnostic criteria by [109]. These datasets are referred to as Dataset A and Dataset B. Dataset A, which comprises 71 patients, is the same dataset utilized in the state-of-the-art method proposed by [108]. Dataset B consists of 97 patients. Every patient underwent two MRI examinations: one at baseline and a follow-up scan after a 12-month period. Patient demographics of these datasets are given in Table 6.1.

For further validation, we procured an external public dataset comprising 19 patients from [110]. We call this Dataset C in our study. All patients were imaged with a high-resolution T1-weighted Magnetization Prepared Rapid Gradient Echo Image (MPRAGE) sequence. Please see the acquisition details in Table 6.2.

Property	Dataset A	Dataset B
Gender (female:male)	64:13	70:27
Mean age (years)	37.67 ± 12.5	54.5 ± 14.6
PPMS	1	15
SPMS	4	19
RRMS	62	63
Mean disease duration (years)	5.71 ± 8.44	10 ± 14.32
Median EDSS (range)	2.0 (0 - 6.5)	3.5 (0 -7.5)
Treatment (DMTs)	49/67	81/97

Table 6.1: Patient demographics of Dataset A and B. In the original study that compiled Dataset A, 4 patients were excluded, thus no patient demographics for these 4 patients were recorded. However, in this work, the imaging data of these patients were used. PPMS = primary progressive MS, SPMS = secondary progressive MS, RRMS = relapsing-remitting MS, EDSS = Expanded Disability Status Scale, DMTs = individually selected immune therapies)

Property	Dataset A	Dataset B	Dataset C
Scanner	Magnetom Skyra, Siemens	Magnetom Allegra, Siemens	Unknown scanner, Philips
Field strength (T)	3.0	3.0	3.0
Sequence	T1-w MPRAGE	T1-w MPRAGE	T1-w MPRAGE
TR (ms)	1900	2080	10.3
TE (ms)	2.49	3.93	6
TI (ms)	900	1100	835
Flip angle	9°	15°	8°
Spatial resolution (mm^3)	0.94 × 0.94 × 2.00	1 × 0.98 × 0.98	0.82 × 0.82 × 1.17
Volume size (voxels)	256 × 256 × 70	160 × 240 × 256	256 × 256 × 120

Table 6.2: Image acquisition characteristics for each dataset. (MPRAGE: Magnetization Prepared Rapid Gradient Echo)

6.3.2 Ground truth VGM generation

VGM is a technique used for aligning 3D MRI images and generating maps that reveal global and regional changes in the brain between two sets of 3D MRI data collected at different time points. It utilizes T1-weighted MRI data. To initiate the process, high-quality brain masks are required, which can be generated using the FreeSurfer software package (refer to [111] for details). The VGM process unfolds in four sequential steps:

1. Coarse Linear Alignment: In this initial step, VGM determines an affine transformation that maximizes the overlap between the brain masks of the two time points. This coarse linear alignment helps bring the images into initial alignment.
2. Inhomogeneity Correction: To eliminate low-frequency bias in the images, a correction is applied by comparing the coarsely aligned images, as described in [106].
3. Fine Linear Alignment: After inhomogeneity correction, a cross-correlation-based technique is employed to achieve finer alignment between the images. This step further refines the alignment achieved in the previous coarse alignment step.

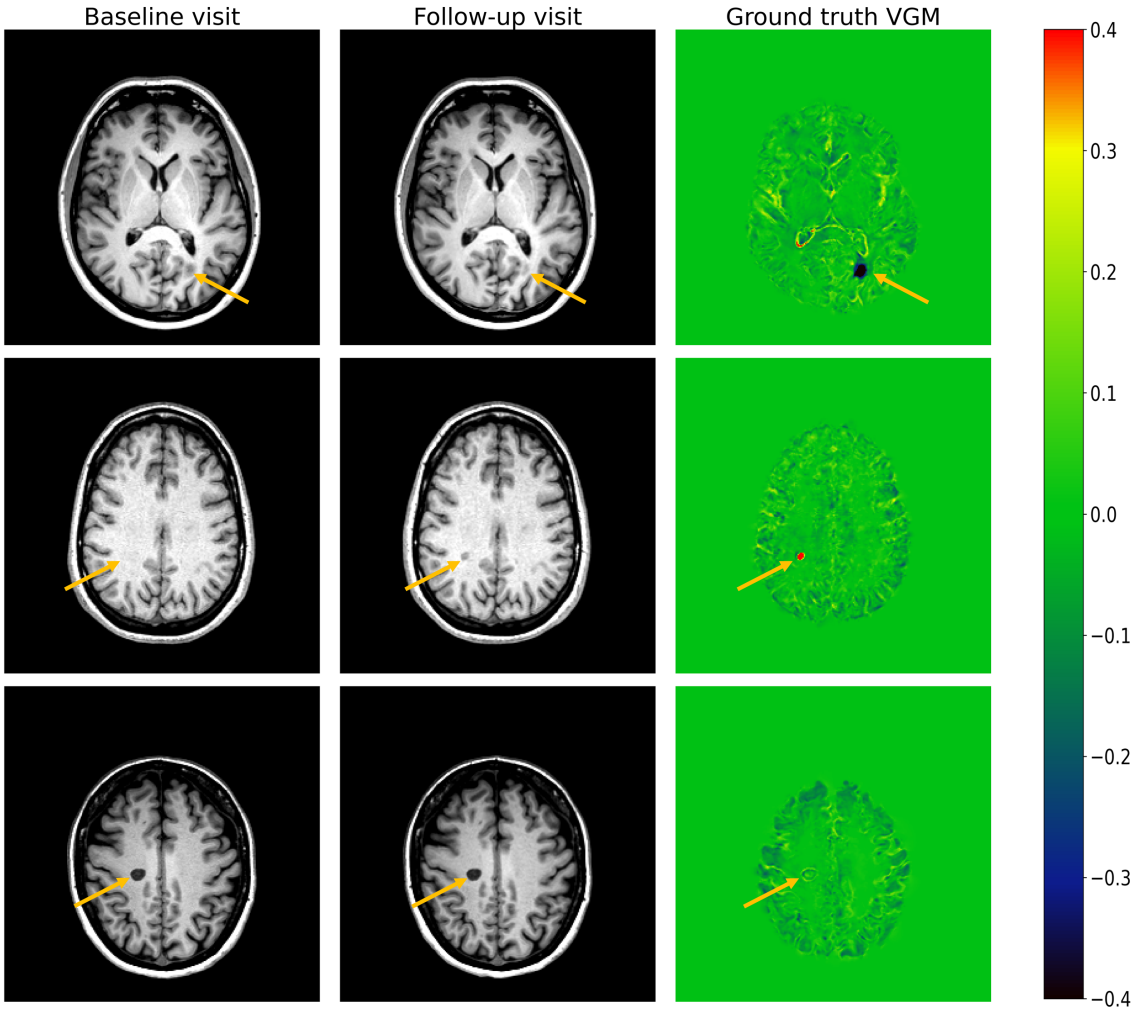


Figure 6.1: Three examples of VGM (ground truth) from a patient scan. Left column: baseline image; middle column: follow-up image; right column: corresponding VGM map. The top slice depicts a lesion decreasing in volume (dark region). The middle slice VGM shows that one new lesion appeared in the follow-up visit (small bright-red region). The third slice shows a lesion with a very small change between the baseline and the follow-up visits. Arrows indicate the location of lesions in each case.

4. The final step involves the application of a high-dimensional multiresolution full multigrid method. This step is crucial for capturing nonlinear deformations in the brain structures, allowing for comprehensive exploitation of information and effective image processing, as explained in [107].

It's worth noting that typical computation times for these four steps on a CPU are approximately 4 minutes for steps (i) to (iii) and 7 minutes for step (iv).

In the ensuing stage, the VGM process orchestrates a guided movement for each voxel based on its grey value, aligning it from the source to the target image. The ultimate objective is to extract volume alterations for each voxel from the high-dimensional deformation field. The outcome is a map that assigns a quantified value to each voxel, indicating whether the corresponding brain region has undergone an increase or decrease in volume.

To illustrate the application of VGM, we provide an example case comprising a baseline image, a follow-up image, and the resulting VGM map in Figure 6.1. Initially designed for stroke data analysis [107, 112], recent research has demonstrated its efficacy in the context of MS [113–116]. However, it is important to note that the clinical utilization of VGM is currently impeded by the relatively long computation time of approximately 11 minutes per case.

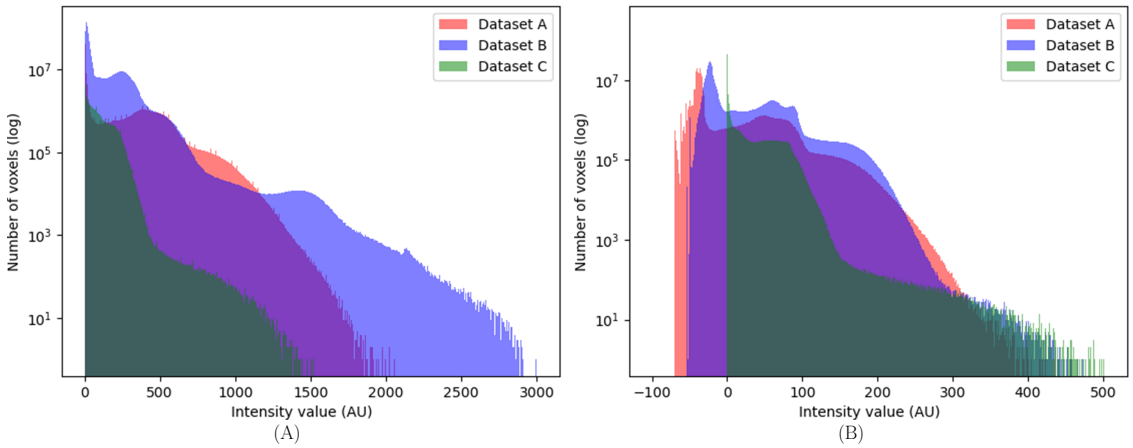


Figure 6.2: Histogram of image intensities of the original data distribution (A) and after Nyul normalization (B). The y-axis scale is in logarithms. (x-axis; AU: arbitrary units)

6.3.3 Image Preprocessing

Initially, we perform histogram matching normalization using the Nyul normalization technique [80]. Specifically, we train the Nyul normalizer using Dataset A. Subsequently, we apply the trained normalizer to Datasets A, B, and C ensuring that the data distributions become identical. Following this step, we resample Dataset B and C to match the voxel spacing of Dataset A, which is 0.94 × 0.94 × 2.00 mm. Figure 6.2 illustrates the data distribution both before and after the application of histogram matching and resampling techniques. The visual comparison demonstrates a greater degree of similarity between the data distributions following the implementation of these techniques. Moreover, the calculated Wasserstein distances between the datasets A and B prior to and post-normalization are 62.77 and 25.48, respectively. Similarly, the distances between datasets A and C pre- and post-normalization are 100.19 and 31.39, respectively.

For the purpose of network training, we apply multiple preprocessing steps to the MRI volumes. These steps include bias correction, skull-stripping, and rigid registration (brain images of the two time points). Additionally, we adjust the image intensities to be confined within the interval of [0, 100] and then rescale them to the range of [-1, 1]. As for the VGM maps (labels), they are truncated to fit within the range of [-1, 1], while any values falling within the range of [-0.01, 0.01] are set to 0.

6.3.4 Attention Mechanisms

In this work, we incorporate three attention modules into the U-Net architecture for improved VGM map prediction. These are described briefly in the following.

6.3.4.1 Attention Block from Attention U-Net

In their work, [58] introduced attention gates within the U-Net architecture. These attention gates serve as a mechanism to guide the network's decision-making process by selectively choosing relevant features while disregarding irrelevant ones. The authors achieved this by leveraging higher-level features as a guide to suppress trivial and noisy responses present in the lower-level skip connections. By incorporating attention gates, the network gains the ability to focus its attention on more informative features, thus enhancing its discriminative power and improving the overall performance of the U-Net model.

6.3.4.2 Squeeze and Excitation Block

The Squeeze and Excitation (*SE*) block was introduced by [59]. This block is designed to enhance the representational power of Convolutional Neural Networks (*CNNs*) by adaptively recalibrating feature maps. It consists of two main operations: squeezing and exciting. In the squeezing step, global spatial information is extracted by applying global average pooling to the input feature maps. This operation reduces the spatial dimensions of the feature maps. In the exciting step, the squeezed information is used to learn channel-wise dependencies and recalibrate the feature maps. This recalibration process enables the network to emphasize more informative channels and suppress less relevant ones, thereby improving the discriminative power of the network.

6.3.4.3 Convolutional Block Attention Module

The Convolution Block Attention Module (*CBAM*) was first developed by [60]. It consists of two attention sub-modules: the channel attention module (CAM) and the spatial attention module (SAM). The CAM captures interdependencies between channels by adaptively recalibrating feature maps based on channel-wise information (similar to the SE block). It employs a combination of global average pooling and fully connected layers to compute channel attention weights. The SAM, on the other hand, captures spatial dependencies by adaptively highlighting informative spatial locations within feature maps. It utilizes the max-pooling and average-pooling operations followed by convolutional layers to generate spatial attention weights. By integrating both channel and spatial attention, the CBAM module enhances the discriminative power of CNNs and allows them to focus on salient features during image classification or object detection tasks.

6.3.5 Network Architecture

We implemented three 3D U-Nets utilizing the aforementioned attention mechanisms to compute VGM maps from input volumes [28, 81]. These U-Nets are equipped with residual and skip connections to facilitate the seamless flow of information and gradients. The encoder/decoder structure consists of five levels, with the first two levels comprising two convolution layers each, and the subsequent three levels consisting of three convolution layers each. The number of filters starts at 8 at the initial level and progressively increases to 128 at the bottom level. The VGM prediction map is generated through a final convolution layer of size $1 \times 1 \times 1$,

while all other convolutions employ $3 \times 3 \times 3$ kernels. The inputs to the network consist of baseline and follow-up volumes.

We trained the following networks:

1. Attention U-Net: The U-Net architecture is enhanced with attention blocks in the decoder [58].
2. SE-Attention U-Net: The U-Net architecture incorporates SE blocks in the encoder and attention blocks in the decoder [59, 81].
3. CBAM-Attention U-Net: The U-Net architecture integrates CBAM blocks in the encoder and attention blocks in the decoder [60, 81].

Furthermore, we compare our trained networks' performance with the baseline U-Net [28] model obtained from the work of [108].

For a visual representation of these U-Net architectures, refer to Figure 6.3.

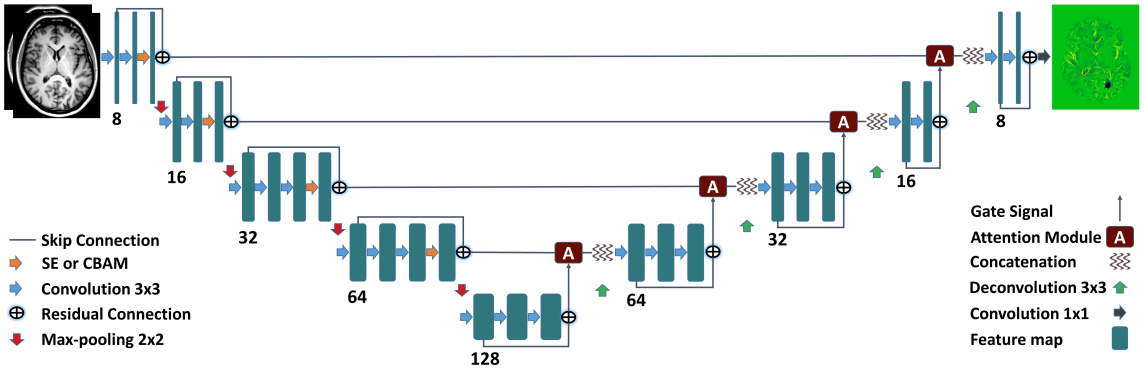


Figure 6.3: Proposed 3D U-Net incorporating attention mechanisms in the encoder and decoder parts. The attention mechanism from [58] is part of the decoder. The SE/CBAM blocks [59, 60] are part of the encoder of the U-Net.

6.3.6 Loss Function

The networks in our study are trained using a combination of MAE and gradient loss. The MAE loss, defined in Equation 6.1, calculates the average absolute difference between the predicted output $\hat{Y}$ and the ground truth Y:

$$\mathcal{L}_{MAE}\left(Y, \hat{Y}\right) = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|, \tag{6.1}$$

To further improve the training process, we incorporate the combination of MAE and gradient loss, which is described in Equation 6.2.

$$\begin{aligned}\mathcal{L}_{Grad}\left(Y, \hat{Y}\right) = &\frac{1}{N_x N_y N_z}\sum_{i,j,k}\left(|y_{i,j,k} - y_{i-1,j,k}| - |\hat{y}_{i,j,k} - \hat{y}_{i-1,j,k}|\right)^2 \\ &+ \frac{1}{N_x N_y N_z}\sum_{i,j,k}\left(|y_{i,j,k} - y_{i,j-1,k}| - |\hat{y}_{i,j,k} - \hat{y}_{i,j-1,k}|\right)^2 \\ &+ \frac{1}{N_x N_y N_z}\sum_{i,j,k}\left(|y_{i,j,k} - y_{i,j,k-1}| - |\hat{y}_{i,j,k} - \hat{y}_{i,j,k-1}|\right)^2\end{aligned} \tag{6.2}$$

The additional gradient loss term incorporates gradient information to guide the network's learning:

$$\mathcal{L}_{MAE+Grad}\left(Y, \hat{Y}\right) = \mathcal{L}_{MAE} + \lambda \cdot \mathcal{L}_{Grad} \tag{6.3}$$

Where, $\hat{y}_i$ and y_i represent the predicted and label voxel values, respectively. The total number of voxels in a batch is denoted by N, where N_x, N_y, and N_z represent the number of voxels along each dimension of the 3D MRI volume. Additionally, the parameter λ is set to one in Equation 6.2 to equally weight the two loss functions. We selected the $\mathcal{L}_{MAE+Grad}$ loss function based on its demonstrated effectiveness in predicting VGM maps, as reported in [108].

6.3.7 Training and Implementation

In our study, we employed a training approach using 3D patches of size 128 × 128 × 32. The patches were sampled randomly, with the constraint that their centers lie within the brain mask, and were oriented along the transverse plane. Each training batch consisted of 8 samples.

For optimization, we utilized the Adam optimizer with a learning rate of 10^{-3}. To mitigate overfitting, we applied L2 regularization with a weight of 10^{-10}. Each network underwent training for a total of 860 epochs in a 5-fold cross-validation scheme. The 5-fold cross-validation was performed on Dataset A, which was split into train:validation:test sets with a ratio of 54:2:15 cases.

To assess the generalizability of our trained models, we applied them to Datasets B and C. This additional evaluation aimed to determine how well the models could perform on an independent, previously unseen dataset. It is also worth noting that for the baseline state-of-the-art method, we use the settings described in the baseline approach [108] to make predictions unless specified otherwise. In one scenario, we followed the same preprocessing steps as described in the baseline approach. In another case, we implemented our preprocessing steps, including intensity truncation in [0, 100], Nyul normalization, and resampling, for inference using the baseline model.

The neural networks were trained using Tensorflow 2.3.0 [117] and Python 3.6.13, employing an Nvidia RTX A4000 as the GPU.

6.3.8 Evaluation

6.3.8.1 Quantitative Evaluation

To facilitate a more meaningful comparison, we utilize the same evaluation metrics described by [108]. These metrics allow us to assess the performance of our approach consistently.

The first metric we employ is the MAE, which quantifies the average absolute difference between the predicted and ground truth VGM map within the brain mask. This metric provides insight into the accuracy of the predicted VGM values at the voxel level. The second metric used for evaluation is the Structural Similarity Index

Measure (*SSIM*), a measure that assesses the similarity of structures between the predicted and ground truth VGM maps. The SSIM compares three components of images: luminance, contrast, and structure [118]. This metric provides information about the overall structural preservation in the predicted VGM map compared to the ground truth. We further utilize the Dice Similarity Coefficient (*DSC*), specifically for non-change regions within the brain mask. The DSC is calculated for voxel values falling within the range of [-0.01, 0.01] in both the predicted and ground truth VGM maps. This metric allows us to evaluate the similarity and overlap between these regions, further assessing the accuracy of the predicted VGM map. Finally, we perform a paired t-test to find a statistically significant difference (p-value < 0.05) between the results from the baseline and our proposed methods.

6.3.8.2 Qualitative Evaluation

Two expert neuroimagers (A.G., P.E.) performed a joined qualitative review by consensus. The predictions from our best-performing network were compared to conventional VGM maps (ground truth) and source T1-weighted data. The expert checked 5 patients from each dataset by comparing the predicted VGM and ground truth VGM in conjunction with the baseline and follow-up visits' MRI volumes. Based on the visual analysis, each patient's prediction in comparison to the ground truth VGM was categorized into 4 categories: 1. Missing information, where the prediction does not have enough details as compared to the ground truth, 2. loss of lesion to background contrast, where the lesion is lost to background and is not visible in prediction, 3. original result well presented, where the prediction is of similar quality to the ground truth, and 4. additional lesion details offered, where the predicted VGM gives additional lesion information that might not be present in the ground truth.

6.4 Results

6.4.1 Quantitative Results

The results for each dataset are described in Table 6.3. For Dataset A, the CBAM-Attention U-Net attains the highest SSIM, DSC, and MAE of 0.9177, 0.9814, and 0.0335, respectively. In comparison, the baseline state-of-the-art method obtains SSIM, DSC, and MAE of 0.9139, 0.9800, and 0.0350, respectively. For Dataset B, the best SSIM, DSC, and MAE are 0.9416 (Attention U-Net), 0.9882 (CBAM-Attention U-Net), and 0.0337 (SE-Attention U-Net), respectively. Moreover, each network's metric for both Datasets A and B surpasses the corresponding metric for the baseline method. Furthermore, each of the proposed networks outperforms the baseline significantly in the case of Datasets A and B (p-value < 0.05). For Dataset C, the baseline method outperforms other networks in SSIM (0.9102) and MAE (0.0422) metrics. However, the DSC of the baseline (0.9287) is considerably lower than our networks' DSCs (best:0.9784). The poor DSC of the baseline method is due to the network outputting more values close to 0 in larger regions. The visual results in Figure 6.4 for Dataset C also confirm this, showing that although the baseline has slightly better metrics, the outputs are not useful for a physician.

Dataset	Network	SSIM ↑	DSC ↑	MAE ↓
Dataset A	U-Net (baseline)	0.9139 ± 0.0216	0.9800 ± 0.0033	0.0350 ± 0.0112
	Attention U-Net	0.9172 ± 0.0213	0.9812 ± 0.0032	0.0338 ± 0.0107
	CBAM-Attention U-Net	**0.9177 ± 0.0212**	**0.9814 ± 0.0031**	**0.0335 ± 0.0105**
	SE-Attention U-Net	0.9172 ± 0.0212	0.9810 ± 0.0031	0.0337 ± 0.0106
Dataset B	U-Net (baseline)	0.9207 ± 0.0187	0.9816 ± 0.0034	0.0364 ± 0.0064
	Attention U-Net	**0.9416 ± 0.0143**	0.9881 ± 0.0023	0.0344 ± 0.0052
	CBAM-Attention U-Net	0.9406 ± 0.0144	**0.9882 ± 0.0023**	0.0351 ± 0.0053
	SE-Attention U-Net	**0.9416 ± 0.0143**	0.9880 ± 0.0022	**0.0337 ± 0.0052**
Dataset C	U-Net (baseline)	**0.9102 ± 0.0377**	0.9287 ± 0.0093	**0.0422 ± 0.0091**
	Attention U-Net	0.9097 ± 0.0288	0.9780 ± 0.0041	0.0465 ± 0.0087
	CBAM-Attention U-Net	0.9010 ± 0.0287	**0.9784 ± 0.0042**	0.0519 ± 0.0112
	SE-Attention U-Net	0.9091 ± 0.0291	0.9780 ± 0.0042	0.0459 ± 0.0086

Table 6.3: Results for each Dataset from networks trained on Dataset A only. The metrics for Dataset A are congregated from the results of 5-fold training. The metrics for Datasets B and C are calculated using an ensemble of 5 models that were trained on Dataset A. The baseline state-of-the-art method is outperformed in each metric for Datasets A and B by CBAM- or SE-Attention U-Net. However, for Dataset C, the SSIM and MAE of the baseline are higher than our approach. But as can be seen in Figure 6.4, the baseline method outputs VGM without any details. The MAE in this case is lower as most of the predicted values from the baseline method are close to 0. Further, the DSC of 0.92 is comparatively lower than the average 0.98 DSC from other methods. Numbers in bold signify the best metric for each dataset for each category. Underlined values depict significantly better metrics in comparison to the corresponding baseline metric with p-value < 0.05.

Dataset	Network	SSIM ↑	DSC ↑	MAE ↓
Dataset A	U-Net (baseline)	0.9071 ± 0.0227	**0.9809 ± 0.0034**	0.0379 ± 0.0120
Dataset B	U-Net (baseline)	**0.9396 ± 0.0144**	**0.9879 ± 0.0023**	**0.0348 ± 0.0054**
Dataset C	U-Net (baseline)	0.9063 ± 0.0300	**0.9778 ± 0.0042**	0.0467 ± 0.0097

Table 6.4: Results from the baseline method (ensemble) after applying image preprocessing described in our study. As can be seen, the MAE for Dataset B improves, however, for Datasets A and C, it gets worse. The DSC for Dataset C increases by 6% as compared to the result in Table 6.3. The numbers in bold depict comparatively higher metrics to the corresponding baseline method from Table 6.3.

Furthermore, in Table 6.4, we depict results from the baseline method after applying the preprocessing methods from our approach (Section 6.3.3; i.e. truncation in [0, 100] range instead of baseline truncation in [200, 700] range, Nyul normalization, and resampling). In this case, the method improves on all metrics for Dataset B as compared to the corresponding result in Table 6.3. However, for Datasets A and C, SSIM and MAE numbers are lower in comparison. Noteworthily, the DSC of Dataset C improves and reaches 0.9778. After applying our preprocessing approach,

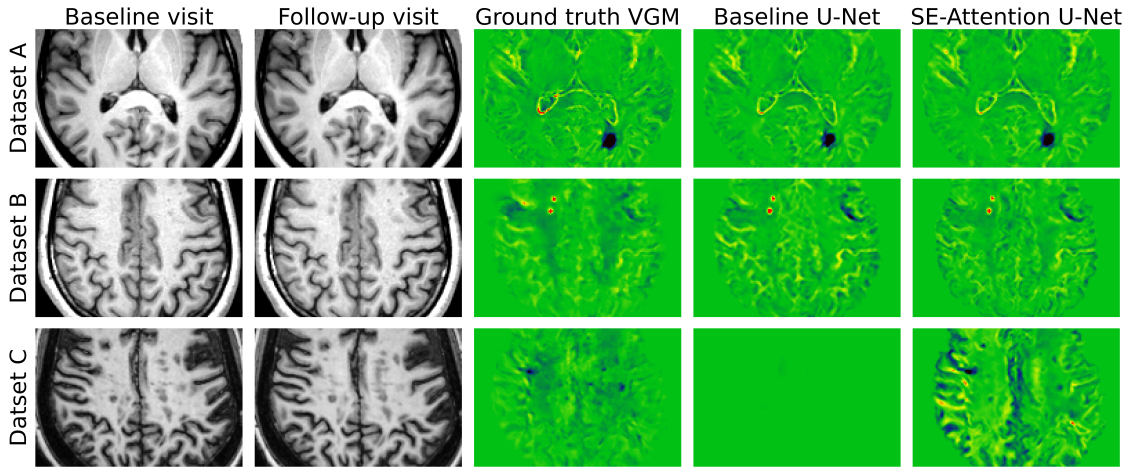


Figure 6.4: Top row: Dataset A, middle row: Dataset B, and bottom row: Dataset C. Qualitative results for each dataset from the baseline U-Net and our proposed approach with SE-Attention U-Net.

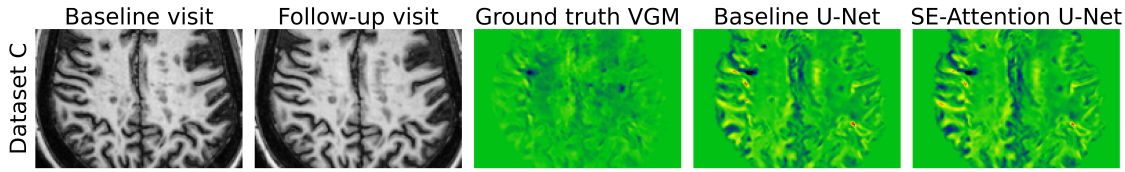


Figure 6.5: Visual result from the baseline U-Net for Dataset C (for the same patient from Figure 6.4) after applying our approach's preprocessing steps.

Dataset	Missing information	Loss of lesion to background contrast	Original result well presented	Additional lesion detail offered
Dataset A	0/5	0/5	5/5	0/5
Dataset B	0/5	0/5	5/5	1/5
Dataset C	0/5	0/5	5/5	3/5

Table 6.5: Visual inspection result by an expert neuro-radiologist for SE-Attention U-Net predictions compared to ground truth VGM, baseline and, follow-up MRIs.

the baseline network is able to produce meaningful predictions for Dataset C as depicted in Figure 6.5, suggesting that our preprocessing method is important for generalizability in this case.

The prediction of each VGM map takes approx. 2.75 seconds. With the inclusion of preprocessing, the total time taken for VGM prediction is about 4 minutes (same as the SOTA method [108]).

6.4.2 Qualitative Results

Table 6.5 shows the qualitative analysis result of five cases of each dataset that were visually analyzed by the experts. Since it achieved the best mean MAE score of 0.0377 across all the datasets, we selected SE-Attention U-Net as the best network for visual analysis. None of the predictions (0/15) showcased any missing information details. There is also no loss of lesion to background contrast (0/15) for any case

in each dataset. Furthermore, all the analyzed predictions (15/15) show that they present the ground truth VGM well. Interestingly, one case in Dataset B (1/5) and three cases in Dataset C (3/5), show additional lesion details in comparison to the ground truth. However, for dataset A, the predicted VGMs do not offer (0/5) any additional lesion details in comparison to the ground truth.

Visual results are depicted in Figure 6.4 and Figure 6.5. In Figure 6.4, it can be seen that for Datasets A and B, the prediction from SE-Attention U-Net is similar to the ground truth, having dark (Dataset A (n=1)) and bright spots (Dataset B (n=2)) for changes in lesions in the same regions. For Dataset C, we show an example case for which our prediction offers better detail and more lesion information as compared to the ground truth. However, in this case, the baseline prediction is worse and does not show any lesion details. Interestingly, when we swap the baseline method's preprocessing with our preprocessing approach, the output is more informative and depicts VGM in higher quality (see Figure 6.5).

6.5 Discussion

We aimed to develop a generalizable approach to predict VGM maps for the longitudinal assessment of MS patients. Our work builds upon previous research [108] by addressing the crucial issue of generalizability. Some interesting aspects emerge from this work. The VGM maps help to detect subtle changes in lesions between baseline and follow-up visits. Our approach did calculate VGM maps in a short time and across three different datasets (2/3 unseen datasets) with high accuracy.

In the development process, we integrated advanced deep-learning techniques, image preprocessing, and careful model evaluation. Several steps were performed, that appeared useful and were able to improve the process incrementally. We began by describing the image preprocessing steps, which involved histogram matching normalization and voxel spacing resampling to ensure data consistency across different imaging centers. These preprocessing steps are crucial for enhancing the model's ability to generalize to diverse datasets. Our deep learning model is based on a 3D residual U-Net architecture, which incorporates attention mechanisms to highlight salient regions in the brain volumes. The application of attention mechanisms in MS lesion change detection is warranted as they have been shown to enhance lesion detection algorithms in previous works [119, 120].To evaluate the effectiveness of our approach, we conducted extensive experiments using a dataset of 71 longitudinal MRI brain volumes of MS patients. We compared our model's performance to the SOTA method [108]. Additionally, we evaluated the generalizability of our model on two unseen datasets, to test its robustness and potential for broader clinical applicability.

In our approach, we show that across all the datasets, the method attains SSIM and DSC higher than 0.91 and 0.98, respectively. Furthermore, the differences in the MAEs for each dataset are not considerably high. The current state-of-the-art (baseline) for predicting VGM maps using deep learning, as proposed by [108], involves the implementation of a U-Net model without attention mechanisms. This baseline method does not adequately address the generalizability across different scanners and sites, a limitation our approach seeks to overcome. Employing the

trained model from [108] on Dataset A, we obtained an SSIM of 0.9139, a DSC of 0.9800, and an MAE of 0.0350 (Table 6.3) [1]. In comparison, each of our proposed networks surpassed the state-of-the-art result for Dataset A. Similarly, we extended the baseline to Datasets B and C and found that our approach outperforms the baseline for Dataset B. However, for Dataset C, the baseline attained higher SSIM and MAE scores (Table 6.3). In contrast, the baseline DSC in this case is considerably lower (0.92 v/s 0.97). When visually analyzed (Figure 6.4), we found that the baseline result could not replicate the ground truth VGM and contained values very close to 0 (therefore lower DSC), and hence it attained better MAE in comparison. In this case (Dataset C), the numbers did not reflect the results visually and were deemed not useful in a clinical setting. However, when we applied our preprocessing approach (i.e. Nyul normalization + truncation in [0, 100] range), the baseline method yielded a much more convincing VGM map as seen in Figure 6.5. This suggests that the preprocessing method is key to the generalizability of deep VGM maps. The resulting average MAE is 0.0467 (Table 6.4), which is worse than our SE-Attention U-Net. Furthermore, we found that for Datasets B and C, the predicted VGM maps could offer extra lesion details in comparison to the ground truth maps without the loss of important information in 4/15 cases (Table 6.5). This also suggests that the ground truth VGM maps of Dataset A are of higher quality and training on them could help the network learn high-quality features.

Moreover, there exist multiple automated new lesion segmentation algorithms based on deep learning [120–127]. These methodologies primarily leverage the MSSEG-2 dataset [128], encompassing FLAIR images from baseline and follow-up visits for each patient, either with or without the integration of synthetic data. Notably, these approaches focus on segmenting new lesions during follow-up visits. In contrast, our proposed approach distinguishes itself by its fundamental objective: quantifying the change in lesion activity between baseline and follow-up measurements. Unlike the aforementioned methods, which aim to delineate new lesions, our approach offers a distinctive perspective, providing quantitative insights into the variations in lesion size between visits. This ability to quantitatively showcase the decrease or increase in lesion activity enhances the depth and specificity of our methodology in assessing lesion dynamics over time.

Nevertheless, methodologies analogous to ours have been proposed in the literature, specifically targeting the identification of change maps in MS patients between baseline and follow-up examinations. [129] presented an algorithm that concurrently optimized image registration and local intensity change detection within FLAIR volumes.[130] computed lesion changes utilizing T1, T2, and FLAIR sequences. Their approach involved estimating a dissimilarity map between two visits and subsequently incorporating logistic regression with neighborhood information and local texture descriptors. It is noteworthy that a direct comparison between our approach and these existing methodologies is challenging due to the fact that both

[1] In our study, discrepancies were observed in the metric values for the baseline method when compared to those reported in [108]. Despite employing the same trained models and adhering to the preprocessing protocol detailed in their paper, the variation in results could be attributed to small details in their methodology that were not described in their paper. Additionally, the divergence in outcomes may also stem from variances in the versions of the libraries used between the two studies.

aforementioned approaches utilize lesion segmentation maps as the ground truth for evaluation. In contrast, our study necessitates expert annotation and an optimal threshold for generating lesion maps (derived from VGM maps) to facilitate segmentation evaluation. Such a comparative task extends beyond the scope of the current work.

Addressing the challenge of MRI data heterogeneity across sources is essential for the widespread adoption of deep learning-based tools in clinical practice, and our model's independence from MRI sources demonstrates its potential as a versatile clinical asset. Furthermore, the high accuracy and generalizability of our deep learning approach hold great promise for clinical practitioners, as it offers a valuable tool for detecting and monitoring subtle changes in MS lesions, facilitating more informed treatment decisions. The ability to identify even the most discreet changes in brain structures could significantly impact the clinical management of MS patients, potentially leading to earlier interventions and improved patient outcomes.

Nonetheless, our approach has a few limitations. Firstly, it needs accurately registered brain volumes for baseline and follow-up visits. If the registration is of low quality, then the VGM maps will be less accurate and might display less precise information. However, we found that all the cases in our study were registered with high quality. Another limitation could be the (partial) loss of lesion when resampling datasets to have the same spacing as Dataset A. Dataset A has anisotropic spacing where the slice thickness is 2 mm. Resampling a higher-resolution MRI volume to a 2mm slice thickness could result in partial volume effects, i.e., loss of some detail. In this case, it would mean that the VGM maps might be less precise in detecting some subtle lesion changes.

In future work, we will analyze if the VGM maps can be produced from a single scan of MS patients with a comparison scan from an age- and sex-matched group of healthy individuals for detecting lesions. Furthermore, we would also test our algorithm for detecting structural changes in other neurological diseases such as stroke, neurodegenerative diseases, or brain tumors.

In conclusion, we present a generalizable approach that can produce VGM maps in a fast and robust manner across datasets from various sources. Our algorithm can be helpful in detecting subtle lesion changes in brains of MS patients.

Conflict of Interest Statement

A.D. and M.K. are employed by VGMorph GmbH. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This research project is funded by the Ministry of Economic Affairs Baden Württemberg within the framework "KI für KMU".

7. Streamlining acute Abdominal Aortic Dissection management - an AI based CT imaging workflow

Anish Raj [1,2], Ahmad Allababidi [3], Hany Kayed [3], Andreas LH Gerken [4], Julia Müller [5], Stefan O. Schönberg [3], Frank G. Zöllner [1,2] and Johann S. Rink [3]

[1]Computer Assisted Clinical Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany
[2]Mannheim Institute for Intelligent Systems in Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany
[3]Department of Radiology and Nuclear Medicine, University Medical Center Mannheim, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany
[4]Department of Surgery, University Medical Center Mannheim, Medical Faculty Mannheim, Heidelberg University, Mannheim, Baden Württemberg, Germany
[5]mediri GmbH, Heidelberg, Baden Württemberg, Germany

7.1 Abstract

Life-threatening acute Aortic Dissection (*AD*) demands timely diagnosis for effective intervention. To streamline intrahospital workflows, automated detection of AD in abdominal Computed Tomography (*CT*) scans seems useful to assist humans. We aimed at creating a robust Convolutional Neural Network (*CNN*) based pipeline capable of real-time screening for signs of abdominal AD in CT.

In this retrospective study, abdominal CT data from AD patients presenting with AD and from non-AD patients were collected (n: 195, AD cases: 94, mean age: 65.9 years, female ratio: 35.8%). A CNN-based algorithm was developed with the goal of enabling a robust, automated, and highly sensitive detection of abdominal AD. Two sets from internal (n=32, AD cases: 16) and external sources (n=1189, AD cases: 100) were procured for validation. The abdominal region was extracted, followed by the automatic isolation of the aorta Region of Interest (*ROI*) and highlighting of the membrane via edge extraction, followed by classification of the aortic ROI as dissected/healthy. A 5-fold cross-validation was employed on the internal set, and an ensemble of the 5 trained models was used to predict the internal and external validation set. Evaluation metrics included Area Under the Curve (*AUC*) and balanced accuracy.

The AUC, balanced accuracy, and sensitivity scores of the internal dataset were 0.932 (CI: 0.891-0.963), 0.860, and 0.885, respectively. For the internal validation dataset, the AUC, balanced accuracy, and sensitivity scores were 0.887 (Confidence Interval (*CI*): 0.732-0.988), 0.781, and 0.875, respectively. Furthermore, for the external validation dataset, AUC, balanced accuracy, and sensitivity scores were 0.993 (CI: 0.918-0.994), 0.933, and 1.000, respectively.

The proposed automated pipeline could assist humans in expediting acute aortic dissection management when integrated into clinical workflows.

7.2 Introduction

Acute aortic syndrome (AAS) consists of the life-threatening conditions of AD, intramural hematoma (IMH) and penetrating atherosclerotic ulcer (PAU) [131]. In acute AD, tearing of the aortic vessel intima leads to uncontrolled blood inflow into the aortic wall. Incidence has been reported to range from 2.6 to 7.2 cases per 100,000 person-years and is associated with a reported in-hospital mortality of 30.1% in women and 21.0% in men, creating a substantial healthcare burden [132].

The event of acute AD can cause severe abdominal or back pain, however, in many cases symptoms are nonspecific and driven by secondary complications like visceral ischemia, resulting in a relatively high rate of patients not initially being suspected of AD, thus receiving abdominal imaging for other reasons [133]. Modern medical management of acute AD aims at early CT-based diagnosis and prompt therapy stratification. Whereas thoracic AD including the ascending part of the vessel (type A) typically requires immediate intervention, AD located more distally in the aorta (type B) in the absence of complications like rupture, malperfusion of visceral organs, spinal ischemia or lower limb ischemia [134] may be managed conservatively or with elective intervention. In CT imaging, acute AD can in many cases be distinguished from its chronic form [135], bearing the risk of development of Abdominal Aortic Aneurysm (*AAA*), representing a chronic risk for rupture.

In many cases, the unspecific clinical symptoms and even radiological misdiagnosis under emergency conditions which were reported at 35% for type A AD and 17% for type B AD in a British setting [136] possibly compromise timely suspicion, diagnosis, and treatment of this rare but time-critical medical condition in a substantial proportion of patients [137].

Multiple factors have been addressed to streamline AD management workflows [138]. Artificial Intelligence (*AI*)-based techniques have been described as promising tools, capable of detecting critical findings, prioritizing cases accordingly and eventually reducing delays [139]. Multiple groups have aimed to create algorithms capable of detecting AD, mainly in thoracic CT scans. Yi et al. [140] developed a method using a 2.5D U-Net to extract aorta masks and subsequently used a 3D ResNet34 CNN [141], pre-trained on MedicalNet [142], for feature extraction and final prediction via a Gaussian Naive Bayes algorithm by combining radiomics and CNN features. The results showed high performance on internal and external datasets (AUC=0.948, sensitivity=0.862, specificity=0.923 for internal (341 patients); AUC=0.969, sensitivity=0.978, specificity=0.554 for external (111 patients)). Despite their robust results on the internal set, they attained low specificity on a small external dataset. Hata et al. [143] utilized a 2D Xception architecture [144], pre-trained on ImageNet [145], to classify AD in non-contrast-enhanced CT scans (170 patients). They achieved an AUC of 0.940, sensitivity of 0.918, and specificity of 0.882 by classifying consecutive slices of the aorta. However, their study's limitation lies in the lack of external validation, raising concerns about the generalizability of their model.

Huang et al. [146] proposed a 2-step hierarchical model involving an attention U-Net for initial AD detection (AD case; if 5 or more slices were detected with AD) followed by a ResNext model [147] for Stanford type classification. Their internal results showed excellent performance (AUC=0.980, recall=0.960, specificity=1.000 for AD detection; AUC=0.950, recall=0.947, specificity=0.953 for Stanford types). Despite high internal metrics, their approach was not tested on an external dataset, questioning its robustness. Harris et al. developed a 2D five-layer CNN as a screening algorithm, yielding a sensitivity of 0.878 and a specificity of 0.960 [148] with a training dataset of 778 patients, and demonstrated a reduction of turnover time (395s) by prioritizing worklist in a teleradiology setting in the US. Cheng et al. employed a U-Net to first segment the aorta and then analyze its circularity for AD classification. They obtained a sensitivity of 0.900 and specificity of 0.800 on a comparably small Chinese dataset of n=20 patients (10 AD) [149]. Yellapragada et al. described a wider focussed 3D deep learning model for the detection of AAS solely in CTA scans, trained on a dataset of 3500 cases (500 containing AAS, unclear number of AD cases), overall yielding promising results [150]. Guo et al. [151] segmented the aorta ROI manually and extracted 396 radiomic features, including texture features, gray-level co-occurrence matrix, gray-level run-length matrix, gray-level size zone matrix, form factor features, and histograms. They selected the top 20 features using max-relevance min-redundancy and constructed a radiomic signature via LASSO logistic regression, followed by logistic regression for classification. Their results indicated consistent performance on internal (304 patients, AUC=0.92, recall=0.941, specificity=0.753) and external (74 patients, AUC=0.90, recall=0.857, specificity=0.917) datasets. It is noteworthy that their method involves manual segmentation of the aorta, which is time-consuming. Manual segmentation diminishes the benefits of an automated AD detection algorithm, as the time spent could instead be used by a radiologist to identify dissection cases directly. Additionally, the unavailability of a large external dataset in these studies underscores the necessity of developing robust algorithms that can generalize well across diverse (external) datasets. Furthermore, in the medical domain, the availability of large annotated datasets (for training) is particularly challenging due to factors such as patient privacy concerns, the extensive time required for expert annotation, and the variability in imaging protocols across institutions. Hence, producing a reliable model using a small internal dataset that can be effectively validated on a larger external dataset without manual segmentation is vital.

The aim of this analysis was to develop an easy-to-train CNN based AI pipeline on a small dataset that can be validated on a large external dataset and capable of real-time screening of all intrahospital abdominal CT scans for signs of AD, even at peak times and independent of acquired contrast phase. High robustness, efficient use of computing power, and a high degree of automation are therefore key requirements.

7.3 Materials and Methods

Possible conflicts of interest have been stated elsewhere. This retrospective study was approved by the local institutional review board (2021-635). Written informed consent was not required due to the retrospective nature of the study population enrolled.

7.3.1 Data Collection

7.3.1.1 Internal Training and Validation Dataset

Patients presenting with acute AD with abdominal extent between 2010/01/01 and 2021/03/01 were identified in the Radiology Information System (RIS) and included in the study, if an abdominal CT exam had been performed, intentionally regardless of contrast phase or other parameters like image quality or presence of metallic interferences. Patients with preexisting AD and scans not covering the entire abdomen were excluded. To include AD-negative cases, studies from patients with matching contrast phases and patient characteristics were collected similarly, of which some have already been used in a previous, different study [152], representing patient data overlap (n=85). Datasets were randomly split into training (80%; 163 patients) and validation (20%; 32 patients) datasets. The validation set is used only for testing, and we call it internal validation dataset. Patient-specific data was removed.

7.3.1.2 External Validation Dataset

Anonymous external non-synthetic CT data containing healthy aortic vessels and AD were created from the publicly available ImageTBAD [153], AVT [154], and Abdomenct-1k collections [155]. To avoid data overlap, we discarded the 20 KiTS19 patients from the AVT dataset, since these were included in the Abdomenct-1k set. Example images are displayed in Figure 7.1.

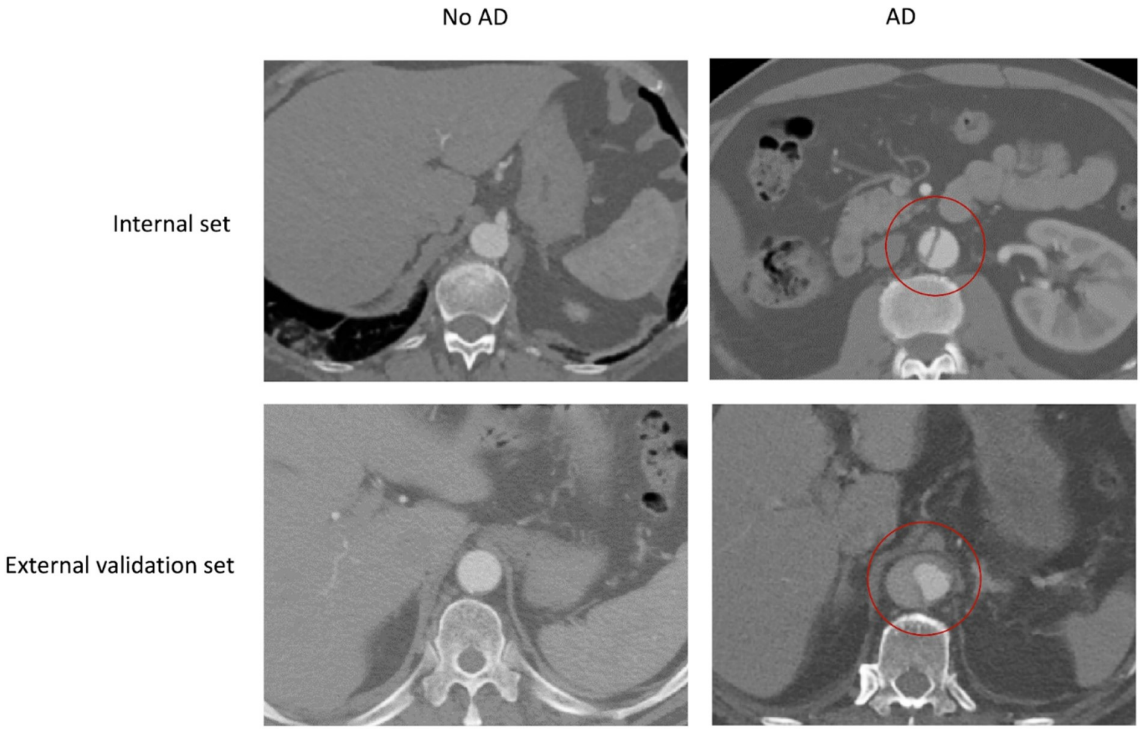


Figure 7.1: Example images from internal and external datasets. The left column is healthy cases and the right column is AD cases.

7.3.2 Image Annotation Strategy

Annotation was performed in Aycan Osirix (Aycan Digitalsysteme, Würzburg, Germany) by an attending radiologist (5 years of experience in cardiovascular imaging)

and a separate second reading by a resident (4 years of experience). Cases with disagreements were resolved by a senior physician with >15 years of experience. Annotation contained age, sex, CT date, contrast phase, presence of metallic interferences, presence of dissection (small and subtle AD were annotated as positive cases), exact dissection extent, occlusion or dissection of side branches (celiac trunk, superior and inferior mesenteric artery, renal arteries, iliac arteries), and presence of signs of visceral ischemia, intramural haematoma or aortic thrombosis. The external dataset and its annotations were validated.

7.3.3 Data Pre-Processing

We extract the abdominal region from each CT case using a heuristic implemented in [152]. The algorithm analyzes Hounsfield Unit (*HU*) distribution along the z-axis in the soft tissue HU range to establish the upper and lower bounds of the abdomen, determining the abdomen center using high HU values, and extracting a subvolume around it. The extracted subvolumes are resampled to a size of 320 x 384 x 224 voxels with a uniform spacing of 0.9 x 0.9 x 1.5 mm^3. The intensities are windowed in the [-200, 400] HU range, corresponding to the soft tissue domain. Next, the CT image is rescaled to be in the range [0, 1]. Subsequently, the mask of the abdominal aorta including iliac arteries is extracted using the TotalSegmentator algorithm [156]. To compensate for the insufficient segmentation quality of dissected vessels, aorta masks were dilated in 10 iterations via a square structuring element with a connectivity of one, ensuring complete coverage. The CT image is then masked using the dilated mask. The dissected aorta has a membrane separating the true and false lumen, which was highlighted by extracting edges (canny detector ($\sigma = 2$)) inside the aortic ROI. The subvolume voxels that do not belong to the edges are weighed down with a factor of 0.6, while the edge voxels remain unchanged. We then extract the bounding box of weighted aorta ROI using the dilated mask with a boundary margin of 5 voxels in each direction. Finally, the bounding box is resized with padding or cropping to a standardized size of 224 x 224 x 224 voxels (CNN input) in order to standardize data and minimize the usage of memory. The subvolume of this size is sufficient to encompass the entire aorta ROI (as seen in Figure 7.2). The aorta mask is used to create a bounding box, which centers the aorta ROI within this subvolume. The preprocessing pipeline is depicted in Figure 7.2. Further data augmentation details are described in the supplemental material (Section 11.2).

7.3.4 Network Architecture

We construct a CNN comprising 5 convolutional blocks and 1 dense block, depicted in Figure 7.3. The network takes as input the masked, edge-weighted aortic ROI. Each convolution block entails a 3 x 3 x 3 convolution, followed by instance normalization, dropout, and Rectified Linear Unit (*ReLU*) activation. Dropout probabilities for the initial to final convolution blocks are 0.00, 0.05, 0.10, 0.10, and 0.10, correspondingly. Post each of the initial four convolution blocks, a 3D max-pooling operation is performed to downsample the feature maps with a stride of 2. The final feature vector of size 256 is produced using a 3D average-pooling operation with a kernel size of 14. Lastly, the dense block processes this vector, yielding network output through a dense layer followed by sigmoid activation, i.e., a probability score

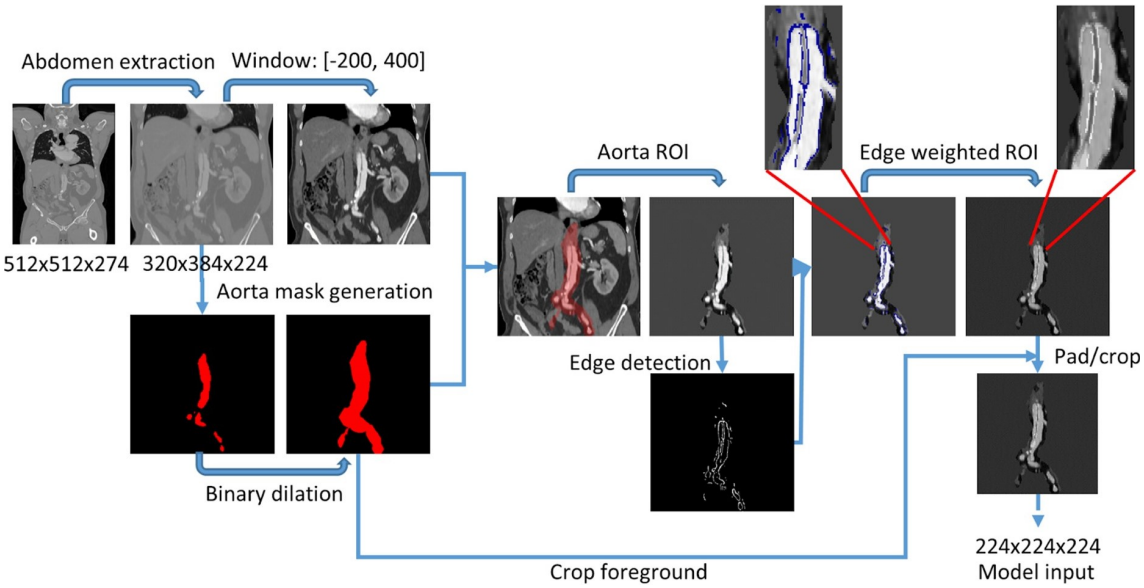


Figure 7.2: CT volume preprocessing pipeline. The abdominal region is automatically extracted and resized. An aorta mask is generated from this region and then dilated to mask out the aorta ROI from the abdomen region. For highlighting aorta bifurcation in AD cases, edges are extracted using the canny detector. The aorta ROI is weighted down by a factor of 0.6 where there is no edge voxel present. Finally, the edge-weighted aorta ROI is standardized by cropping / padding with the bounding box of the dilated aortic mask.

that a patient belongs to either AD or non-AD class. This network was selected after testing various other networks, whose details are provided in the supplements and results.

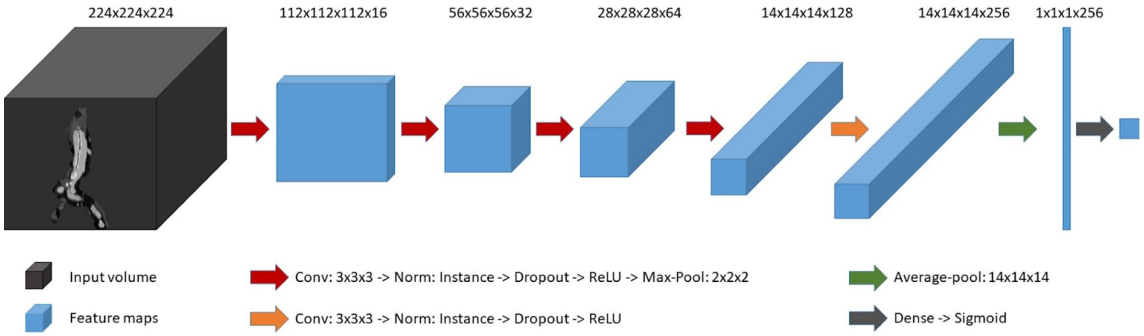


Figure 7.3: Network architecture for AD classification. The input volume (edge-weighted aorta ROI) is processed with a CNN to produce a single output in the range [0, 1]. The network consists of convolutional blocks with convolutions of size three, followed by instance norm, dropout, ReLU activation, and max-pool of size two (except in the last convolution block (orange block)). The final feature vector is produced by an average-pooling. It is then processed by a dense layer and a sigmoid activation to produce the output.

7.3.5 Evaluation

To assess the performance of all different networks, the AUCs were compared to choose the best-performing network for further evaluation. Next, the best network's performance was further assessed using sensitivity, specificity, balanced accuracy ([Sensitivity + Specificity]/2; suited for high class imbalance), and the AUC with a 95% confidence interval. We select the binary threshold of our model based on the high sensitivity and balanced specificity on the internal cross validation dataset, i.e. 0.45. Next, we create an ensemble of five models trained on this set to predict the internal validation and external sets based on our set binary threshold. The ensemble combines probabilities (average) for each case in the internal validation and external sets, and we evaluate its performance using the same threshold.

7.4 Results

7.4.1 Patient and Dataset Characteristics

From the total of n=266 cases, n=163 scans were used for training and n=32 were used for validation, in the rest, either images were not available, or the abdominal aorta was not fully covered. AD patients on average were 65.9 ± 13.5 (29-93) years of age, 36% were female. For the training cases, in 72.0%, suprarenal AD was present, and infrarenal AD was seen in 79.3%. 9.8% presented with intramural haematoma, and 56.1% showed partial aortic lumen thrombosis. The external public dataset used for validation consisted of a dataset with n=1189 cases (100 AD cases). Details are provided in Table 7.1 and patient selection criteria are shown in Figure 7.4.

	Internal set	**Internal validation set**	**External validation set**
Data source	Mannheim University Hospital, Germany		ImageTBAD, AVT dataset, Abdomen CT-1k dataset. Details provided in the supplements.
Patient numbers	163 (78 AD)	32 (16 AD)	1189 (100 AD)
Patient age (years)	65.9 ± 13.5 (29-93)		AD: 52.20 ±11.30 Non-AD: unknown
% female	35,8%		AD: 40,7% Non-AD: unknown
Contrast phase	CTA 73.1%; venous: 26.9%		mixed contrast
in-plane resolution (X/Y) (mm)	0.845 ± 0.106	0.845 ± 0.094	0.817 ± 0.120
Slice thickness (Z) (mm)	2.006 ± 1.341	1.836 ± 1.306	2.565 ± 1.538
Original image size (voxels)	512x512x495 ± 0x0x398	512x512x459 ± 0x0x228	512x514x224 ± 8x22x220

Table 7.1: Patient characteristics from internal and external set.

7.4.2 Classification Results

From the four tested networks, our 5-layer CNN yielded the highest AUC (0.932) in cross-validation (Table 7.2) and was therefore chosen for further investigation. Its performance on internal and external validation sets are provided in Table 7.3. Furthermore, the confusion matrices and the AUC curves are illustrated in Figure 11.1

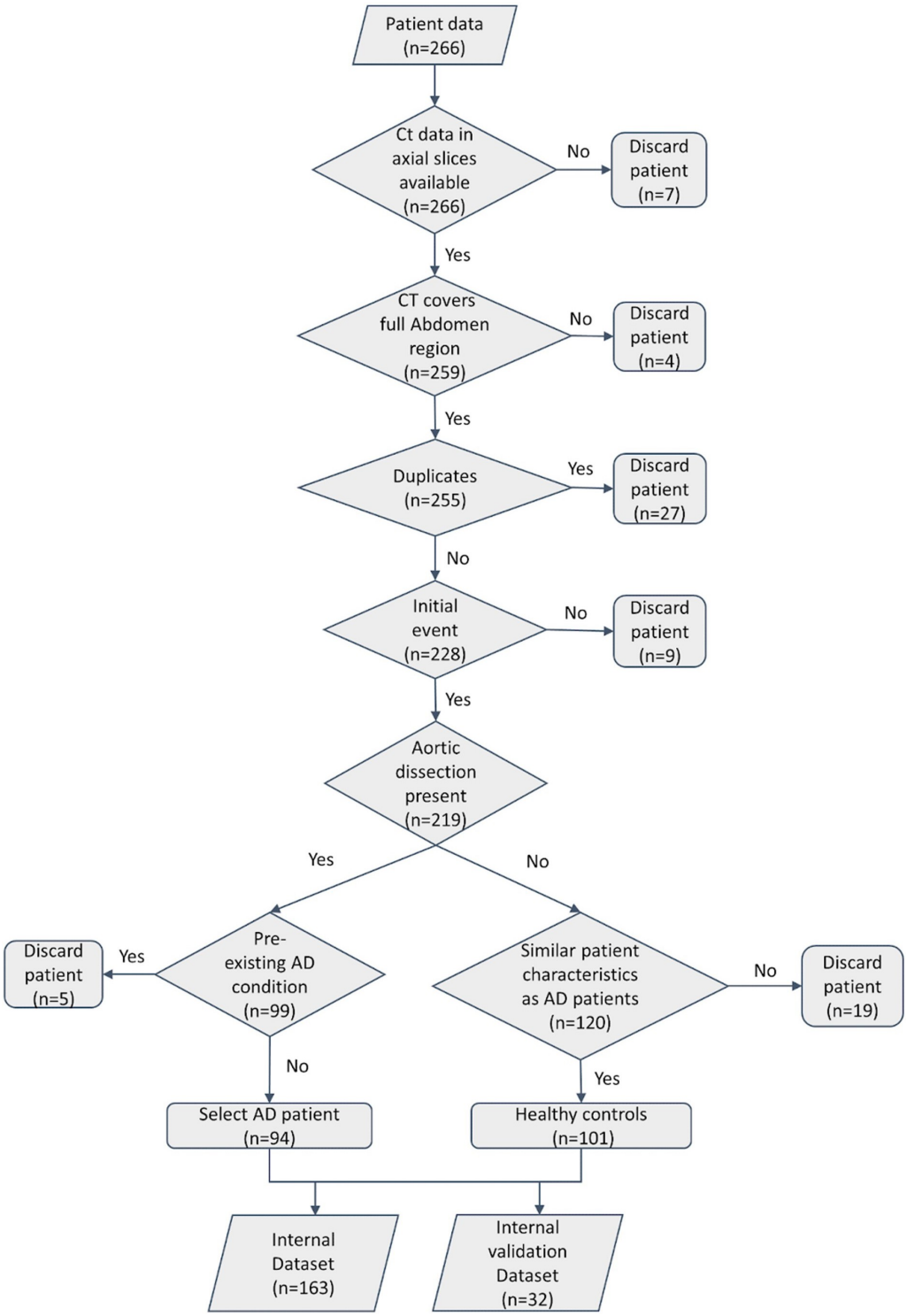


Figure 7.4: Internal dataset patient selection flow-chart. Data acquisition process. The healthy controls (n=101) were collected similarly to the included AD cases (n=94).

and Figure 7.5, respectively. The entire workflow including preprocessing, aorta segmentation, and the prediction was tested on internal training (cross-validation) and validation dataset and takes approximately 45 seconds (model prediction $\approx$ 0.003 seconds).

Network	5-layer CNN (ours)	ResNet10	ResNet34 (pretrained)	SEResNet50
AUC	0.932	0.796	0.520	0.869

Table 7.2: Performance comparison of four different networks on the internal training dataset.

Dataset	Sensitivity	Specificity	Balanced accuracy	AUC (95% CI)
Internal (cross validation)	0.885 (69/78)	0.835 (71/85)	0.860 (1.720/2)	0.932 (0.891-0.963)
Internal (validation)	0.875 (14/16)	0.688 (11/16)	0.781 (1.563/2)	0.887 (0.732-0.988)
External (validation)	1.000 (100/100)	0.865 (942/1089)	0.933 (1.865/2)	0.993 (0.988-0.997)

Table 7.3: Evaluation metrics for internal and external sets. The internal set results are from 5-fold cross-validation test sets, while the external set results are from an ensemble of 5 models that were trained on the internal set. Furthermore, a separate internal validation set results from the ensemble of 5 models is shown.

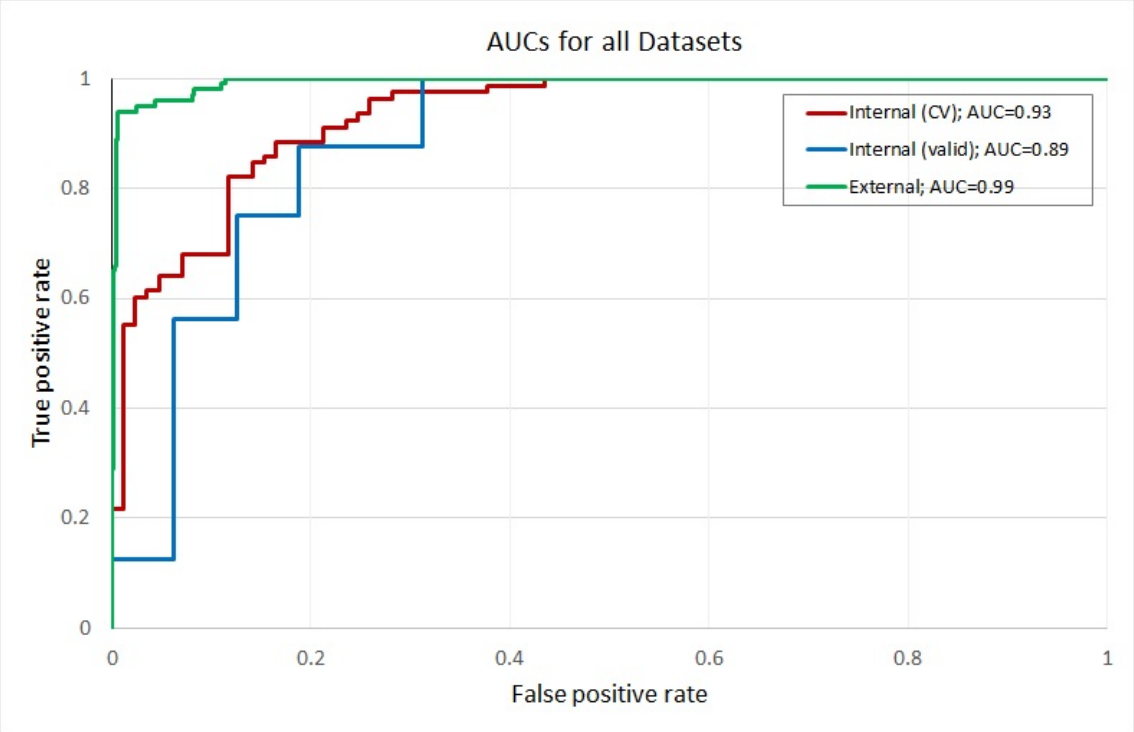


Figure 7.5: AUC curves for internal cross validation (CV), internal validation (valid) and external sets. The internal set AUC is 0.93, internal validation set AUC is 0.89, while the external set AUC is 0.99.

7.4.2.1 Internal Set

The balanced accuracy score for the internal set 5-fold cross-validation (test sets only) is 0.860 (Table 7.3). The corresponding sensitivity value is 0.885 (69/78), with the specificity being 0.835 (71/85) and the AUC of 0.932 (CI: 0.891 - 0.963), which is shown in the confusion matrix (Figure 11.1 (a)) and plotted in Figure 7.5 (red). From the nine missed AD cases, six had a clearly visible AD and three

presented with very subtle signs of AD (further described in supplement material and Figure 11.4).

7.4.2.2 Internal Validation Set

The sensitivity and specificity for the internal validation set are 0.875 (14/16) and 0.688 (11/16), respectively, with a balanced accuracy of 0.789. The AUC value obtained is 0.887 (CI: 0.732 - 0.988), as shown in Figure 11.1 (b) and Figure 7.5 (blues). Out of the five FP cases, two had a very subtle form of AD (examples in supplementary Figure 11.3).

7.4.2.3 External Set

For the external validation dataset, a balanced accuracy score of 0.933 was obtained with sensitivity and specificity of 1.000 (100/100) and 0.865 (942/1089), respectively (Table 7.3, Figure 11.1 (c)). Furthermore, the AUC value is 0.993 (CI: 0.988 - 0.997) (Figure 7.5 (green)).

7.5 Discussion

This study demonstrated that rapid AI-based automated aortic dissection detection from CT images is feasible in an academic-level hospital in Germany. Overall, on all the datasets, an AUC of $> 88.7\%$ and sensitivities and specifities of $> 87.5\%$ and $> 68.8\%$ were consistently achieved. The strengths of this study are the comparably low training effort, and testing on a heterogeneous, real-world internal and external datasets with good overall processing times, indicating promising potential for clinical implementation as an alarming system for AD management.

AI-based approaches have been demonstrated to be of great potential for the detection of various pathologies in abdominal emergency imaging [139] and to support the management of chronic and acute vascular pathologies [157]. Structured text-based clinical data and imaging data have both been leveraged for AD detection, its rupture risk assessment, segmentation, therapy planning, and prediction of mortality [158–160]. AI-based regular chest radiography analysis has been shown to offer a precision of 90.2% in the detection of AD [141]. Early approaches of CT imaging-based AD characterization tools used rule-based technologies and small datasets of n<20 ADs [161]. Recently, CNN based algorithms were proposed as mentioned in Section 7.2. In contrast to most other algorithms, this study focussed on the abdominal region and comparably heterogeneous data with respect to CT scanners, contrast phase and morphological characteristics of AD, resulting in sensitivity from 87.5-100% and specificity from 68.8-86.5%. It is important to mention, that when used as a detection algorithm to improve prioritization of AD-positive cases, high sensitivity remains paramount and optimizing thresholds accordingly represents a very important design decision, but it also comes with the cost of a higher false-positive rate. An understanding of human performance is of high interest when interpreting AI performance. Nienaber et al. found the human sensitivity and specificity for the detection of acute AD in the thoracic region to be 93.8% and 87.1% [162], which is further underlined by the findings of Dreisbach et al., which describe substantial error rates of CT reading in acute AD under emergency conditions, depending on

reader experience [136]. An important goal for further research is therefore to create a detailed understanding of the performance of human performance alone, but also with AI support under realistic conditions in a prospective set-up.

Clinical implementation of AI lagged behind expectations in the past [163]. The reasons are limited performance, lack of trust in AI systems, and poor workflow integration, amongst others [164]. A practical way of implementing this algorithm would be to automatically prioritize acute AD cases by re-ordering radiology reading lists [160] which has been shown to reduce delays in pulmonary embolism management [165] or to alarm specialized vascular care teams. Importantly, physicians would realistically expect the algorithm to not miss any cases of AD and a failure in this area can be expected to negatively impact trust in a clinical setting, therefore optimization of parameters contributing to sensitivity, but also the quality of user training require high attention. Moreover, a broader AI solution also covering chronic AD, PAU, and IMH would increase clinical use and therefore should be added with priority. To overcome the black box problem, a graphical visualization of the dissection membrane within the aorta in each detected case could offer substantial merit.

The results of this study are limited by various factors. First, due to the retrospective nature, there was an inevitable selection bias. As AD remains a rare condition, the dataset inevitably contains a limited number of AD patient cases. This limitation is particularly evident in our internal cases and extends to the variety of CT scanners and morphological differences between cases. Even though thorough external validation on a newly created dataset was performed, generalizability required confirmation and prospective multicenter testing with more patient cases therefore represents an important next goal. Additionally, the generation of synthetic training data using latent diffusion models could contribute to overcome these limitations. Second, even though results are promising, balancing sensitivity and specificity remains a major issue. External validation demonstrated perfect sensitivity, but due to limited specificity, over half of the positive cases would yield false-positive results. On the internal training and validation datasets, on the other hand, six and two cases of AD were not detected. While some of these cases contained very subtle ADs and may have been missed for anatomical reasons, others presented with clear ADs where, for example, the isolation of the aortic vessel did not work correctly. Both the inclusion of larger training datasets as well as detailed refining and improvement of automated segmentation of the aortic lumen independent of external components like TotalSegmentator [156] or canny detector (for AD cases) might help improve performance. Third, due to the design-decision to not create a segmentation-based algorithm, possibilities of graphical visualization remain limited which could in the future possibly be added. Last, the focus only on abdominal imaging and only on AD limits clinical benefits to patients that present with abdominal AD, and therefore in the future, the inclusion of the full range of vascular pathologies and the thoracic region is needed.

7.6 Conclusion

In conclusion, the proposed algorithm yielded sensitivity >87.5% for the detection of AD within heterogeneous abdominal CT scans, which could be confirmed on

an internal as well as external validation dataset. It therefore seems promising as a detection tool for AD in the abdomen which offers the potential for earlier detection of AD, especially in patients with unclear symptoms, leading to improved management. Further in-hospital testing is encouraged and necessary.

Conflict of Interest Statement

SOS: The Department of Radiology and Nuclear Medicine has general research agreements with Siemens Healthineers. FGZ: The Department of Computer Assisted Clinical Medicine has general research agreements with Siemens Healthineers. Others: No conflicts of interest declared.

Funding

This research has been funded as part of the "DeepRAY" project under the funding code BW1_1523/02 by the Ministry of Economy, Employment and Tourism of the State of Baden-Württemberg, Germany.

8. Summary

Deep Learning (*DL*) has emerged as a transformative force in the realm of medical image analysis, offering unprecedented capabilities for automated, robust, and high-accuracy image interpretation. The advent of Convolutional Neural Network (*CNN*)s in recent years has proven vital for fast and automated image analysis in natural as well as medical domains [166]. Despite the proven efficacy of DL, its deployment in medical image analysis faces a notable challenge: the requirement for large, annotated datasets. Reasons for difficulty in obtaining large medical datasets include privacy concerns, the rarity of certain conditions, and the logistical difficulties of gathering large annotated datasets. Therefore, this work focuses on leveraging DL techniques for medical image analysis using small datasets.

In this work, various techniques have been employed to deal with small dataset problems. Firstly, attention mechanisms that help in guiding the CNNs to focus on Regions of Interest (*ROIs*) and discard noisy signals [58] were implemented so that the networks learn as much as possible from small data sizes. The cosine loss function has been shown to be more effective for image classification for small datasets [65]. This loss function was adapted and implemented for a segmentation task. Another technique, known as Sharpness Aware Minimization (*SAM*), has shown that smoother minima are better for generalizability than sharper ones [57]. Furthermore, to improve generalizability across datasets, image preprocessing techniques like resampling and histogram matching were implemented. These image processing techniques create similar image distribution among datasets that originate from different sources and hence help in avoiding/reducing data distributional shifts that usually hinders the ability of neural networks to adapt to data from multiple sources. The data distributional shift is more important for consideration in the medical domain due to the lack of availability of large datasets from multiple sources for creating robust DL models.

In Chapter 3 to Chapter 7 multiple scientific studies that employed the aforementioned techniques are exhibited and have been shown to produce effective results. These techniques have been applied across 3 sub-tasks: 1. Segmentation, 2. Classification, and 3. Regression. A comprehensive summary of each study is provided in the following paragraphs.

Deep Learning-Based Total Kidney Volume Segmentation in Autosomal Dominant Polycystic Kidney Disease Using Attention, Cosine Loss, and Sharpness Aware Minimization,
Diagnostics, doi: https://doi.org/10.3390/diagnostics12051159

Chapter 3 focuses on the development of a deep learning methodology for segmenting Total Kidney Volume (*TKV*) in patients with Autosomal Dominant Polycystic Kidney Disease (*ADPKD*) using Magnetic Resonance Imaging (*MRI*) data. The significance of accurate TKV estimation is underscored by its role as a biomarker for disease progression in ADPKD, a leading cause of end-stage renal disease. Traditional manual segmentation methods are time-consuming and subject to variability, highlighting the need for automated techniques.

The proposed solution integrates attention mechanisms into the U-Net architecture, enhancing its ability to focus on relevant features for better segmentation. The cosine loss function is employed to tackle the challenge posed by small datasets, a common limitation in medical imaging. Furthermore, SAM is incorporated to enhance model generalizability, a critical aspect for clinical applications. The methodology is validated on a dataset comprising 100 MRI scans, demonstrating significant improvements in segmentation accuracy over the reference U-Net model. Key findings include achieving a Dice Similarity Coefficient (*DSC*) of 0.918, a Mean Symmetric Surface Distance (*MSSD*) of 1.20 mm, and a mean TKV difference of -1.72%, with an R^2 of 0.96, using only a limited dataset for training and testing.

The study also explores the efficacy of ensemble models, revealing further improvements in segmentation accuracy, thereby underscoring the potential of combining multiple deep learning models for enhanced performance. This study presents a promising advancement in the field of medical imaging for ADPKD, offering a methodology that not only improves segmentation accuracy but also addresses the challenges associated with small datasets and model generalizability.

Generalizable Kidney Segmentation for Total Volume Estimation,
Proc. Bildverarbeitung für die Medizin 2024, doi: https://doi.org/10.1007/978-3-658-44037-4_75

This study presents a deep learning framework designed for the automated segmentation of kidneys from T1-weighted MRI scans in patients with ADPKD, aiming to streamline the TKV estimation process - a crucial marker for monitoring disease progression.

The proposed approach integrates Nyul normalization [80] , resampling, and attention mechanisms within a CNN framework to craft a generalizable model capable of accurate kidney delineation.

The model's validation was conducted using two distinct datasets (93 and 41 patients, respectively), showcasing its ability to generalize across different data sources. The results demonstrated a significant enhancement over the baseline model, with an average improvement of 9.45% in Dice scores across both datasets. Such improvements highlight the model's potential in reliably calculating TKV from MRI images in ADPKD patients, a task pivotal for disease monitoring and management.

The incorporation of Nyul normalization and resampling techniques plays a crucial role in harmonizing intensity distributions across datasets, thereby bolstering the model's generalizability - an essential trait for clinical application.

Automated Prognosis of Renal Function Decline in ADPKD Patients using Deep Learning,
Zeitschrift für Medizinische Physik, doi: https://doi.org/10.1016/j.zemedi.2023.08.001

An application of classification and regression is presented in Chapter 5. It presents a deep learning approach for predicting renal function decline in patients with ADPKD using MRI data (135 patients). Recognizing the limitations of current biomarkers:

Height-adjusted Total Kidney Volume (*HtTKV*), estimated Glomerular Filtration Rate (*eGFR*), and patient age, in accurately predicting disease progression, the research implements a two-fold deep learning approach to enhance prognostic accuracy.

Firstly, an automated kidney volume segmentation model is developed, utilizing a CNN equipped with attention mechanisms. This model accurately calculates HtTKV from T2-weighted MRI images, addressing the challenge of manual measurement's time consumption and observer variability. Secondly, leveraging the segmented kidney volumes, the study employs a Multi-Layer-Perceptron (*MLP*) alongside a CNN to predict the progression to Chronic Kidney Disease (*CKD*) stages >=3A, >=3B, and a 30% decline in eGFR after 8 years from baseline. This dual-model approach integrates automatically generated features from MRI images with conventional biomarkers, presenting a more comprehensive method for ADPKD prognosis.

The study's findings demonstrate high prognostic accuracy with Area Under the Curve (*AUC*) scores > 0.95 for predicting CKD stages >=3A and >=3B, and a 30% decline in eGFR. Furthermore, the research extends its analysis to predict distinct CKD stages after eight years, achieving an AUC of 0.97.

In addition to classification tasks, the study also explores regression analysis by predicting the percent change in eGFR after 8 years. The prognosis network attains a Pearson correlation coefficient of 0.81 between predicted and measured eGFR decline. This aspect of the study emphasizes the model's versatility not only in classification but also in providing quantitative predictions regarding eGFR changes, surpassing the performance of existing prognostic models.

This study presents a potential to improve patient management by predicting disease progression more accurately. The study's findings suggest that deep learning models can complement traditional clinical biomarkers, providing a more comprehensive and automated approach to ADPKD prognosis.

A Generalizable Deep Voxel-Guided Morphometry Algorithm for the Detection of Subtle Lesion Dynamics in Multiple Sclerosis,
Frontiers in Neuroscience, doi: https://doi.org/10.3389/fnins.2024.1326108

Chapter 6 introduces a deep learning algorithm designed to enhance the analysis of Multiple Sclerosis (*MS*) disease activity through Voxel-Guided Morphometry (*VGM*) maps generated from longitudinal MRI brain volumes. Emphasizing the need for accurate detection and monitoring of MS-related changes in brain structures, the study focuses on creating a generalizable model capable of effectively being applied to unseen datasets.

The study employed a 3D residual U-Net architecture integrated with attention mechanisms. This model facilitates spatial feature extraction from MRI volumes, with attention mechanisms highlighting salient regions critical for analysis. Image normalization techniques like histogram matching and resampling were implemented to bolster the model's generalizability across different MRI systems, ensuring robust performance regardless of the data source.

The algorithm's effectiveness was validated on a primary dataset (Dataset A; 71 patients) and further evaluated for generalizability on two additional unseen datasets (Datasets B and C; 116 patients), showcasing an average improvement of 4.2% in Mean Absolute Error (*MAE*) over the state-of-the-art (SOTA) method. This signifies its ability to adapt and perform consistently across diverse imaging conditions.

Quantitative results from the study reveal the proposed approach's capability to outperform existing methods, with the Convolution Block Attention Module (*CBAM*)-Attention U-Net achieving the highest scores in structural similarity and accuracy metrics across multiple datasets. Qualitative evaluations by expert neuroradiologists confirm the high similarity of Deep VGM maps to traditional VGM maps, further validated by the low lesion error rate and dramatically reduced computation times.

The proposed model's independence from specific MRI systems and its robust performance across varied datasets underscore its potential as a valuable tool for clinical practitioners, offering fast, accurate, and reliable insights into subtle MS disease activity.

Streamlining acute Abdominal Aortic Dissection management - an AI based CT imaging workflow,
Journal of Imaging Informatics in Medicine, doi: https://doi.org/10.1007/s10278-024-01164-0

In Chapter 7, a CNN-based framework was developed to automate the detection of acute Abdominal Aortic Dissection (*AD*) in Computed Tomography (*CT*) scans, aiming to streamline the diagnostic process for this critical condition. Utilizing a retrospective dataset comprising 195 cases, including 94 AD instances, the study sought to enhance intrahospital workflow efficiency through early and accurate AD identification.

The methodology centered on a CNN architecture designed for the robust, automated detection of AD within the abdominal region. The pipeline involved preprocessing steps such as abdominal region extraction, aorta ROI isolation, and edge extraction to highlight the AD membrane, culminating in the classification of the aortic ROI into dissected or healthy categories. The model underwent validation on the internal set (cross-validation set; n=163, with 78 AD cases), an internal validation set (n=32, with 16 AD cases), and an expansive external set (n=1189, with 100 AD cases), testing its capability to generalize across varying datasets.

Quantitative metrics from the internal set (cross-validation) demonstrated an AUC of 0.932, balanced accuracy of 0.860, and a sensitivity score of 0.885. While, for the internal validation set an AUC of 0.887, balanced accuracy of 0.781, and a sensitivity score of 0.875 was obtained. External validation further confirmed the model's effectiveness, with an AUC of 0.993, balanced accuracy of 0.933, and a perfect sensitivity score of 1.000.

This study offers a possible solution for improving the efficiency and effectiveness of AD management, paving the way for further research and development in emergency radiology Artificial Intelligence (*AI*)-assisted diagnosis.

9. Outlook

Developing algorithms for Deep Learning (*DL*) based medical image analysis faces two primary challenges: the scarcity of large, annotated datasets and the heterogeneity among data sources. This thesis has endeavored to address these challenges by integrating techniques related to image normalization, attention mechanisms, and adaptive loss functions. The work has shown that adopting techniques like attention mechanisms, cosine loss function, Sharpness Aware Minimization (*SAM*), and histogram matching (data harmonization) can enhance the performance and robustness of medical image analysis algorithms. These methods have proven to be generalizable and effective across various tasks, including segmentation, classification, and regression, particularly in scenarios involving small datasets.

Looking ahead, to further improve the robustness and accuracy of DL models, it is essential to focus on expanding and diversifying training datasets. This can be achieved through collaborative efforts across institutions to collect data, leveraging synthetic data generation, and exploring self-supervised learning methods to utilize large amounts of unannotated data. Additionally, while normalization techniques have improved model generalizability, incorporating advanced methods such as Generative Adversarial Networks (*GANs*) for data augmentation and domain adaptation can further reduce data distributional shifts. These techniques can help models better adapt to variations in data from different sources, enhancing their applicability in diverse clinical environments.

Furthermore, the integration of these algorithms into clinical settings necessitates further validation on multi-center data. Large-scale studies involving diverse clinical environments are essential to ensure the robustness and reliability of these models. Furthermore, developing user-friendly software tools and interfaces will facilitate the seamless integration of DL models into clinical workflows, enhancing their practical utility. This will not only improve the efficiency of medical image analysis but also support clinicians in making more accurate and timely diagnoses.

While important progress has been made in this work for conditions like Autosomal Dominant Polycystic Kidney Disease (*ADPKD*) and Multiple Sclerosis (*MS*), future research should extend these techniques to other medical imaging challenges, such as oncology, cardiology, and rare diseases. By focusing on developing models that can handle complex cases and anomalies, the scope and impact of DL in medical image analysis can be further broadened. This will pave the way for more comprehensive and versatile diagnostic tools, ultimately benefiting a wider range of patients.

Reflecting on the objectives set forth in this thesis, incremental progress has been made in enhancing the performance and generalizability of DL algorithms in medical image analysis. By addressing the challenges of small datasets and data heterogeneity, this work lays a foundation for future research to advance the field of automated medical image analysis. Ultimately, these advancements will contribute to improving patient care by offering more efficient, accurate, and scalable solutions for medical image analysis.

10. Bibliography

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS 2012* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), vol. 25, pp. 1097–1105, Curran Associates, Inc., 2012.

[2] M. A. Abdou, "Literature review: Efficient deep neural networks techniques for medical image analysis," *Neural Computing and Applications*, vol. 34, no. 8, pp. 5791–5812, 2022.

[3] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[4] G. Mohan and M. M. Subashini, "Mri based medical image analysis: Survey on brain tumor grade classification," *Biomedical Signal Processing and Control*, vol. 39, pp. 139–161, 2018.

[5] F. Altaf, S. M. Islam, N. Akhtar, and N. K. Janjua, "Going deep in medical image analysis: concepts, methods, challenges, and future directions," *IEEE Access*, vol. 7, pp. 99540–99572, 2019.

[6] H.-P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, "Deep learning in medical image analysis," *Deep learning in medical image analysis: challenges and applications*, pp. 3–21, 2020.

[7] E. Kondrateva, M. Pominova, E. Popova, M. Sharaev, A. Bernstein, and E. Burnaev, "Domain shift in computer vision models for mri data analysis: an overview," in *Thirteenth International Conference on Machine Vision*, vol. 11605, pp. 126–133, SPIE, 2021.

[8] Y. L. Thian, D. W. Ng, J. T. P. D. Hallinan, P. Jagmohan, S. Y. Sia, J. S. A. Mohamed, S. T. Quek, and M. Feng, "Effect of training data volume on performance of convolutional neural network pneumothorax classifiers," *Journal of Digital Imaging*, vol. 35, no. 4, pp. 881–892, 2022.

[9] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, pp. 843–852, 2017.

[10] S. W. Atlas, *Magnetic resonance imaging of the brain and spine*, vol. 1. Lippincott Williams & Wilkins, 2009.

[11] G. D. Rubin, C. J. Ryerson, L. B. Haramati, N. Sverzellati, J. P. Kanne, S. Raoof, N. W. Schluger, A. Volpi, J.-J. Yim, I. B. Martin, *et al.*, "The role of chest imaging in patient management during the covid-19 pandemic: a multinational consensus statement from the fleischner society," *Radiology*, vol. 296, no. 1, pp. 172–180, 2020.

[12] P. Lauterbur, "Image formation by induced local interactions," *Nature*, vol. 242, no. 5394, pp. 190–191, 1973.

[13] K. E. Keenan, J. R. Biller, J. G. Delfino, M. A. Boss, M. D. Does, J. L. Evelhoch, M. A. Griswold, J. L. Gunter, R. S. Hinks, S. W. Hoffman, *et al.*, "Recommendations towards standards for quantitative mri (qmri) and outstanding needs," *Journal of magnetic resonance imaging: JMRI*, vol. 49, no. 7, p. e26, 2019.

[14] G. N. Hounsfield, "Computerized transverse axial scanning (tomography): Part 1. description of system," *The British Journal of Radiology*, vol. 46, no. 552, pp. 1016–1022, 1973.

[15] T. Buzug, *Computed Tomography: From Photon Statistics to Modern Cone-Beam CT*. Springer Berlin Heidelberg, 2008.

[16] A. Katsevich, "An improved exact filtered backprojection algorithm for spiral computed tomography," *Advances in Applied Mathematics*, vol. 32, no. 4, pp. 681–697, 2004.

[17] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, 1943.

[18] E. Grossi and M. Buscema, "Introduction to artificial neural networks," *European journal of gastroenterology & hepatology*, vol. 19, no. 12, pp. 1046–1054, 2007.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[20] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, p. 106, 1962.

[21] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," *Advances in neural information processing systems*, vol. 2, 1989.

[22] C. C. Aggarwal, "Convolutional neural networks," in *Neural networks and deep learning: a textbook*, ch. 8, pp. 315–370, Springer, 2018.

[23] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833, Springer, 2014.

[24] L. Deng and D. Yu, "Deep learning: Methods and applications," Tech. Rep. MSR-TR-2014-21, Microsoft, May 2014.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE, 2016.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, IEEE, 2015.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Intl. Conf. Med. Image Comput. Comput-Assist. Intervent. (MICCAI)*, pp. 234–241, Springer, 2015.

[29] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 2016 4th Intl Conf. 3D Vis. (3DV)*, pp. 565–571, IEEE, 2016.

[30] Ö. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, vol. 9901 of *LNCS*, pp. 424–432, Springer, 2016.

[31] C. C. Aggarwal, "Neural networks and deep learning," *Cham: Springer International Publishing*, 2018.

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

[33] A. Géron, *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc.", 2017.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[35] BioRender.com, "Created with biorender.com," 2024. https://www.biorender.com.

[36] J. J. Grantham, "Autosomal dominant polycystic kidney disease," *New England Journal of Medicine*, vol. 359, no. 14, pp. 1477–1485, 2008.

[37] E. Cornec-Le Gall, A. Alam, and R. D. Perrone, "Autosomal dominant polycystic kidney disease," *The Lancet*, vol. 393, no. 10174, pp. 919–935, 2019.

[38] M. P. McGinley, C. H. Goldschmidt, and A. D. Rae-Grant, "Diagnosis and treatment of multiple sclerosis: a review," *Jama*, vol. 325, no. 8, pp. 765–779, 2021.

[39] C. A. Nienaber, R. E. Clough, N. Sakalihasan, T. Suzuki, R. Gibbs, F. Mussa, M. P. Jenkins, M. M. Thompson, A. Evangelista, J. S. Yeh, *et al.*, "Aortic dissection," *Nature reviews Disease primers*, vol. 2, no. 1, pp. 1–18, 2016.

[40] J. J. Grantham, "Polycystic kidney disease: from the bedside to the gene and back," *Curr. Opin. Nephrol. Hy.*, vol. 10, no. 4, pp. 533–542, 2001.

[41] A. B. Chapman *et al.*, "Renal structure in early autosomal-dominant polycystic kidney disease (adpkd): The consortium for radiologic imaging studies of polycystic kidney disease (crisp) cohort," *Kidney Intl.*, vol. 64, no. 3, pp. 1035–1045, 2003.

[42] O. Z. Dalgaard, "Bilateral polycystic disease of the kidneys : A follow-up of two hundred and eighty four paients and their families.," *Acta Med. Scand.*, vol. 328, pp. 1–251, 1957.

[43] M. V. Irazabal, L. J. Rangel, E. J. Bergstralh, S. L. Osborn, A. J. Harmon, J. L. Sundsbak, K. T. Bae, A. B. Chapman, J. J. Grantham, M. Mrug, *et al.*, "Imaging classification of autosomal dominant polycystic kidney disease: a simple model for selecting patients for clinical trials," *Journal of the American Society of Nephrology*, vol. 26, no. 1, pp. 160–172, 2015.

[44] J. J. Grantham, V. E. Torres, A. B. Chapman, L. M. Guay-Woodford, K. T. Bae, B. F. King Jr, L. H. Wetzel, D. A. Baumgarten, P. J. Kenney, P. C. Harris, *et al.*, "Volume Progression in Polycystic Kidney Disease," *N. Engl. J. Med.*, vol. 354, pp. 2122–2130, May 2006.

[45] Center for Drug Evaluation and Research, "Qualification of biomarker total kidney volume in studies for treatment of autosomal dominant polycystic kidney disease draft guidance for industry." https://www.fda.gov/regulatory-information/search-fda-guidance-documents/qualification-biomarker-total-kidney-volume-studies-treatment-autosomal-dominant-polycystic-kidney, 2016.

[46] F. G. Zöllner, M. Kociński, L. Hansen, A.-K. Golla, A. Š. Trbalić, A. Lundervold, A. Materka, and P. Rogelj, "Kidney segmentation in renal magnetic resonance imaging - current status and prospects," *IEEE Access*, vol. 9, pp. 71577–71605, 2021.

[47] F. G. Zöllner, E. Svarstad, A. Z. Munthe-Kaas, L. R. Schad, A. Lundervold, and J. Rørvik, "Assessment of kidney volumes from mri: acquisition and segmentation techniques," *Am. J. Roentgenol.*, vol. 199, no. 5, pp. 1060–1069, 2012.

[48] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Z. Med. Phys.*, vol. 29, pp. 102–127, May 2019.

[49] T. L. Kline, P. Korfiatis, M. E. Edwards, J. D. Blais, F. S. Czerwiec, P. C. Harris, B. F. King, V. E. Torres, and B. J. Erickson, "Performance of an artificial multi-observer deep neural network for fully automated segmentation of polycystic kidneys," *J. Digit. Imag.*, vol. 30, no. 4, pp. 442–448, 2017.

[50] M. D. van Gastel, M. E. Edwards, V. E. Torres, B. J. Erickson, R. T. Gansevoort, and T. L. Kline, "Automatic measurement of kidney and liver volumes

from mr images of patients affected by autosomal dominant polycystic kidney disease," *J. Am. Soc. Nephrol.*, vol. 30, no. 8, pp. 1514–1522, 2019.

[51] V. Bevilacqua, A. Brunetti, G. D. Cascarano, F. Palmieri, A. Guerriero, and M. Moschetta, "A deep learning approach for the automatic detection and segmentation in autosomal dominant polycystic kidney disease based on magnetic resonance images," in *Proc. Intl. Conf. Intel. Comput.*, pp. 643–649, Springer, 2018.

[52] G. Mu, Y. Ma, M. Han, Y. Zhan, X. Zhou, and Y. Gao, "Automatic mr kidney segmentation for autosomal dominant polycystic kidney disease," in *Proc. Med. Imag. 2019: Comput. Aided Diag.*, vol. 10950, p. 109500X, International Society for Optics and Photonics, 2019.

[53] A. J. Daniel, C. E. Buchanan, T. Allcock, D. Scerri, E. F. Cox, B. L. Prestwich, and S. T. Francis, "Automated renal segmentation in healthy and chronic kidney disease subjects using a convolutional neural network," *Magn. Reson. Med.*, vol. 86, no. 2, pp. 1125–1136, 2021.

[54] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[55] D. F. Bauer *et al.*, "Generation of annotated multimodal ground truth datasets for abdominal medical image registration," *Intl J. Comput. Assist. Rad. Surg.*, vol. 16, no. 8, pp. 1277–1285, 2021.

[56] T. Russ, S. Goerttler, A. Schnurr, D. Bauer, S. Hatamikia, L. R. Schad, F. G. Zöllner, and K. Chung, "Synthesis of ct images from digital body phantoms using cyclegan," *Int J CARS*, vol. 14, no. 10, pp. 1741–1750, 2019.

[57] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *Proc. Intl. Conf. Learn. Represent. (ICLR)*, 2021.

[58] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, *et al.*, "Attention u-net: Learning where to look for the pancreas," in *Medical Imaging with Deep Learning (MIDL)*, 2022.

[59] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[60] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Europ. Conf. Comput. Vis. (ECCV)*, pp. 3–19, 2018.

[61] T. Zhou, L. Li, X. Li, C.-M. Feng, J. Li, and L. Shao, "Group-wise learning for weakly supervised semantic segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 799–811, 2021.

[62] T. Zhou, J. Li, S. Wang, R. Tao, and J. Shen, "Matnet: Motion-attentive transition network for zero-shot video object segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 8326–8338, 2020.

[63] A.-K. Schnurr, C. Drees, L. R. Schad, and F. G. Zöllner, "Comparing sample mining schemes for cnn kidney segmentation in t1w mri," in *Proceedings of the International Conference on Functional Renal Imaging*, 2019.

[64] Z. Yaniv, B. C. Lowekamp, H. J. Johnson, and R. Beare, "Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research," *J. Digit. Imag.*, vol. 31, no. 3, pp. 290–303, 2018.

[65] B. Barz and J. Denzler, "Deep learning on small datasets without pre-training using cosine loss," in *Proc. IEEE/CVF Winter Conf. App. Comput. Vis.*, pp. 1371–1380, 2020.

[66] A.-K. Golla, D. F. Bauer, R. Schmidt, T. Russ, D. Nörenberg, K. Chung, C. Tönnes, L. R. Schad, and F. G. Zöllner, "Convolutional neural network ensemble segmentation with ratio-based sampling for the arteries and veins in abdominal ct scans," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 5, pp. 1518–1526, 2021.

[67] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[68] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 903–921, 2004.

[69] C. Payer, D. Štern, T. Neff, H. Bischof, and M. Urschler, "Instance segmentation and tracking with cosine embeddings and recurrent hourglass networks," in *Proc. Intl. Conf. Med. Image Comput. Comput-Assist. Intervent. (MICCAI)*, pp. 3–11, Springer, 2018.

[70] F. G. Zöllner, R. Sance, P. Rogelj, M. J. Ledesma-Carbayo, J. Rørvik, A. Santos, and A. Lundervold, "Assessment of 3D DCE-MRI of the kidneys using non-rigid image registration and segmentation of voxel time courses.," *Comp. Med. Imag. Graph.*, vol. 33, pp. 171–81, Apr. 2009.

[71] T. Zhou, W. Wang, E. Konukoglu, and L. Van Gool, "Rethinking semantic segmentation: A prototype view," *arXiv preprint arXiv:2203.15102*, 2022.

[72] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 633–641, 2017.

[73] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 3213–3223, 2016.

[74] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1209–1218, 2018.

[75] A. E. Kavur, L. I. Kuncheva, and M. A. Selver, "Basic ensembles of vanilla-style deep learning models improve liver segmentation from ct images," *arXiv preprint arXiv:2001.09647*, 2020.

[76] C. Zhang and Y. Ma, *Ensemble machine learning: methods and applications*, pp. 1–2. Springer, 2012.

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

[78] J. J. Grantham, V. E. Torres, A. B. Chapman, L. M. Guay-Woodford, K. T. Bae, B. F. King Jr, L. H. Wetzel, D. A. Baumgarten, P. J. Kenney, P. C. Harris, *et al.*, "Volume progression in polycystic kidney disease," *N Engl J Med.*, vol. 354, no. 20, pp. 2122–2130, 2006.

[79] A. Raj, F. Tollens, A. Caroli, D. Nörenberg, and F. G. Zöllner, "Automated prognosis of renal function decline in adpkd patients using deep learning," *Z Med Phy*, 2023.

[80] L. G. Nyúl and J. K. Udupa, "On standardizing the mr image intensity scale," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 42, no. 6, pp. 1072–1081, 1999.

[81] A. Raj, F. Tollens, L. Hansen, A.-K. Golla, L. R. Schad, D. Nörenberg, and F. G. Zöllner, "Deep learning-based total kidney volume segmentation in autosomal dominant polycystic kidney disease using attention, cosine loss, and sharpness aware minimization," *Diagnostics*, vol. 12, no. 5, p. 1159, 2022.

[82] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2020.

[83] X. He, Z. Hu, H. Dev, D. J. Romano, A. Sharbatdaran, S. I. Raza, S. J. Wang, K. Teichman, G. Shih, J. M. Chevalier, *et al.*, "Test retest reproducibility of organ volume measurements in adpkd using 3d multimodality deep learning," *Acad Radiol*, 2023.

[84] R. Perrone, Y. Pirson, R. Schrier, R. Torra, V. Torres, T. Watnick, D. Wheeler, and on behalf of the Conference Participants, "Autosomal-dominant polycystic kidney disease (adpkd): executive summary from a kidney disease: Improving global outcomes (kdigo) controversies conference," *Kidney Int*, vol. 88, no. 1, pp. 17–27, 2015.

[85] F. T. Chebib and V. E. Torres, "Assessing risk of rapid progression in autosomal dominant polycystic kidney disease and special considerations for disease-modifying therapy," *American Journal of Kidney Diseases*, vol. 78, no. 2, pp. 282–292, 2021.

[86] E. Cornec-Le Gall, M.-P. Audrézet, A. Rousseau, M. Hourmant, E. Renaudineau, C. Charasse, M.-P. Morin, M.-C. Moal, J. Dantal, B. Wehbe, R. Perrichot, T. Frouget, C. Vigneau, J. Potier, P. Jousset, M.-P. Guillodo, P. Siohan, N. Terki, T. Sawadogo, D. Legrand, V. Menoyo-Calonge, S. Benarbia, D. Besnier, H. Longuet, C. Férec, and Y. Le Meur, "The propkd score: A new algorithm to predict renal survival in autosomal dominant polycystic kidney disease," *J Am Soc Neph*, vol. 27, no. 3, pp. 942–951, 2016.

[87] E. Cornec-Le Gall, J. D. Blais, M. V. Irazabal, O. Devuyst, R. T. Gansevoort, R. D. Perrone, A. B. Chapman, F. S. Czerwiec, J. Ouyang, C. M. Heyer, S. R. Senum, Y. Le Meur, V. E. Torres, and P. C. Harris, "Can we further enrich autosomal dominant polycystic kidney disease clinical trials for rapidly progressive patients? Application of the PROPKD score in the TEMPO trial," *Nephrol Dial Transplant*, vol. 33, pp. 645–652, 07 2017.

[88] S. Riyahi, H. Dev, J. D. Blumenfeld, H. Rennert, X. Yin, H. Attari, I. Barash, I. Chicos, W. Bobb, S. Donahue, *et al.*, "Hemorrhagic cysts and other mr biomarkers for predicting renal dysfunction progression in autosomal dominant polycystic kidney disease," *Journal of Magnetic Resonance Imaging*, vol. 53, no. 2, pp. 564–576, 2021.

[89] T. L. Kline, P. Korfiatis, M. E. Edwards, K. T. Bae, A. Yu, A. B. Chapman, M. Mrug, J. J. Grantham, D. Landsittel, W. M. Bennett, *et al.*, "Image texture features predict renal function decline in patients with autosomal dominant polycystic kidney disease," *Kidney international*, vol. 92, no. 5, pp. 1206–1216, 2017.

[90] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[91] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, "The elements of statistical learning: data mining, inference, and prediction," vol. 2, ch. 10, Springer, 2009.

[92] C. C. Aggarwal, *Recommender Systems*. Springer International Publishing, 2016.

[93] "Imaging classification of adpkd: A simple model for selecting patients for clinical trials," Accessed 18 October 2022.

[94] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[95] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[96] X. Cao, Y. Lin, B. Yang, Y. Li, and J. Zhou, "Comparison between statistical model and machine learning methods for predicting the risk of renal function decline using routine clinical data in health screening," *Risk Management and Healthcare Policy*, vol. 15, p. 817, 2022.

[97] J. Zhao, Y. Zhang, J. Qiu, X. Zhang, F. Wei, J. Feng, C. Chen, K. Zhang, S. Feng, and W.-D. Li, "An early prediction model for chronic kidney disease," *Scientific reports*, vol. 12, no. 1, pp. 1–9, 2022.

[98] J. Norouzi, A. Yadollahpour, S. Mirbagheri, M. Mazdeh, and S. Hosseini, "Predicting renal failure progression in chronic kidney disease using integrated intelligent fuzzy expert system," *Comput Math Methods Med*, p. 6080814, 2016.

[99] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," *Journal of Big Data*, vol. 9, no. 1, pp. 1–19, 2022.

[100] H. Ilyas, S. Ali, M. Ponum, O. Hasan, M. T. Mahmood, M. Iftikhar, and M. H. Malik, "Chronic kidney disease diagnosis using decision tree algorithms," *BMC nephrology*, vol. 22, no. 1, pp. 1–11, 2021.

[101] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked*, vol. 15, p. 100178, 2019.

[102] A. Dal-Bianco, G. Grabner, C. Kronnerwetter, M. Weber, R. Höftberger, T. Berger, E. Auff, F. Leutmezer, S. Trattnig, H. Lassmann, *et al.*, "Slow expansion of multiple sclerosis iron rim lesions: pathology and 7 t magnetic resonance imaging," *Acta neuropathologica*, vol. 133, no. 1, pp. 25–42, 2017.

[103] U. W. Kaunzner and S. A. Gauthier, "Mri in the assessment and monitoring of multiple sclerosis: an update on best practice," *Therapeutic advances in neurological disorders*, vol. 10, no. 6, pp. 247–261, 2017.

[104] M. Filippi, M. A. Rocca, O. Ciccarelli, N. De Stefano, N. Evangelou, L. Kappos, A. Rovira, J. Sastre-Garriga, M. Tintorè, J. L. Frederiksen, *et al.*, "Mri criteria for the diagnosis of multiple sclerosis: Magnims consensus guidelines," *The Lancet Neurology*, vol. 15, no. 3, pp. 292–303, 2016.

[105] J. M. Frischer, S. D. Weigand, Y. Guo, N. Kale, J. E. Parisi, I. Pirko, J. Mandrekar, S. Bramow, I. Metz, W. Brück, *et al.*, "Clinical and pathological insights into the dynamic nature of the white matter multiple sclerosis plaque," *Annals of neurology*, vol. 78, no. 5, pp. 710–721, 2015.

[106] E. B. Lewis and N. C. Fox, "Correction of differential intensity inhomogeneity in longitudinal mr images," *Neuroimage*, vol. 23, no. 1, pp. 75–83, 2004.

[107] T. Schormann and M. Kraemer, "Voxel-guided morphometry ("vgm") and application to stroke," *IEEE Transactions on Medical Imaging*, vol. 22, no. 1, pp. 62–74, 2003.

[108] A.-K. Schnurr, P. Eisele, C. Rossmanith, S. Hoffmann, J. Gregori, A. Dabringhaus, M. Kraemer, R. Kern, A. Gass, and F. G. Zöllner, "Deep voxel-guided morphometry (vgm): learning regional brain changes in serial mri," in *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology: Third International Workshop, MLCN 2020, and Second International Workshop, RNO-AI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3*, pp. 159–168, Springer, 2020.

[109] C. Polman, S. Reingold, B. Banwell, M. Clanet, J. Cohen, M. Filippi, K. Fujihara, E. Havrdova, M. Hutchinson, L. Kappos, F. Lublin, X. Montalban, P. O'Connor, M. Sandberg-Wollheim, A. Thompson, E. Waubant, B. Weinshenker, and J. Wolinsky, "Diagnostic criteria for multiple sclerosis: 2010 revisions to the mcdonald criteria," *Annals of Neurology*, vol. 69, pp. 292–302, Feb 2011.

[110] A. Carass, S. Roy, A. Jog, J. L. Cuzzocreo, E. Magrath, A. Gherman, J. Button, J. Nguyen, F. Prados, C. H. Sudre, *et al.*, "Longitudinal multiple sclerosis lesion segmentation: resource and challenge," *NeuroImage*, vol. 148, pp. 77–102, 2017.

[111] F. Segonne, A. Dale, E. Busa, M. Glessner, D. Salat, H. Hahn, and B. Fischl, "A hybrid approach to the skull stripping problem in mri," *Neuroimage*, vol. 22, pp. 1060–1075, 2004.

[112] M. Kraemer, T. Schormann, G. Hagemann, B. Qi, O. W. Witte, and R. J. Seitz, "Delayed shrinkage of the brain after ischemic stroke: preliminary observations with voxel-guided morphometry," *Journal of Neuroimaging*, vol. 14, no. 3, pp. 265–272, 2004.

[113] J. Fox, M. Kraemer, T. Schormann, A. Dabringhaus, J. Hirsch, P. Eisele, K. Szabo, C. Weiss, M. Amann, K. Weier, *et al.*, "Individual assessment of brain tissue changes in ms and the effect of focal lesions on short-term focal atrophy development in ms: a voxel-guided morphometry study," *Int. Journal of Molecular Sciences*, vol. 17, no. 4, p. 489, 2016.

[114] M. Kraemer, T. Schormann, A. Dabringhaus, J. Hirsch, K. Stephan, V. Hömberg, L. Kappos, and A. Gass, "Individual assessment of chronic brain tissue changes in mri–the role of focal lesions for brain atrophy development. a voxel-guided morphometry study," *Klinische Neurophysiologie*, vol. 39, no. 01, p. A178, 2008.

[115] C. E. Weber, M. Wittayer, M. Kraemer, A. Dabringhaus, M. Platten, A. Gass, and P. Eisele, "Quantitative mri texture analysis in chronic active multiple sclerosis lesions," *Magnetic Resonance Imaging*, vol. 79, pp. 97–102, 2021.

[116] C. E. Weber, M. Wittayer, M. Kraemer, A. Dabringhaus, K. Bail, M. Platten, L. Schirmer, A. Gass, and P. Eisele, "Long-term dynamics of multiple sclerosis iron rim lesions," *Multiple Sclerosis and Related Disorders*, vol. 57, p. 103340, 2022.

[117] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[118] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[119] B. Sarica, D. Z. Seker, and B. Bayram, "A dense residual u-net for multiple sclerosis lesions segmentation from multi-sequence 3d mr images," *International Journal of Medical Informatics*, vol. 170, p. 104965, 2023.

[120] B. Sarica and D. Z. Seker, "New ms lesion segmentation with deep residual attention gate u-net utilizing 2d slices of 3d mr images," *Frontiers in Neuroscience*, vol. 16, p. 912000, 2022.

[121] O. Commowick, B. Combès, F. Cervenansky, and M. Dojat, "Automatic methods for multiple sclerosis new lesions detection and segmentation," *Frontiers in Neuroscience*, vol. 17, p. 1176625, 2023.

[122] S. Hitziger, W. X. Ling, T. Fritz, T. D'Albis, A. Lemke, and J. Grilo, "Triplanar u-net with lesion-wise voting for the segmentation of new lesions on longitudinal mri studies," *Frontiers in Neuroscience*, vol. 16, p. 964250, 2022.

[123] J. Andresen, H. Uzunova, J. Ehrhardt, T. Kepp, and H. Handels, "Image registration and appearance adaptation in non-correspondent image regions for new ms lesions detection," *Frontiers in Neuroscience*, vol. 16, p. 981523, 2022.

[124] R. A. Kamraoui, B. Mansencal, J. V. Manjon, and P. Coupé, "Longitudinal detection of new ms lesions using deep learning," *Frontiers in Neuroimaging*, vol. 1, p. 948235, 2022.

[125] P. Ashtari, B. Barile, S. Van Huffel, and D. Sappey-Marinier, "New multiple sclerosis lesion segmentation and detection using pre-activation u-net," *Frontiers in Neuroscience*, vol. 16, p. 975862, 2022.

[126] B. D. Basaran, P. M. Matthews, and W. Bai, "New lesion segmentation for multiple sclerosis brain images with imaging and lesion-aware augmentation," *Frontiers in Neuroscience*, vol. 16, p. 1007453, 2022.

[127] M. Schmidt-Mengin, T. Soulier, M. Hamzaoui, A. Yazdan-Panah, B. Bodini, N. Ayache, B. Stankoff, and O. Colliot, "Online hard example mining vs. fixed oversampling strategy for segmentation of new multiple sclerosis lesions from longitudinal flair mri," *Frontiers in Neuroscience*, vol. 16, p. 1004050, 2022.

[128] O. Commowick, F. Cervenansky, F. Cotton, and M. Dojat, "Msseg-2 challenge proceedings: Multiple sclerosis new lesions segmentation challenge using a data management and processing infrastructure," in *MICCAI 2021-24th International Conference on Medical Image Computing and Computer Assisted Intervention*, p. 126, 2021.

[129] E. Dufresne, D. Fortun, S. Kremer, and V. Noblet, "A unified framework for focal intensity change detection and deformable image registration. application to the monitoring of multiple sclerosis lesions in longitudinal 3d brain mri," *Frontiers in Neuroimaging*, vol. 1, p. 1008128, 2022.

[130] M. Cheng, A. Galimzianova, Ž. Lesjak, Ž. Špiclin, C. B. Lock, and D. L. Rubin, "A multi-scale multiple sclerosis lesion change detection in a multi-sequence mri," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pp. 353–360, Springer, 2018.

[131] I. Vilacosta, J. A. San Román, R. di Bartolomeo, K. Eagle, A. L. Estrera, C. Ferrera, S. Kaji, C. A. Nienaber, V. Riambau, H.-J. Schäfers, *et al.*, "Acute aortic syndrome revisited: Jacc state-of-the-art review," *Journal of the American College of Cardiology*, vol. 78, no. 21, pp. 2106–2125, 2021.

[132] E. Bossone and K. A. Eagle, "Epidemiology and management of aortic disease: aortic aneurysms and acute aortic syndromes," *Nature Reviews Cardiology*, vol. 18, no. 5, pp. 331–348, 2021.

[133] M. Wundram, V. Falk, J.-J. Eulert-Grehn, H. Herbst, J. Thurau, B. A. Leidel, E. Göncz, W. Bauer, H. Habazettl, and S. D. Kurz, "Incidence of acute type a aortic dissection in emergency departments," *Scientific Reports*, vol. 10, no. 1, p. 7434, 2020.

[134] A. R. Benkert and J. G. Gaca, "Initial medical management of acute aortic syndromes," *Aortic Dissection and Acute Aortic Syndromes*, pp. 119–129, 2021.

[135] N. A. Orabi, L. E. Quint, K. Watcharotone, B. Nan, D. M. Williams, and K. M. Kim, "Distinguishing acute from chronic aortic dissections using ct imaging features," *The International Journal of Cardiovascular Imaging*, vol. 34, pp. 1831–1840, 2018.

[136] J. G. Dreisbach, J. C. Rodrigues, and G. Roditi, "Emergency ct misdiagnosis in acute aortic syndrome," *The British Journal of Radiology*, vol. 94, no. 1126, p. 20201294, 2021.

[137] C. A. Nienaber and R. E. Clough, "Management of acute aortic dissection," *The Lancet*, vol. 385, no. 9970, pp. 800–811, 2015.

[138] K. M. Harris, C. E. Strauss, K. A. Eagle, A. T. Hirsch, E. M. Isselbacher, T. T. Tsai, H. Shiran, R. Fattori, A. Evangelista, J. V. Cooper, *et al.*, "Correlates of delayed recognition and treatment of acute type a aortic dissection: the international registry of acute aortic dissection (irad)," *Circulation*, vol. 124, no. 18, pp. 1911–1918, 2011.

[139] J. Liu, B. Varghese, F. Taravat, L. S. Eibschutz, and A. Gholamrezanezhad, "An extra set of intelligent eyes: application of artificial intelligence in imaging of abdominopelvic pathologies in emergency radiology," *Diagnostics*, vol. 12, no. 6, p. 1351, 2022.

[140] Y. Yi, L. Mao, C. Wang, Y. Guo, X. Luo, D. Jia, Y. Lei, J. Pan, J. Li, S. Li, *et al.*, "Advanced warning of aortic dissection on non-contrast ct: the combination of deep learning and morphological characteristics," *Frontiers in Cardiovascular Medicine*, vol. 8, p. 762958, 2022.

[141] D. K. Lee, J. H. Kim, J. Oh, T. H. Kim, M. S. Yoon, D. J. Im, J. H. Chung, and H. Byun, "Detection of acute thoracic aortic dissection based on plain chest radiography and a residual neural network (resnet)," *Scientific Reports*, vol. 12, no. 1, p. 21884, 2022.

[142] S. Chen, K. Ma, and Y. Zheng, "Med3d: Transfer learning for 3d medical image analysis," *arXiv preprint arXiv:1904.00625*, 2019.

[143] A. Hata, M. Yanagawa, K. Yamagata, Y. Suzuki, S. Kido, A. Kawata, S. Doi, Y. Yoshida, T. Miyata, M. Tsubamoto, *et al.*, "Deep learning algorithm for detection of aortic dissection on non-contrast-enhanced ct," *European radiology*, vol. 31, pp. 1151–1159, 2021.

[144] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[145] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[146] L.-T. Huang, Y.-S. Tsai, C.-F. Liou, T.-H. Lee, P.-T. P. Kuo, H.-S. Huang, and C.-K. Wang, "Automated stanford classification of aortic dissection using a 2-step hierarchical neural network at computed tomography angiography," *European Radiology*, pp. 1–9, 2022.

[147] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

[148] R. J. Harris, S. Kim, J. Lohr, S. Towey, Z. Velichkovich, T. Kabachenko, I. Driscoll, and B. Baker, "Classification of aortic dissection and rupture on post-contrast ct images using a convolutional neural network," *Journal of Digital Imaging*, vol. 32, no. 6, pp. 939–946, 2019.

[149] J. Cheng, S. Tian, L. Yu, X. Ma, and Y. Xing, "A deep learning algorithm using contrast-enhanced computed tomography (ct) images for segmentation and rapid automatic detection of aortic dissection," *Biomedical Signal Processing and Control*, vol. 62, p. 102145, 2020.

[150] M. S. Yellapragada, Y. Xie, B. Graf, D. Richmond, A. Krishnan, and A. Sitek, "Deep learning based detection of acute aortic syndrome in contrast ct images," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1474–1477, IEEE, 2020.

[151] Y. Guo, X. Chen, X. Lin, L. Chen, J. Shu, P. Pang, J. Cheng, M. Xu, and Z. Sun, "Non-contrast ct-based radiomic signature for screening thoracic aortic dissections: a multicenter study," *European Radiology*, vol. 31, pp. 7067–7076, 2021.

[152] A.-K. Golla, C. Tönnes, T. Russ, D. F. Bauer, M. F. Froelich, S. J. Diehl, S. O. Schoenberg, M. Keese, L. R. Schad, F. G. Zöllner, *et al.*, "Automated screening for abdominal aortic aneurysm in ct scans under clinical conditions using deep learning," *Diagnostics*, vol. 11, no. 11, p. 2131, 2021.

[153] Z. Yao, W. Xie, J. Zhang, Y. Dong, H. Qiu, H. Yuan, Q. Jia, T. Wang, Y. Shi, J. Zhuang, *et al.*, "Imagetbad: A 3d computed tomography angiography image dataset for automatic segmentation of type-b aortic dissection," *Frontiers in Physiology*, vol. 12, p. 732711, 2021.

[154] L. Radl, Y. Jin, A. Pepe, J. Li, C. Gsaxner, F.-h. Zhao, and J. Egger, "Avt: Multicenter aortic vessel tree cta dataset collection with ground truth segmentation masks," *Data in brief*, vol. 40, p. 107801, 2022.

[155] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu, *et al.*, "Abdomenct-1k: Is abdominal organ segmentation a solved problem?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6695–6714, 2021.

[156] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, *et al.*, "Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, 2023.

[157] A. P. Javidan, A. Li, M. H. Lee, T. L. Forbes, and F. Naji, "A systematic review and bibliometric analysis of applications of artificial intelligence and machine learning in vascular surgery," *Annals of Vascular Surgery*, vol. 85, pp. 395–405, 2022.

[158] L. D. Hahn, G. Mistelbauer, K. Higashigaito, M. Koci, M. J. Willemink, A. M. Sailer, M. Fischbein, and D. Fleischmann, "Ct-based true-and false-lumen segmentation in type b aortic dissection using machine learning," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 3, p. e190179, 2020.

[159] B. Li, T. Feridooni, C. Cuen-Ojeda, T. Kishibe, C. de Mestral, M. Mamdani, and M. Al-Omran, "Machine learning in vascular surgery: a systematic review and critical appraisal," *NPJ Digital Medicine*, vol. 5, no. 1, p. 7, 2022.

[160] D. Mastrodicasa, M. Codari, K. Bäumler, V. Sandfort, J. Shen, G. Mistelbauer, L. D. Hahn, V. L. Turner, B. Desjardins, M. J. Willemink, *et al.*,

"Artificial intelligence applications in aortic dissection imaging," in *Seminars in roentgenology*, vol. 57, pp. 357–363, Elsevier, 2022.

[161] D. Chen, X. Zhang, Y. Mei, F. Liao, H. Xu, Z. Li, Q. Xiao, W. Guo, H. Zhang, T. Yan, *et al.*, "Multi-stage learning for segmentation of aortic dissections using a prior aortic anatomy simplification," *Medical image analysis*, vol. 69, p. 101931, 2021.

[162] C. A. Nienaber, Y. von Kodolitsch, V. Nicolas, V. Siglow, A. Piepho, C. Brockhoff, D. H. Koschyk, and R. P. Spielmann, "The diagnosis of thoracic aortic dissection by noninvasive imaging procedures," *New England Journal of Medicine*, vol. 328, no. 1, pp. 1–9, 1993.

[163] R. Fujimori, K. Liu, S. Soeno, H. Naraba, K. Ogura, K. Hara, T. Sonoo, T. Ogura, K. Nakamura, T. Goto, *et al.*, "Acceptance, barriers, and facilitators to implementing artificial intelligence–based decision support systems in emergency departments: quantitative and qualitative evaluation," *JMIR formative research*, vol. 6, no. 6, p. e36501, 2022.

[164] F. A. Eltawil, M. Atalla, E. Boulos, A. Amirabadi, and P. N. Tyrrell, "Analyzing barriers and enablers for the acceptance of artificial intelligence innovations into radiology practice: a scoping review," *Tomography*, vol. 9, no. 4, pp. 1443–1455, 2023.

[165] L. Topff, E. R. Ranschaert, A. Bartels-Rutten, A. Negoita, R. Menezes, R. G. Beets-Tan, and J. J. Visser, "Artificial intelligence tool for detection and worklist prioritization reduces time to diagnosis of incidental pulmonary embolism at ct," *Radiology: Cardiothoracic Imaging*, vol. 5, no. 2, p. e220163, 2023.

[166] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical image analysis using convolutional neural networks: a review," *Journal of medical systems*, vol. 42, pp. 1–13, 2018.

[167] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[168] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, *et al.*, "Monai: An open-source framework for deep learning in healthcare," *arXiv preprint arXiv:2211.02701*, 2022.

[169] B. Zhao, L. H. Schwartz, and M. G. Kris, "Data from rider_lung ct," *The Cancer Imaging Archive*, 2015.

[170] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaissis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand, *et al.*, "The liver tumor segmentation benchmark (lits)," *Medical Image Analysis*, vol. 84, p. 102680, 2023.

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

[172] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, *et al.*, "A large annotated medical image dataset for the development and evaluation of segmentation algorithms," *arXiv preprint arXiv:1902.09063*, 2019.

[173] H. R. Roth, A. Farag, E. Turkbey, L. Lu, J. Liu, and R. M. Summers, "Data from pancreas-ct. the cancer imaging archive," *IEEE Transactions on Image Processing*, 2016.

[174] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*, pp. 556–564, Springer, 2015.

[175] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, pp. 1045–1057, 2013.

11. Appendix

11.1 Supplementary Material Chapter 3

Architecture	Loss	DSC (Left) ↑ 96 × 96	DSC (Right) ↑ 96 × 96	MSSD (mm) (Left) ↓ 96 × 96	MSSD (mm) (Right) ↓ 96 × 96	DSC (Left) ↑ 128 × 128	DSC (Right) ↑ 128 × 128	MSSD (mm) (Left) ↓ 128 × 128	MSSD (mm) (Right) ↓ 128 × 128
Baseline U-Net	$\mathcal{L}_{CE+DSC}$	0.886 ± 0.061	0.877 ± 0.099	1.992 ± 1.828	2.578 ± 4.388	0.901 ± 0.058	0.907 ± 0.083	1.381 ± 1.121	1.447 ± 3.727
SE U-net	$\mathcal{L}_{CE+DSC}$	0.899 ± 0.064	0.894 ± 0.076	1.594 ± 2.330	1.614 ± 2.123	0.900 ± 0.060	0.897 ± 0.071	1.630 ± 2.286	1.494 ± 2.232
	$\mathcal{L}_{COS}$	0.887 ± 0.111	0.882 ± 0.111	2.140 ± 4.858	1.694 ± 2.248	0.899 ± 0.059	0.901 ± 0.068	1.592 ± 2.116	1.377 ± 2.000
	$\mathcal{L}_{CE+DSC} + SAM$	0.888 ± 0.080	0.900 ± 0.087	1.512 ± 1.416	1.559 ± 3.308	0.912 ± 0.044	0.893 ± 0.095	1.129 ± 0.847	1.344 ± 1.941
	$\mathcal{L}_{COS} + SAM$	0.912 ± 0.045	0.891 ± 0.099	1.222 ± 0.957	1.547 ± 2.447	0.895 ± 0.112	0.905 ± 0.077	1.813 ± 4.790	1.367 ± 2.210
CBAM U-Net	$\mathcal{L}_{CE+DSC}$	0.894 ± 0.070	0.890 ± 0.098	1.682 ± 2.310	1.845 ± 3.532	0.905 ± 0.064	0.907 ± 0.064	1.285 ± 1.164	1.363 ± 2.223
	$\mathcal{L}_{COS}$	0.905 ± 0.053	0.900 ± 0.082	1.316 ± 0.984	1.293 ± 1.678	0.905 ± 0.061	0.904 ± 0.077	1.404 ± 1.727	1.429 ± 2.537
	$\mathcal{L}_{CE+DSC} + SAM$	0.891 ± 0.068	0.896 ± 0.080	1.893 ± 2.575	1.709 ± 2.596	0.917 ± 0.046	0.904 ± 0.081	1.072 ± 0.889	1.346 ± 2.486
	$\mathcal{L}_{COS} + SAM$	0.897 ± 0.067	0.899 ± 0.085	1.832 ± 2.626	1.701 ± 3.604	0.910 ± 0.054	0.908 ± 0.072	1.265 ± 1.406	1.270 ± 2.429
Attn. U-Net	$\mathcal{L}_{CE+DSC}$	0.894 ± 0.074	0.890 ± 0.093	1.662 ± 1.637	1.771 ± 2.655	0.914 ± 0.043	0.907 ± 0.061	1.224 ± 0.914	1.403 ± 2.080
	$\mathcal{L}_{COS}$	0.901 ± 0.057	0.907 ± 0.072	1.741 ± 2.455	1.366 ± 2.212	0.910 ± 0.049	0.909 ± 0.073	1.286 ± 1.573	1.3042.253
	$\mathcal{L}_{CE+DSC} + SAM$	0.894 ± 0.058	0.897 ± 0.093	1.910 ± 2.745	1.776 ± 3.212	0.913 ± 0.051	0.909 ± 0.069	1.353 ± 1.697	1.399 ± 2.374
	$\mathcal{L}_{COS} + SAM$	0.905 ± 0.059	0.907 ± 0.075	1.442 ± 1.940	1.321 ± 2.018	0.914 ± 0.048	0.915 ± 0.072	1.184 ± 1.190	1.257 ± 2.536
U-Net	$\mathcal{L}_{COS}$	0.898 ± 0.074	0.897 ± 0.093	1.725 ± 2.479	1.794 ± 3.727	0.908 ± 0.059	0.909 ± 0.076	1.208 ± 1.104	1.342 ± 2.909
	$\mathcal{L}_{COS} + SAM$	0.909 ± 0.051	0.910 ± 0.075	1.454 ± 2.109	1.313 ± 2.326	0.921 ± 0.043	0.914 ± 0.062	1.009 ± 0.843	1.274 ± 2.331
Ensemble-4-STAPLE	$\mathcal{L}_{COS} + SAM$	0.911 ± 0.054	0.909 ± 0.077	1.556 ± 2.498	1.351 ± 2.413	0.920 ± 0.046	0.919 ± 0.068	1.112 ± 1.144	1.158 ± 2.240
Ensemble-7-STAPLE	$\mathcal{L}_{CE+DSC} + SAM + \mathcal{L}_{COS} + SAM$	0.911 ± 0.053	0.909 ± 0.077	1.484 ± 2.145	1.399 ± 2.343	0.923 ± 0.045	0.919 ± 0.069	1.079 ± 1.162	1.134 ± 2.224
Ensemble-4-VOTING	$\mathcal{L}_{COS} + SAM$	0.918 ± 0.048	0.910 ± 0.074	1.086 ± 1.037	1.219 ± 2.029	0.919 ± 0.050	0.916 ± 0.066	1.035 ± 0.979	1.102 ± 1.989
Ensemble-7-VOTING	$\mathcal{L}_{CE+DSC} + SAM + \mathcal{L}_{COS} + SAM$	0.916 ± 0.049	0.914 ± 0.073	1.238 ± 1.779	1.213 ± 2.035	0.925 ± 0.044	0.919 ± 0.067	0.993 ± 0.998	1.085 ± 2.072

Table 11.1: The Dice Similarity Coefficient (*DSC*) and Mean Symmetric Surface Distance (*MSSD*) (in mm) values for left and right kidneys for various networks and ensembles with loss functions as DSC ($\mathcal{L}_{DSC}$), cross-entropy+DSC ($\mathcal{L}_{CE+DSC}$), and cosine loss ($\mathcal{L}_{COS}$). The results are post-processed using the largest connected components.

11.2 Supplementary Material Chapter 7

11.2.1 Methods supplement

11.2.1.1 Data augmentation

During training, we apply random rotation, zoom, and translation on the training set. Random rotation is performed with rotation along the Z-axis in the range [-12°, 12°]. Random translation has a range of [-15, 15] voxels for both the X- and Y-axis, while the range for the Z-axis is [-5, 5] voxels. Lastly, random zoom has a zoom range of [0.8, 1.2]. Each linear transformation is applied with a probability of 30% in each iteration.

11.2.1.2 Training

We tested three different networks: ResNet10 [141], ResNet34 (pretrained on MedicalNet [142]) and SEResNet50 [59] with varying network depths, of which all yielded comparably inferior results to our 5 layer Convolutional Neural Network (*CNN*) (Table 7.2).

We follow a 5-fold stratified cross-validation approach to make sure that each training, validation, and test set has the same proportion of dissection and healthy patients. The data split is in a disjoint manner (patient level) with the number of samples in train (72%), validation (8%), and test set (20%) being 117, 13, and 32-33, respectively. We train the network (with random initialization) for 150 epochs, with a batch size of 8. We use Adam optimizer with a learning rate of 1e-4. The objective function is a weighted binary cross-entropy, where the weight for the Aortic Dissection (*AD*) class is (1 - number of AD cases/total number of cases). We further apply L2-regularization on the network weights, having a penalty term of 5e-5. Afterward, we select the model with the lowest validation loss in each fold for application on corresponding test sets.

All the tested networks followed the same training regime, except the pretrained ResNet34 model. Since this model was pretrained on medical data, we trained it on whole volume instead of only aorta Region of Interest (*ROI*).

11.2.1.3 Implementation

The experiments are performed using Pytorch 1.13.1 [167] with Python 3.9.15. We use the Monai (1.0.1) framework [168] to perform data augmentation and model development. Network training and inference were performed on the Nvidia RTX A6000 with a VRAM of 48 GB.

11.2.2 Dataset details

Dataset details are provided in Table 11.2

11.2.3 Clinical details of AD (internal) training cases

Clinical details of AD (internal) dataset are shown in Table 11.3

	Public validation dataset	
	AD cases	**Non-AD cases**
Source	ImageTBAD dataset, Chinese Guangdong Provincial Peoples' Hospital.	AVT dataset [154], originating from the KiTS19 Grand Challenge (excluded in our study), Rider Lung CT dataset [169] and cases from Chinese Dongyang Hospital. Abdomen CT-1K dataset [155] originating from 6 sources: Bilic et al. [170], Heller et al. [171], Simpson et al. [172], Roth et al. [173], Roth et al. [174], Clark et al. [175]
Number of cases	n=100	n=1189
Average age	52.5 ± 11.3	unknown
% female	31%	unknown
CT hardware	Philips (77%) and Siemens (23%)	various
slice spacing (mm)	0.75	0.5/5/5 mm and 0.625/0.625/2.5 mm and 2/3/3 mm
Voxel size	0.25 x 0.25 x 0.25 mm^3	

Table 11.2: Patient characteristics and technical details on external datasets.

Total cases	78
Female	35,4%
Age	66.6 ± 13.39 (29-92)
% presence of suprarenal / infrarenal AD	72,0 / 79,3%
% presence of visceral ischemia	15,9%
% Obstruction celiac trunk / superior / mesenteric artery / left / right renal artery / inferior mesenteric artery / right side / left side limb ischemia	4,9% / 1,2% / 4,9% / 9,8% / 17,1% / 2,4% / 1,2%
% presence of aortic intramural haematoma	9,8%
% presence of aortic (partial) thrombosis	56,1%

Table 11.3: Details of aortic dissection cases in the internal set.

11.2.4 Result supplement

Table 11.4 and Figure 11.2 depict ensemble performance on the external set for an optimal threshold value of 0.745. This threshold value was determined separately for the external dataset. Here, the sensitivity and specificity of 0.940 and 0.993 is achieved, respectively.

11.2.5 Exemplary images of small and subtle cases

Example images containing small and subtle AD are illustrated in Figure 11.3

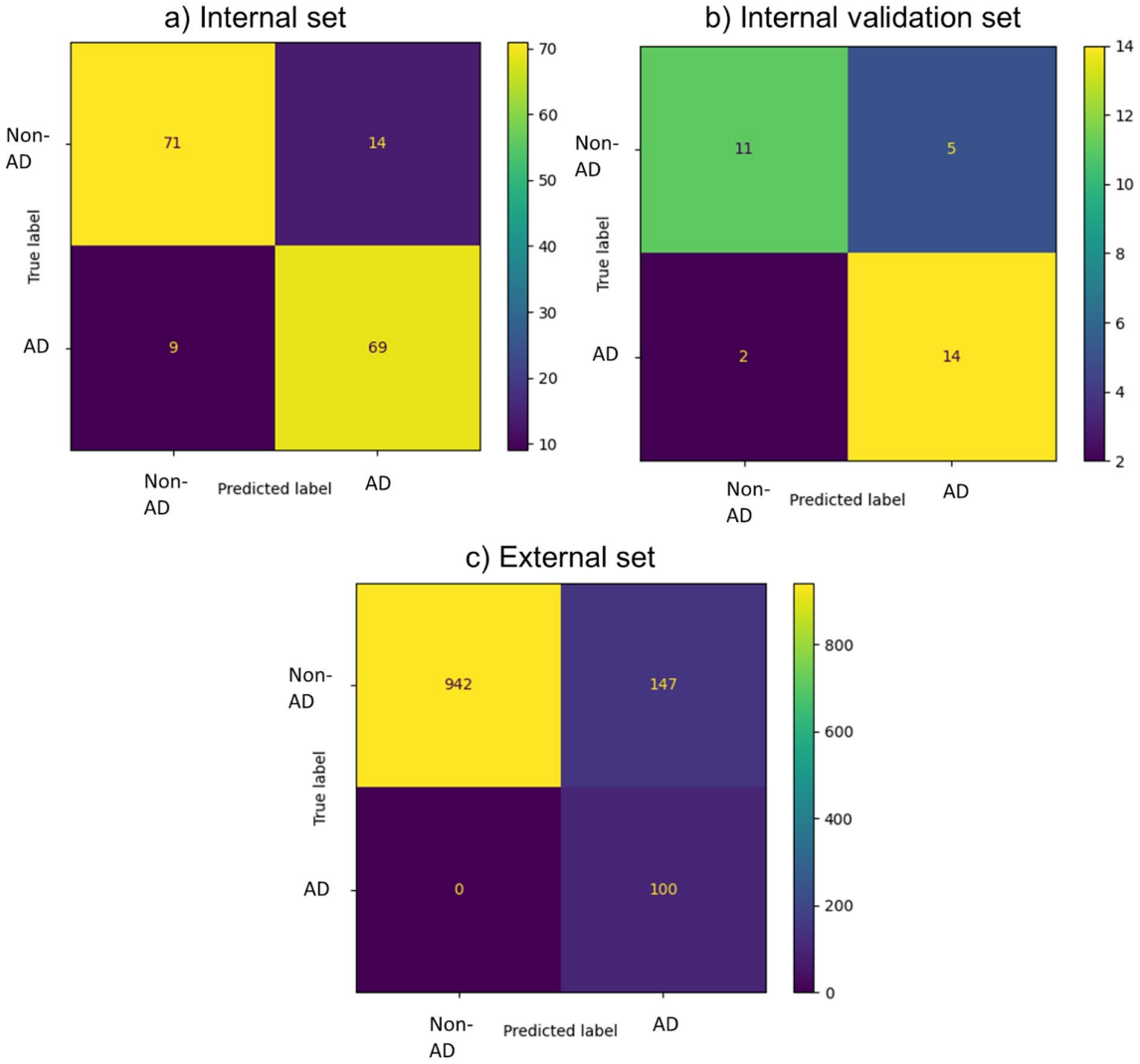


Figure 11.1: Confusion matrices for internal (cross validation), internal validation and external sets. A) Internal set cross-validation, B) internal validation set, and C) external set ensemble results.

Dataset	Sensitivity	Specificity	Balanced accuracy (sensitivity + specificity / 2)	AUC (95% CI)
External (validation)	0.940 (94/100)	0.993 (1082/1089)	0.966 (1.993/2)	0.993 (0.988-0.997)

Table 11.4: External set evaluation metrics for an optimal threshold of 0.745.

11.2.6 Analysis of false negatives on internal training set

Wrong negatives and interpretation

Of the 9 False Negatives (*FNs*), 3 cases had unclear dissections. In these cases, the membrane was not clearly visible/subtle or there was no typical anatomy of dissection visible. In one case, only a small part of the abdomen was extracted, containing only a few slices of the dissected aorta. However, the other six cases had clear dissections. This could be due to the similarity in the aorta size of healthy and dissected cases in this case. The model might consider aorta of specific sizes to belong to a particular class, making it difficult for cases where the aorta sizes are similar in

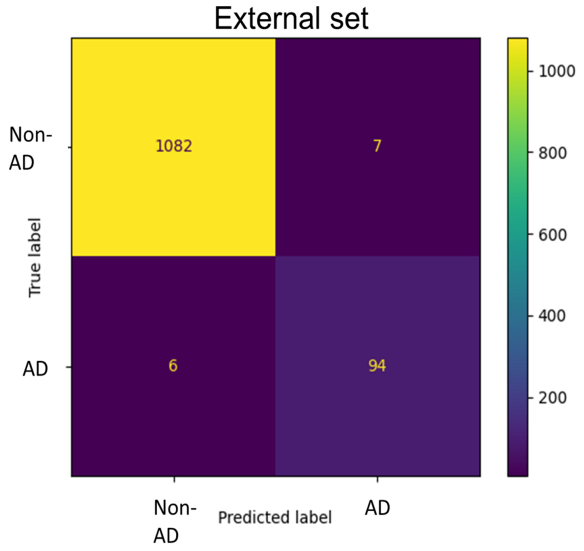


Figure 11.2: Confusion matrix for the external dataset with the optimal threshold of 0.745.

both the healthy and the dissected cases. The reasons behind this, however, remain unclear. Figure 11.4 shows three example cases that were not classified as AD (FN). Interesting findings on each case in the order they appear are:

- Case 1: Dissection anatomy is not typical.
- Case 2: Only part of the abdomen was extracted, not covering a large part of the aorta. However, the dissection has a huge, very broad membrane.
- Case 3: The dissection is clear, and the aorta is large, yet the algorithm failed to correctly classify it.

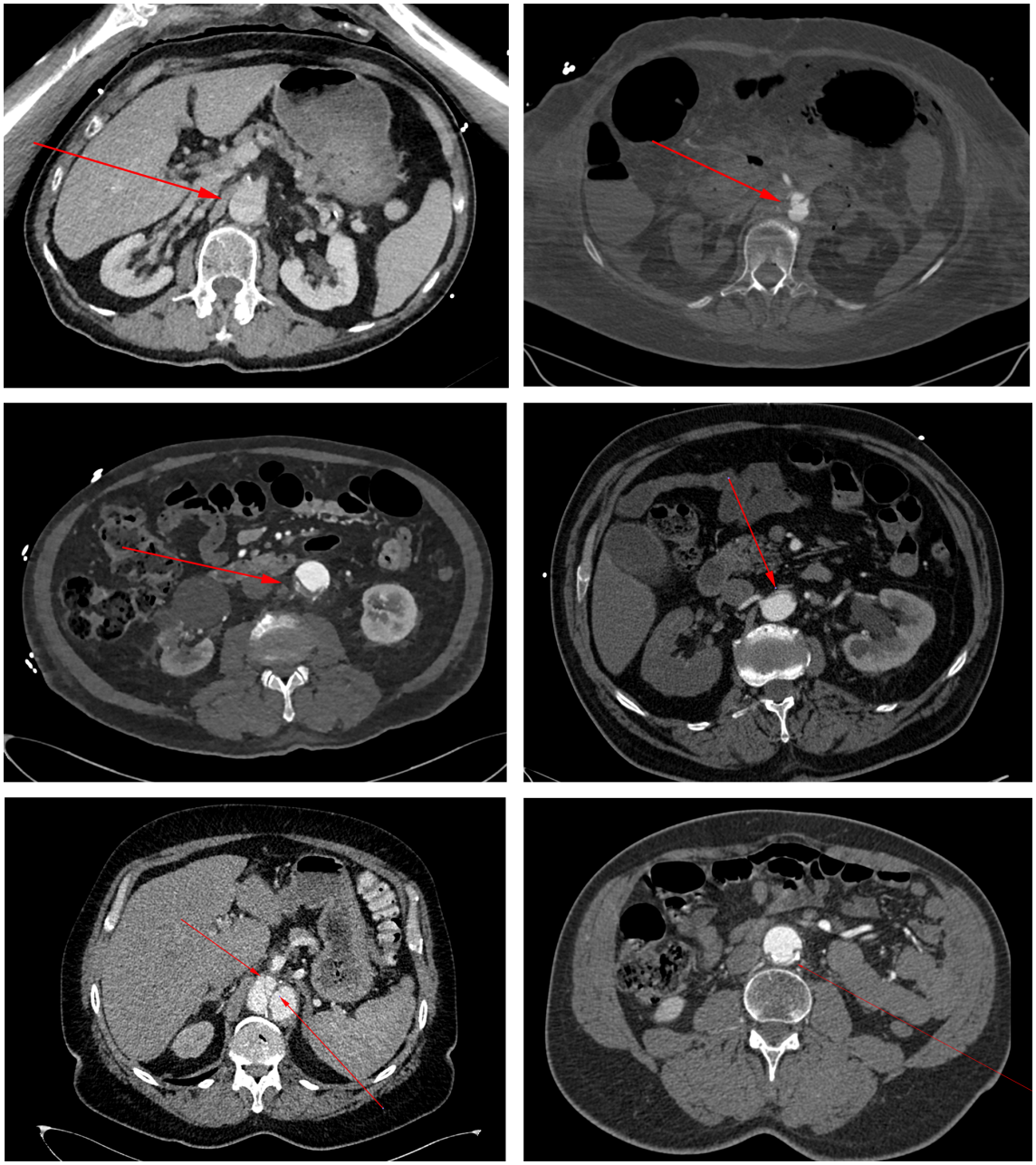


Figure 11.3: Array of atypical and subtle cases from the internal training and validation dataset.

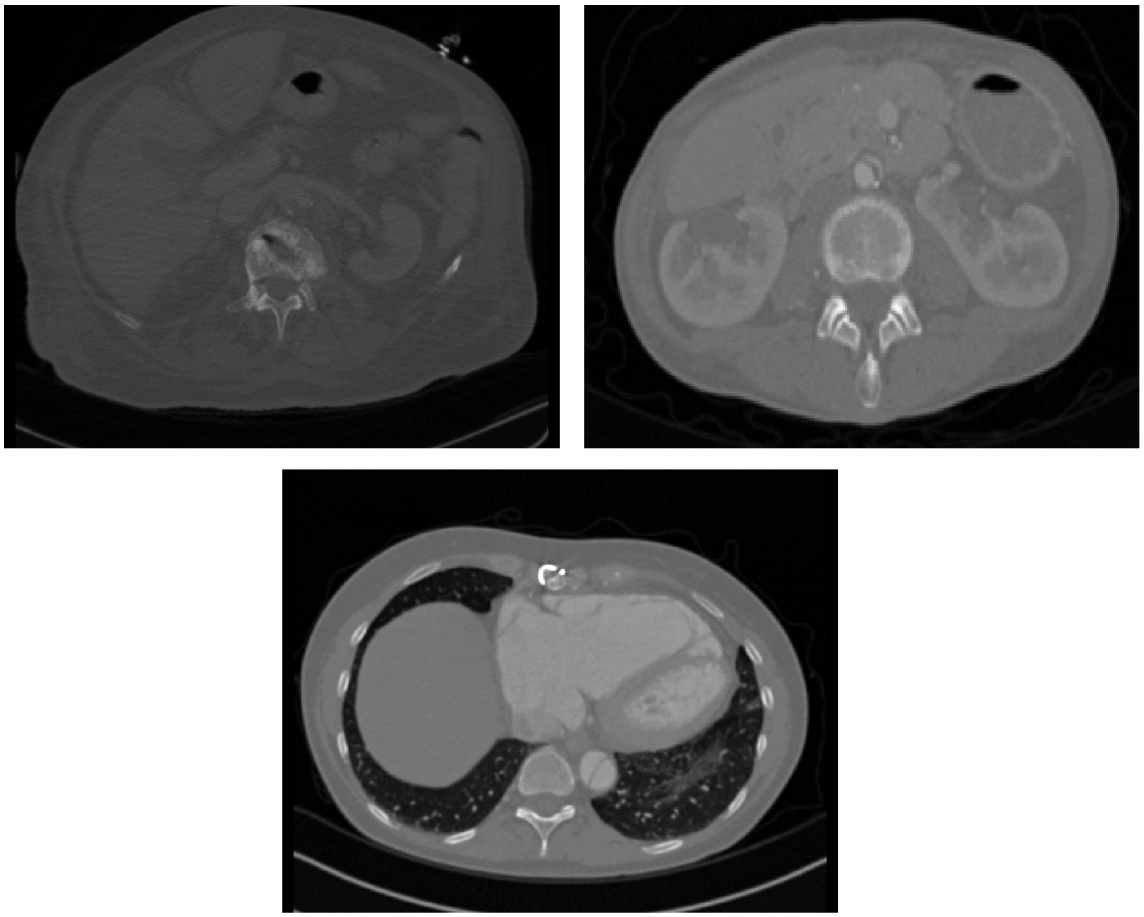


Figure 11.4: Example cases that were falsely classified as non-AD (FN).

12. Publications

Journal Papers

- **Raj, A.**, Allababidi, A., Kayed, H., Gerken, A. LH., Müller, J., Schoenberg, S. O., Zöllner, F. G., & Rink, J. S. (2024). Streamlining acute Abdominal Aortic Dissection management - an AI based CT imaging workflow. Journal of Imaging Informatics in Medicine.
- **Raj, A.**, Gass, A., Eisele, P., Dabringhaus, A., Kraemer, M., & Zöllner, F.G. (2024). A generalizable deep voxel-guided morphometry algorithm for the detection of subtle lesion dynamics in multiple sclerosis. Frontiers in Neuroscience, 18, 1326108.
- **Raj, A.**, Tollens, F., Caroli, A., Nörenberg, D., & Zöllner, F. G. (2023). Automated prognosis of renal function decline in ADPKD patients using deep learning. Zeitschrift für Medizinische Physik, Volume 34, Issue 2, 2024, Pages 330-342,.
- **Raj, A.**, Tollens, F., Hansen, L., Golla, A. K., Schad, L. R., Nörenberg, D., & Zöllner, F. G. (2022). Deep learning-based total kidney volume segmentation in autosomal dominant polycystic kidney disease using attention, cosine loss, and sharpness aware minimization. Diagnostics, 12(5), 1159.

Peer-Reviewed Conference Proceedings

- **Raj, A.**, Gass, A., Eisele, P., Dabringhaus, A., Kraemer, M., & Zöllner, F.G. (2024). A Generalizable Deep Voxel-Guided Morphometry Algorithm for Change Detection in Multiple Sclerosis. Proc. Medical Imaging with Deep Learning 2024, Paris, France, url=https://openreview.net/forum?id=qHrvdiwCmX.
- **Raj, A.**, Hansen, L., Tollens, F., Nörenberg, D., Villa, G., Caroli, A., & Zöllner, F. G. (2024). Generalizable Kidney Segmentation for Total Volume Estimation. Proc. Bildverarbeitung für die Medizin 2024, Erlangen, Germany, pp.285-290, doi: 10.1007/978-3-658-44037-4_75.
- **Raj, A.**, Mothes, O., Sickert, S., Volk, G. F., Guntinas-Lichius, O., & Denzler, J. (2020). Automatic and objective facial palsy grading index prediction using deep feature regression. In Medical Image Understanding and Analysis: 24th Annual Conference, MIUA 2020, Oxford, UK, July 15-17, 2020, Proceedings 24 (pp. 253-266). Springer International Publishing.

Conference Abstracts

- **Raj, A.**, Tollens, F., Strittmatter, A., Hansen, L., Nörenberg, D. & Zöllner, F.G. (2022). Deep Learning based Total Kidney Volume Segmentation in Autosomal Dominant Polycystic Kidney Disease. Proc. ISMRM Congress, London, 2022.
- Caroli, A., Pasini, S., Vandelboe, T., **Raj, A.**, Garcia-Ruiz, L., Strittmatter, A., Echeverria-Chasco, R., Villa, G., Brambilla, P., Hansen, E., Ringgaard, S., Zöllner, F. G., Fernandez-Seara, M., Francis, S. & Laustsen, C (2023). Multicenter and multi-vendor reproducibility of T1, T2 and ADC phantom data on 1.5T and 3T MRI scanners. Proc. ISMRM Congress, Canada, 2023.

- Koçak, M., **Raj, A.**, Sommer, V., Zahn, K., Schaible T., Weis, M., Zöllner, F. (2023). Deep learning-based lung volume segmentation of DCE-MRI data sets of 2-year-old children after congenital diaphragmatic hernia repair. In Book of Abstracts ESMRMB 2023 Online 39th Annual Scientific Meeting 4-7 October 2023. Magn Reson Mater Phy 36 (Suppl 1) (pp. 73-74). Springer. doi: 10.1007/s10334-023-01108-9

Supervised Theses

- **Master's Thesis** "Deep Learning-Based Kidney and Liver Segmentation for Autosomal Dominant Polycystic Kidney Disease patients in T1-weighted Magnetic Resonance Imaging.", Aswathy Biju Sherly, 2023.

13. Curriculum Vitae

Personal Data

Name **Anish Raj**.

Date of Birth **18. April 1994**.

Place of Birth **Jabalpur, India**.

Nationality **Indian**.

Education

04.2022 – today **Doctoral Candidate: Dr. sc. hum.**,
Topic: Deep Learning based Medical Image Analysis using Small Datasets,
Ruprecht Karl University of Heidelberg,
Supervisor: Prof. Dr. Ing. Frank Gerrit Zöllner.

10.2017 – 02.2020 **Master of Science: Medical Photonics**,
Focus: Image Processing, Optics and Machine Learning,
Friedrich Schiller University, Jena, Grade: 1.3.

08.2013 – 06.2017 **Bachelor of Technology: Biomedical Engineering**,
Focus: Medical Imaging and Processing,
Amity University, Gurgaon, Grade: 85%.

04.2010 – 05.2012 **High school**,
Kendriya Vidyalaya, GCF No.1, Jabalpur, Grade: 88%.

Experience

03.2021 – today **Research Associate**,
Computer Assisted Clinical Medicine, Medical Faculty Mannheim, Ruprecht Karl University of Heidelberg,
since 10.2021: Lecturer Biomedical Engineering.

02.2020 – 01.2021 **Research Associate**,
Continental AG & Computer Vision Group, Jena,
- Real-time road condition estimation using CNNs and specular reflection maps..

10.2018 – 01.2020 **Internship and Master's Thesis**,
Computer Vision Group, Jena,
Topic: Predicting Facial Paralysis Grades via 2D Face Images with the aid of Machine and Deep Learning, Grade: 1.0.

Scholarships & Grants

2022 **ISMRM Educational Stipend**.

14. Acknowledgements

Embarking on this PhD journey has been a profoundly transformative chapter in my life, enriching me professionally and personally in ways I had never imagined. Reflecting on this journey, I realize it has been made possible by the unwavering support and contributions of some incredible individuals to whom I owe my deepest gratitude.

First and foremost, my sincere thanks to Prof. Dr. Lothar Schad for giving me this opportunity and welcoming me into his research group.

I am immensely grateful to Prof. Dr. Frank Zöllner for his supervision and guidance throughout this thesis. His support was pivotal in navigating the challenges of research, and his feedback was instrumental in refining my methodologies and enhancing the quality of our publications.

I extend my gratitude to the people at CKM. During my time there I enjoyed the opportunities to grow and learn.

My journey would have been much harder without the early support of my former colleagues: Alena, Dominik, Tom, and Christian. Your warm welcome laid the foundation of my PhD experience. Alena, your deep learning framework helped me save a lot of time with experimental setups and prototyping. Furthermore, I am equally thankful to my current colleagues: Anika, Christoph, and Zeyu, for their constant help and support. The lunch breaks with you have always been fun, filled with the exchange of good ideas and funny conversations. Additionally, your kindness in helping me improve my German skills has made me feel more at home in our shared environment.

I am grateful for the support and kindness my friends Abhishek, Kamal, Kunal, Mubarak, and Veedesh have provided me throughout the years. I cherish the friendship we have built over these years.

I want to thank my wife, Monika, for being boundlessly kind, loving, and supportive from day 1. This journey would have been less enriching without your presence.

Last but not least, I am forever indebted to my parents and brother. They have been a beacon of kindness and empathy throughout my life. I would not be here if not for their unconditional love and hard work.