Automatic Migration Percentage Measurements on Ultrasound Images

by

Reza Yousefvand

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Software Engineering and Intelligent Systems

Department of Electrical and Computer Engineering
University of Alberta

Abstract

Hip displacement is a prevalent disorder in children with cerebral palsy, defined as the lateral displacement of the femur head from under the acetabulum, and leads to severe pain and difficulties in daily activities. As a result, hip surveillance programs have been developed to monitor and diagnose hip displacement in children with cerebral palsy, and ensure that appropriate interventions are made at the right time. These programs involve regular assessments of hip displacement. Migration Percentage (MP), defined as the ratio of the distance between the lateral borders of the femur head and the acetabulum (A) to the total width of the femur head (B), is the gold standard parameter of hip displacement measurement and is measured using anteroposterior X-ray imaging (MP=A/B). However, the frequent X-ray imaging in hip surveillance programs exposes children to ionizing radiation, increasing the risk of cancer development later in life.. Recently, ultrasound (US) has been proposed as a non-invasive and widely accessible alternative for hip assessments. Yet, the inherently fuzzy and noisy nature of US images makes accurately identifying edges and landmarks challenging, leading to time-consuming and user-dependent manual measurements. To tackle these issues, this study aimed to develop a fast, reliable, and fully automatic algorithm to measure the MP from US images.

In the developed methodology, for each of the "A" and "B" measurements, UNets were trained to segment the hip features on the coronal hip scans. A convolution neural network was trained to score and select frames for measurement. More UNets were trained to identify the measurement features on the selected frames, and statistical analysis was applied to aggregate and finalize the measurements.

To verify the developed method, a total of 38 children with an average age of 9±3.4 years old were recruited, and 62 hips were scanned in total. From the 62 scanned hips, 36 were utilized for training, 8 for validation, and 18 for testing. An experienced rater provided the X-ray measurements for all scanned hips as the ground truth. The mean absolute difference (MAD) and intra-class correlation coefficient (ICC(2,1)) of the test measurements were 6.5% ± 5.5% and 0.86. The clinical acceptance rate was 72%, and the sensitivity and specificity of classification of displaced hips (MP>30) were found to be 100% and 93%. The measurement time for each hip was 105.6 seconds on average, which was 3 times faster than manual measurements. Hence, the developed method demonstrated good reliability, accuracy, and speed in MP measurements, marking a significant step towards replacing X-ray with US in hip assessments.

Preface

This thesis represents original work conducted by the author, Reza Yousefvand. The research project underlying this thesis received ethics approval (Pro000111361) from the University of Alberta Health Research Ethics Board.

Portions of the material in Chapters 4 and 5 have been submitted in the following paper:

- Yousefvand, R., Pham, T-T., Le, L.H., Andersen, J., Lou, E. H, "Applying Deep Learning for Automatic Measurement of Migration Percentage from Ultrasound Images in Children with Cerebral Palsy," Medical & Biological Engineering & Computing, 2024, (Submitted).

I developed the methodology, conducted the analysis, and authored the manuscript. Dr. Edmond H. Lou provided supervision, review, and editorial guidance. Thanh-Tu Pham contributed data essential to the study. Drs. Lawrence H. Le and John Andersen offered valuable insights into the medical aspects of this research.

Acknowledgements

First and foremost, I wish to express my profound gratitude to my supervisor Dr. Edmond Lou for his trust during my most challenging moments, and for his endless support, patience, and guidance.

I sincerely thank Drs. Lawrence Le and Jun Jin for taking the time to read this thesis and provide insightful comments and feedback.

I extend my appreciation to my teammate, Thanh-Tu Pham, for her collaboration and assistance, especially with the data collection.

Special thanks to funders of this research, Glenrose Hospital Foundation, Alberta Health Services, Stollery Children's Hospital Foundation, and Women & Children's Health Research Institute. I am also thankful to Industry Sandbox & AI Computing for providing the essential computational resources.

I would like to heartfully acknowledge my friends in Iran, Canada, and around the world for their continuous emotional support.

Lastly, my deepest gratitude goes to my dearest family, especially my beloved parents. Their enduring love and support have been my foundation, and I am forever grateful.

Contents

List of Tables

List of Figures

Abbreviations

BCE Binary Cross Entropy.

CA Clinical Acceptance.

CDH Congenital Dislocation of the Hip.

CNN Convolutional Neural Network.

CP Cerebral Palsy.

DBSCAN Density-Based Spatial Clustering of Applications with Noise.

DDH Developmental Dysplasia of the Hip.

GMFCS Gross Motor Function Classification System.

ICC Intraclass Correlation Coefficient.

LHC Lateral Head Coverage.

LHD Lateral Head Distance.

LHDR Lateral Head Distance by Radiography.

LoA Limits of Agreement.

MA Moving Average.

MAD Mean Absolute Difference.

MP Migration Percentage.

RNN Recurrent Neural Network.

SCCE Sparse Categorical Cross Entropy.

US Ultrasound.

Chapter 1

Introduction

1.1 Background

Cerebral palsy (CP) is a neurological condition caused by non-progressive brain damage, resulting in movement disorders [1]. According to 49 relevant studies, the pooled average prevalence of CP is estimated at 2.11 per 1,000 live births, with 95% confidence intervals of 1.98-2.25 [2]. Although the damage is neurological, CP creates various musculoskeletal issues, with hip displacement being the second most prevalent one, affecting approximately 35% of the population [3]. Hip displacement is defined as the gradual displacement of the femur head from under the acetabulum [4]. Hip displacement can cause significant physical and functional limitations, including chronic pain, osteoarthritis, and a severely diminished health-related quality of life [5–7], and if left untreated, it could lead to complete dislocation of the femur head. Hip dislocation is preventable through early identification and appropriate intervention [8], which includes postural management, orthoses, tone management and surgery. Interventions should be selected in agreement with the child's clinical and functional status [9]. As a result, hip surveillance programs have been developed to monitor and track the hip displacement progression, incorporating regular hip assessments. Currently, The gold standard for diagnosing hip displacement is the migration percentage (MP), defined as the ratio of the distance between the acetabulum and the femur head ("A") to the total width of the femur head ("B") [4, 10].

The accepted method for MP measurement is anteroposterior X-ray imaging. Although a single radiograph of the hip has a radiation dose of 0.1-0.7 mSv, regular exposure to ionizing radiation during surveillance programs increases the risk of cancer development in later stages. The risk is even higher for children, especially when the exposure area is close to the reproductive organs. Ultrasound (US) is a widely accessible, safe, and non-invasive diagnostic tool that uses mechanical waves for imaging and does not harm the tissues. A recent study has proposed a method for measuring MP from US scans, suggesting the measurement of the lateral head distance from coronal scans and the total width of the femur head from transverse scans [11].

1.2 Motivation

Each US hip scan contains 100-1000 frames, with only a few suitable for measurement. The inherent fuzziness and noise in US images make identifying the necessary features for measurement time-consuming and user-dependent. For a single hip, the measurement process could take up to 10 minutes. Therefore, developing a fully automated method could conserve valuable clinician time, standardize measurements, and pave the way for replacement of radiography with US imaging in hip surveillance programs.

1.3 Objectives

1. To precisely label and prepare a dataset for developing an automated AI-based algorithm for MP measurements using US images.

2. Develop a model for automatic and fast measurement of lateral head distance ("A") from coronal US scans.

3. Develop a model for automatic and fast measurement of total width of the femur head ("B") from transverse US scans.

4. To evaluate and analyse the developed models and measurements.

1.4 Thesis Outline

This thesis consists of five chapters. It starts with an introduction to hip displacement, surveillance programs, and MP in Chapter 1. The objectives of this research are also listed.

Chapter 2 provides in-depth background about CP, hip displacement, its prevalence and consequences, prevention methods, and hip surveillance programs. Furthermore, a literature review on MP, including its accuracy and reliability, is given. It also includes details about related studies on measurement modalities, including X-ray and US. Finally, an overview of the applications of machine learning and deep learning in hip assessments is provided.

Chapter 3 presents three initial studies on manual measurements and labeling by the author. These studies establish the author's intra- and inter-rater reliability in 1) MP measurements using X-ray, 2) frame selection in US scans, and 3) lateral head distance measurement in US frames. The purpose of these studies was to provide the author with the required background and reliability for manual labeling.

Chapter 4 reports the details of dataset preparation, labeling, model development, and the developed measurement algorithm.

Chapter 5 describes the training procedure and model optimization for the proposed method, the performance of the trained networks, the results of the test dataset in "A", "B", and MP measurements, and a detailed discussion of the results and comparison with the literature.

Chapter 6 summarizes the thesis work, provides concluding remarks about the research, and suggests future recommendations.

Chapter 2

Literature Review

Cerebral Palsy (CP) is the leading cause of disability in children, and hip displacement ranks as the second most prevalent orthopedic problem in children with CP. Given this, many studies have been conducted on the diagnosis and prevention of hip displacement. In this chapter, we present a literature review of the related studies. We begin in Section 2.1 with an introduction to CP, outlining the diagnostic criteria and the various levels of severity. Section 2.2, focuses on hip displacement and dislocation, highlighting preventive strategies and the pivotal role of migration percentage (MP) in diagnostics. Section 2.3 discusses the reliability of using MP in radiographical studies. Section 2.4 reviews studies on the use of ultrasound (US) in hip assessments, comparing it with the radiological method. Finally, in Section 2.5, we explore studies employing machine learning in hip examinations.

2.1 Cerebral Palsy

CP is defined as a group of disorders that affect the ability of a person to move and maintain balance and posture. It is caused by abnormal development or damage the during fetal or neonatal period. The motor impairments seen in CP frequently occur alongside issues with sensation, perception, cognitive functions, and communication, as well as behavioral disturbances. Individuals with CP may also experience epilepsy and additional musculoskeletal complications [12]. Although the damage to the brain

is static, the consequent musculoskeletal problems can vary over time [3, 10]. The main features of such problems are joint instability, torsional deformity of bones, and muscle-tendon contracture [13]. The prevalence of CP is approximately 2 per 1000 among newborn infants, making it the primary cause of physical disability affecting children [3, 10, 13].

Identifying CP is primarily based on clinical evaluations. Key indicators that can collectively point towards a CP diagnosis include delay in achieving motor milestones, irregular neurologic assessments, retention of primitive responses, and abnormal postural reactions [14].

2.1.1 Gross Motor Function Classification System

The Gross Motor Function Classification System (GMFCS) is a five-level classification system used to assess the severity of CP based on the individual's movement abilities. Level I indicates normal mobility, while level II indicates the ability to walk independently but with limitations in running or jumping. Children classified at level III require assistive devices to walk and use a wheelchair for longer distances. Level IV indicates limited walking abilities and a reliance on a wheelchair for mobility, and level V indicates an inability to sit, walk, or stand independently [15].

2.2 Hip Displacement and Dislocation

Hip displacement is the displacement of the femur head from the lateral border of the acetabulum. Hip dislocation refers to the condition when the femoral head is completely displaced from under the acetabulum [16].

2.2.1 Hip Displacement in Children with Cerebral Palsy

In children with CP, hip displacement is the second most frequent musculoskeletal problem after equinus, affecting approximately 35% of children with CP, compared to 0.2% in the general population [3, 10].

Children with CP have normal hips up to the age of approximately 18 months, but the spasticity, contracture and imbalance of the hip muscles leads to gradual displacement of the femur head from under the acetabulum and increased pain and functional disability over time [3, 17]. As a result of the asymptomatic nature of the hip displacement and increased focus on other problems such as seizure and feeding difficulties, hip displacement is usually diagnosed late [3].

2.2.2 Migration Percentage

Migration percentage (MP) refers to the distance between the lateral margin of the femur head and the Perkins' line (A) divided by the total width of the femur head (B, measured parallel to Hilgenreiner's line) multiplied by 100. Hilgenreiner's line is the line connecting the triradiate cartilages, and Perkins' line is drawn perpendicularly to Hilgenreiner's line from the most lateral point of the acetabulum [4, 18]. Figure 2.1. depicts a schematic view of the pelvis, Hilgenreiner's and Perkins' lines, and "A" and "B" distances.

MP is considered the gold standard method [10] for hip displacement measurement because of its higher intra- and inter-observer agreement (intraclass correlation coefficient (ICC) = 0.95-0.97 and 0.91-0.93) compared to other hip assessment parameters such as the acetabular index (intra-rater ICC = 0.91-0.92 and inter-rater ICC = 0.80-0.81) and the femoral neck-shaft angle (intra-rater ICC = 0.76-0.95 and inter-rater ICC = 0.58-0.89) [20, 21]. Additionally, MP is easier to measure, especially when the femur head is not entirely spherical and less influenced by the rotational position of the femur head [17].

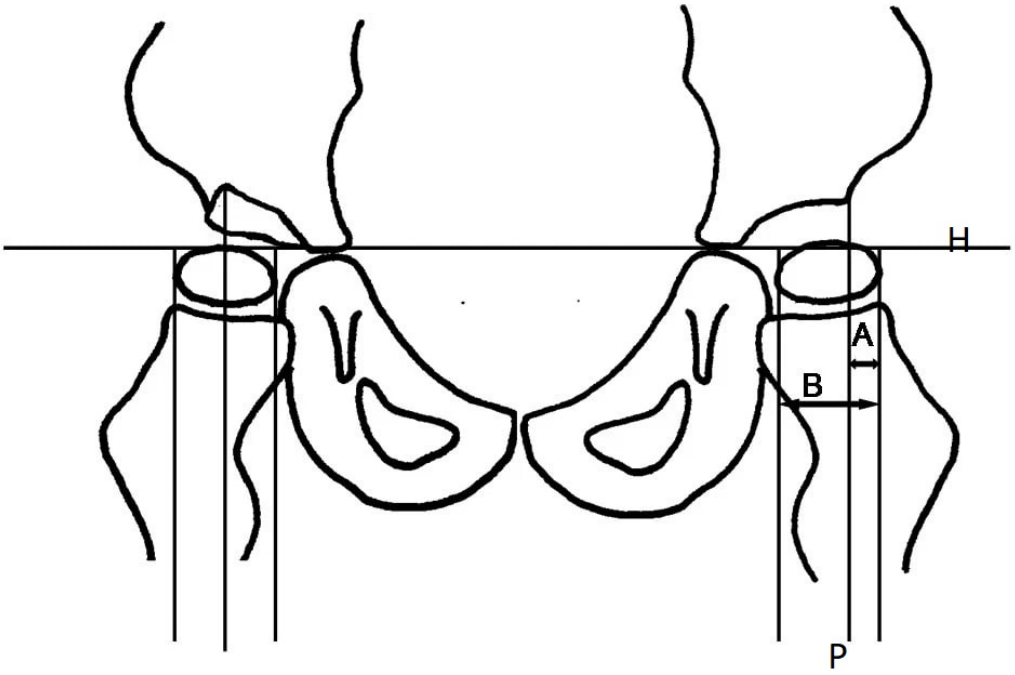


Figure 2.1: A schematic view of pelvis. The Hilgenreiner's and Perkins' are indicated as H and P, and "A" and "B" distances are depicted [19].

2.2.3 Definition of Hip Displacement and Dislocation by MP

There have been slightly different definitions of hip displacement and dislocation based on MP value. In [16], hip displacement or subluxation is defined as 10%<MP<99% and hip dislocation is defined as MP>100%. In [22, 23], 33%<MP<80% denotes hip displacement and hip dislocation refers to MP>80%. The most commonly accepted definition of hip displacement in the literature is an MP>30% [3].

A classification scale based on MP is proposed to describe hip morphology in patients with CP for both clinical and research purposes [24]:

- I: Normal hip – MP<10%
- II: Near normal hip – 10%<MP<15%
- III: Dysplastic hip - 15%<MP<30%

- IV: Subluxated hip - 30%<MP<100%
- V: Dislocated hip – MP>100%
- VI: Salvage surgery

2.2.4 MP Progression and Hip Displacement Factors in CP Children

Terjesen *et al.* [17] studied 76 children with CP and reported gait function level and age as the most influential variables affecting MP progression. The children who could not walk showed an MP progression rate of 12% per year, while for those who could walk with or without support, the progression rate was only 2% per year. In addition, MP progression in younger children tended to be higher than in older children.

Research has shown correlation between the GMFCS level and the severity and incidence of hip displacement. Robin *et al.* [13] measured the overall MP to be 8.1%, 13.0%, 25.0%, 36.8%, and 46.2% for GMFCS level I to level V, respectively. Soo *et al.* [3] reported the incidence of hip displacement, defined as MP>30%, to be 0%, 15,1%, 41.3%, 69.2%, and 89.7% GMFCS level I to level V, respectively. Figure 2.2 demonstrates the mean MP and the incidence of hip displacement in different levels of GMFCS. In addition, 6.5% of children with CP in the study had hip dislocation, all having a GMFCS level of IV or V, and the incidences were 12% and 26%, respectively.

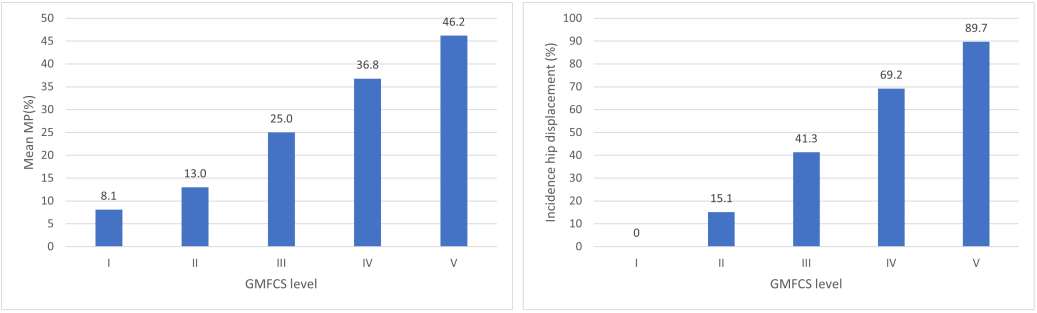


Figure 2.2: The mean MP (left) and the incidence of hip displacement (right) according to different levels of GMFCS [3, 13].

2.2.5 Hip Displacement Consequences

Progressive hip displacement can lead to hip dislocation [16]. Hip dislocation results in severe pain and significant problems with mobility, sitting balance, perineal hygiene, and decubitus ulceration [13, 17]. Unilateral hip dislocation is associated with the development of scoliosis, and pelvic obliquity [3, 22]. Some authors have reported that an MP progression rate of 7% or greater can be correlated with a future inability to walk [25]. An MP of 15% at 30 months of age carries 50% chance of developing hip dislocation, while an MP of 60% is considered unstable and requires immediate action [26].

2.2.6 Hip Dislocation Prevention

Hip dislocation can be prevented by surveillance, early identification, and taking appropriate measures. Studies have demonstrated a substantial decrease in hip dislocation incidence by implementing a prevention program [8, 27]. Early intervention has been reported to yield better long-term outputs and decrease the chance of treatment failure in children at the risk of developing hip displacement. Onimus *et al.* [28] reported the chance of a successful operation to be 90% when the MP is less than 33%, and the patient has an age of less than 4 years. Hence, early diagnosis of hip displacement in children with CP is crucial.

2.2.6.1 Surveillance and Early Identification

Numerous research studies suggest that hip surveillance is a helpful strategy for preventing hip dislocations [26]. Hip surveillance involves identifying any early indications of progressive hip displacement, such as the child's GMFCS level, age, degree of gait function, and high value of MP on radiological examination [16]. Certain studies propose that all children with spastic quadriplegia and those unable to walk independently before reaching 30 months of age should undergo hip surveillance [23, 26]. Other studies suggest that hip surveillance programs should begin as early as

12 months of age [25]. A preventive intervention is recommended before reaching an MP level of 60% in the literature [27].

2.2.6.2 Radiological Examination

The best way to evaluate the degree of hip displacement is by measuring MP using standardized procedures [4, 26]. The standard procedure involves taking pelvic X-ray images of the patient while they are lying in supine position, ensuring both symmetry in the anterior/posterior tilt and the neutral positioning of the femurs relative to the pelvis [20], as shown in Figure 2.3. Regular measurement of MP is part of the hip surveillance programs. The recommended age of the initial radiograph and the frequency of subsequent radiographs is variant among different studies, *e.g.*, starting from 12 months of age and repeating every 6 or 12 months until 8 years of age or skeletal maturity [25], starting from 18 months of age [27], beginning from 30 months of age and repeating every 6 months [23] or every 12 or 24 months [3]. In some sources, the suggested frequency depends on age, GMFCS level, and MP [16].

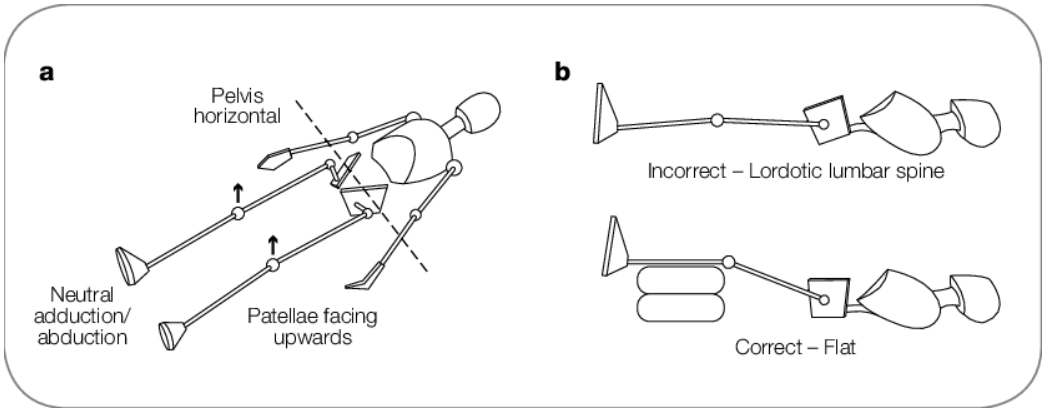


Figure 2.3: Standard positioning for pelvic X-ray imaging [16].

2.3 Studies on MP Measurement on Radiographs

2.3.1 MP Reliability

There have been several studies on the reliability of MP measurements from radiography. Initially, there was some skepticism around the reliability of MP since in one study, the 95% confidence interval for intra- and inter-rater errors were reported as high as 13% and 22.4%, respectively. However, the raters had a limited experience which could have adversely affected the results [29]. Subsequent studies showed that MP is a reliable parameter if measured appropriately. Parrott *et al.* [20] conducted a study with 5 experienced raters and hip radiographs of 20 children to quantify the repeatability of MP measurements. Their study showed that an experienced rater is expected to measure MP within ±5.8% and ±8.3% of the actual value on a single radiograph and on two radiographs of the same hip taken at different times, respectively. Availability of digital images like DICOM files with zooming ability and software-aided measurements has enabled more accurate MP measurements. A study with 16 raters from around the world obtained a 95% confidence interval of 6.4% for MP measurements from digital X-ray images [18]. Another study found the intra- and inter-rater error for software-assisted digital measurements of MP to be 3-7% and 1%, respectively [30]. Kim *et al.* [31] compared a modified method of MP measurement that takes into account the lateral edge of the acetabular sourcil, with the classic method of MP measurement, which uses the acetabular roof as the reference point. Their results showed that the classic method offers more reliability in terms of both intra- and inter-rater reliability, specifically the latter one.

2.3.2 Effect of Positioning on MP Measurements

Cliffe *et al.* [32] showed that with correct positioning under the supervision of an experienced pediatric radiographer and using wedges and pillows for maintenance, the mean variation for a single observer was 3.2% (SD 3.5) and 3.3% (SD 3.2) for

the repeated measurement. In addition, the mean difference for different observers was 3.7% (SD 3.8). Terjesen *et al.* [33] measured the mean variation in MP between supine and standing positions to be less than 1% in individuals who have already been treated for hip dislocation.

2.3.3 Threshold of MP for Operative and Non-operative Intervention

Hägglund *et al.* [34] studied the threshold value of different radiological parameters, including MP, for operative and non-operative intervention in children with CP. They analyzed 1067 radiographs of 272 children aged 6.5-13.5 years at the last examination and recommended 33%<MP<40% as an indicator for non-operative intervention and MP≥40% as the threshold for operative intervention.

Table 2.1 shows a summary of radiological studies on MP measurement.

2.3.4 Limitations

The use of radiography in medical imaging exposes patients to ionizing radiation, which imposes many risks to their health, including an increased chance of developing cancer later in life. The associated risk increases if the patient is younger. As a result, a safer alternative is required for hip surveillance programs in children.

Table 2.1: Summary of radiological studies on MP measurement.

	Num of hips	Num of raters	intra-rater reliability	inter-rater reliability
Faraj *et al.* [29]	44	2	Intra-session: MeAD=3.2% (95^{th}percentile=12.7%) Inter-session: MeAD=1.7-3.2% (95^{th}percentile=12.6-12.9%)	MeAD=2.8% (95^{th}percentile 22%)
Parrott *et al.* [20]	20	5	ICC=0.95-0.97 MD=-1.94-2.78% (SD=1.59-5.75%)	ICC=0.91-0.93 SD=1.28-6.45%
Shore *et al.* [18]	50	16	ICC = 0.95 MAD=3.2% (SD=2.9%)	ICC = 0.94
Segev *et al.* [30]	20	10	SD=3.31–7.91%	ICC=0.83–0.92 SD=1.03–1.04%
Terjesen *et al.* [33]	102	2	MD=-1.3--0.2%	MD=-1.2-1.5% (SD=2.6-3.7%)
Kim *et al.* [31]	152	2	ICC = 0.94-0.97 MD=-1.94-2.78 (SD=1.59-5.75%)	ICC = 0.95
Cliffe *et al.* [32]	40	2	intra-session: ICC = 0.97 MD=3.2% inter-session: MD=3.3%	ICC = 0.96 MD=3.7% (SD=3.8%)

Num, number; MeAD, median absolute difference; MD, mean difference; MAD, mean absolute difference; ICC, intraclass correlation coefficient.

2.4 Ultrasound for Hip Assessment

US is an inexpensive and widely accessible diagnostic tool. The primary advantage of US is that it does not expose children to ionizing radiation, which makes it a safe and repeatable option for medical imaging. Since in neonates and children under 12

months of age, most parts of the pelvis are cartilaginous and not ossified, US can be utilized for the diagnosis and follow-up of hip dysplasia [35]. The application of US as a diagnostic tool for hip examination in infants with potential congenital hip dysplasia is widely acknowledged [36–38]. However, in older children, visualization of most of the acetabulum is impossible in US imaging due to the extensive ossification center within the femoral head. Despite this limitation, since other bony and soft components of the hip joint are visible in US images, it is possible to evaluate the extent of femoral head coverage by the acetabulum by using US images [39].

Terjesen *et al.* [39] realized in their follow-up examinations of children with congenital dislocation of the hip (CDH) that US can be a useful diagnostic tool even for children above two years of age. They conducted a study to evaluate to what extent US can replace radiography in CDH diagnosis. The 95% confidence interval for determining an unknown distance from a single US measurement was found to be the observed value (mm) $\pm$ 1 mm.

2.4.1 Lateral Head Distance

Lateral head distance (LHD) is defined as the distance between the lateral tangent of the ossification center of the femoral head to the lateral bony rim of the acetabulum. LHD corresponds to the variable "A" in MP definition [35, 39–42]. Figure 2.4. illustrates a US image and LHD on a schematic view of the hip.

LHD has been widely explored as a parameter for hip screening in literature. Terjesen *et al.* [39] reported >7mm, >8mm, >10mm, and >12mm as the LHD threshold for hip displacement in children aged 2-3, 4-7, 8-12, and $\geq$13 years, respectively. The resulting accuracy for normal (MP<33%) and dislocated (MP>100%) hips was 99.5% and 100%, respectively, while for subluxated hips (30%<MP<100%) it was only 57.1%. In a subsequent study, Agnar *et al.* [42] obtained a specificity of 89%, but only after excluding a substantial number of uncertain cases in comparison to the abnormal ones. (9 vs. 17). Šmigovec *et al.* [40] reported >5mm and >4.8mm

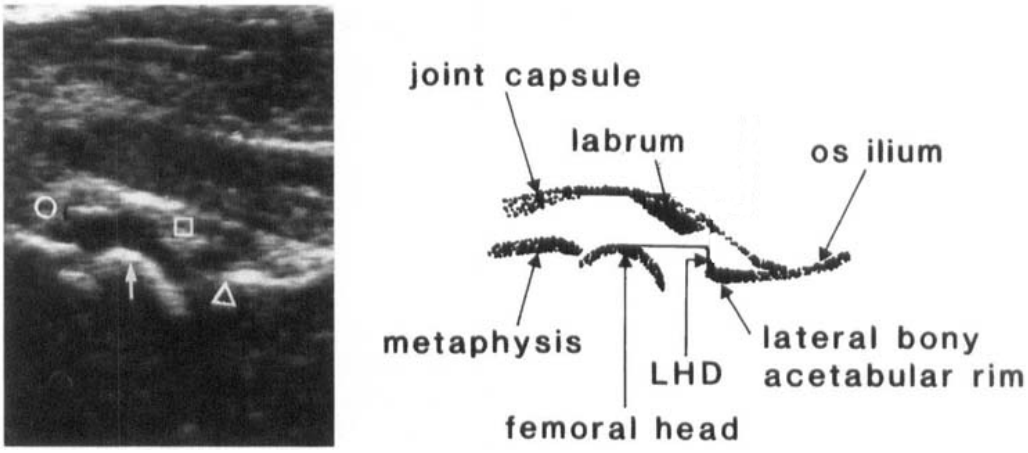


Figure 2.4: A sagittal US image (left) and schematic view (right) of the right hip sonogram), indicating the lateral outline of the femur head (arrow), lateral bony acetabular rim (triangle), joint capsule (circle), and labrum (square). LHD is depicted in the schematic view [41].

as the LHD threshold for hip displacement (MP >33%) in children aged 24-60 and >60 months. Although the resulting classification rate was relatively high (86%), the positive predictive was low, especially in children aged >60 months (46.2%), which means there is a high probability (44.8%) of MP<33% when LHD>5mm.

2.4.2 Comparison of Ultrasound and Radiography measurements

Lateral head distance by radiography (LHDR) is defined as LHD measured from radiography. There have been a few papers in the literature that compare LHD with LHDR. Terjesen *et al.* [39] reported the mean discrepancy, the 90% confidence interval, and the correlation coefficient of LHD and LHDR to be 0.3mm (SD 1.65-1.84), 3mm, and 0.84, respectively. Agnar *et al.* [41] reported -0.3mm, 3mm, and 0.59 as the mean difference, 95% confidence interval, and correlation coefficient of LHD and LHDR in a study of 30 hips, respectively. Another study reported the mean difference and the correlation coefficient of LHD and LHDR measurements as 0.7 (SD 1.6) and 0.85, respectively [42].

To the best of the author's knowledge, only a few studies exist on intra- and

inter-observer reliability of LHD measurements by US. Šmigovec *et al.* [40] reported that the difference between the two raters was within ±1.0mm in 95% of the LHD measurements. The ICC of two raters in LHD measurements was reported as 0.84 [39], 0.94 [35], and 0.98 [40] in the literature.

2.4.3 MP measurement on Ultrasound

Since complete visualization of the femur head is not possible in US images, direct measurement of MP has not been explored until recently. However, the correlation coefficient between LHD and MP was reported as 0.8 [39] and 0.79-0.83 [42].

Kay *et al.* [43] introduced a new index called lateral head coverage (LHC), defined as the portion of the femur head not covered by the acetabulum. LHC was expected to be inversely correlated to MP. In order to measure LHC, a sphere was fitted to the visible parts of the femur head in 3D US images of the hip, and its diameter was taken as an estimation of the total width of the femoral head. They measured MP and LHC from radiographs and sonograms of 24 hips and found that MP is correlated with LHC with a correlation coefficient of -0.86. In addition, the inter- and intra-class correlation coefficients of the measurements were 0.973 (95% CI 0.925-0.998) and 0.982 (95% CI 0.976-0.991).

Pham *et al.* [44] conducted experiments by varying the LHD in two 3D printed hip phantoms for different MP values. Transverse (scanning along the plane that divides the body into left and right halves) and coronal (scanning along the plane that divides the body into front/anterior and back/posterior sections) scans were performed by a 2D US device to measure the variables "A" and the "B". "B" was then estimated by fitting a circle using Taubin's method [45] to visible parts of the femur head in sagittal view. Radiological measurements of MP on these phantoms were obtained as well. In the US measurements, the errors recorded for the two phantoms were 2.20% and 0.68%, while in the X-ray measurements, the errors were higher at 3.23% and 9.83%, respectively. In a subsequent study, Pham *et al.*[11] evaluated their developed

method on 10 hips from children with CP. The MAD between the MP^{US} and MP^{Xray} measurements was 3.5% ± 2.8%, demonstrating good agreement between the imaging modalities. Figure 2.5 displays the Anteroposterior X-ray, coronal and transverse US scans of a participant's left hip, and the measurements "A" and "B". "A" represents the vertical distance (in the rotated coronal image) between landmarks indicating the lateral margins of the acetabulum and femoral head. "B" is the diameter of a circle fitted to the femoral head's upper edge in the transverse US scan. Pham's experiments showed the potential of making accurate MP measurements using US.

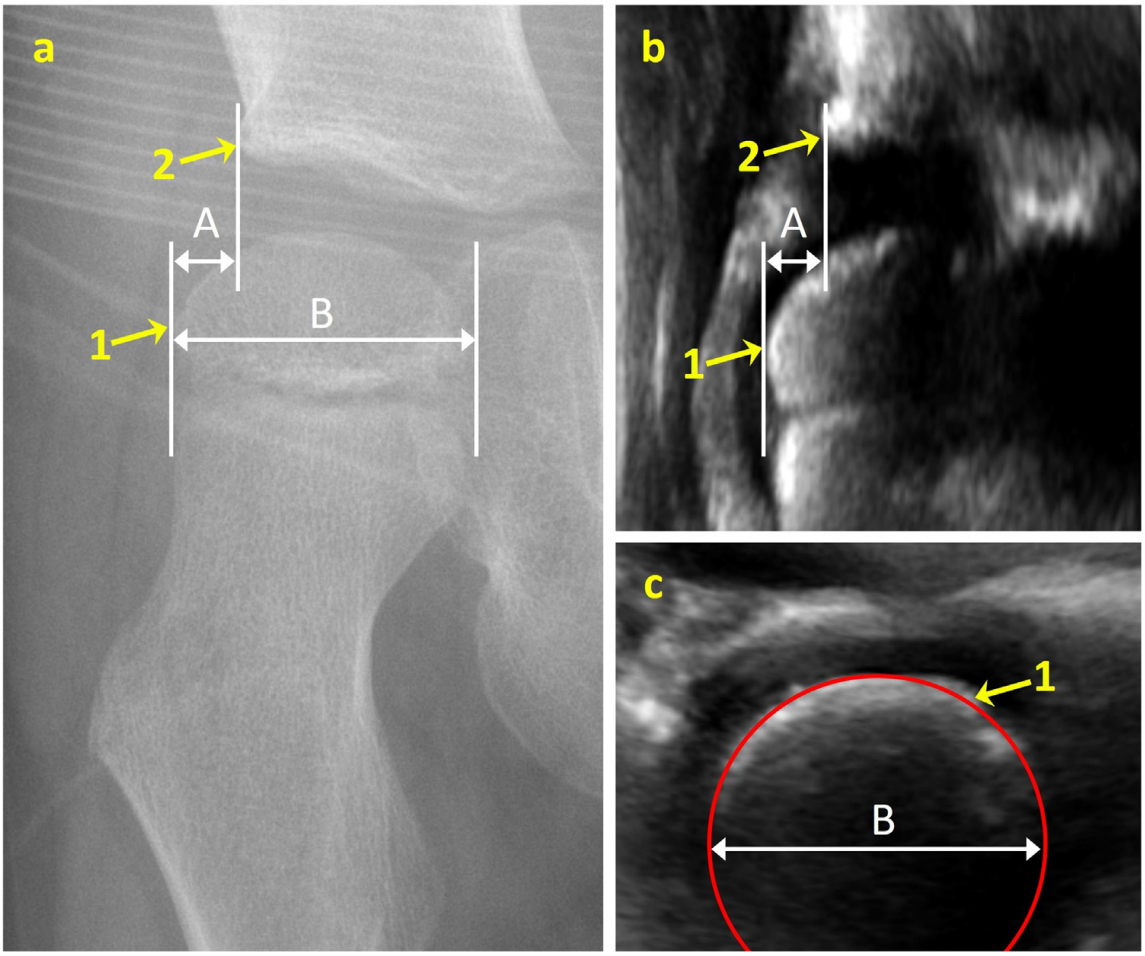


Figure 2.5: (a) Anteroposterior X-ray; US coronal (b) and transverse (c) scans; 1, femoral head; 2, lateral acetabular margin. [11].

2.5 Application of Machine Learning in Hip Examination Image Processing

Convolutional neural networks (CNNs) have been used to segment the acetabular roof and ilium in US images of hip infants. The segmented parts were then used to estimate the alpha angle in Graf's method for developmental dysplasia of the hip (DDH) diagnosis [46–48]. CNNs with residual blocks were employed for osteonecrosis of the femoral head diagnosis [49]. RNNs have been utilized to detect the frames with diagnostic features in 3D US volumes for DDH examination [50]. Segmentation networks have been used for anatomical landmark detection in hip radiographs, *e.g.*, Mask R-CNN [51], UNet [52, 53], Hourglass, and HRNet [53].

Automatic algorithms for MP measurements from radiographs have been developed as well. Liu *et al.* [54] converted the task of landmark detection into a segmentation problem and used a modified UNet to detect the required landmarks for MP measurements from hip radiographs. Pham *et al.* [55] used a two-staged pipeline of CNNs to find the landmarks in hip radiographs from children with CP. The first CNN gave an estimation of the coordinates of the landmarks, and the second network located the landmarks with more accuracy.

2.6 Chapter Summary

CP is a group of disorders that impacts an individual's moving and maintaining balance ability at different levels. Hip displacement is a common issue in children with CP. If not diagnosed and treated early, hip displacement can lead to hip dislocation and cause pain and disability in daily activities. Currently, the best method for measuring the severity of hip displacement is to measure the MP value on anteroposterior pelvis radiograph. However, taking radiographs expose these children to ionizing radiation, which is undesriable. US is a safer alternative for MP measurement. Machine learning has the potential to assist healthcare professionals in reliable and expeditious

measurements of MP from US images.

Chapter 3

Studies on Manual Measurements and Labeling on X-ray and ultrasound Images

In this chapter, we report three validation studies conducted on X-ray and ultrasound (US) images to assess the reliability and accuracy of manual measurements and labelings. The chapter is structured as follows:

Firstly, in Section 3.1, we present a study conducted by the author to evaluate the reliability and accuracy of migration percentage (MP) measurements on radiographs. In Section 3.2, a brief study on the validation of manual frame selection is provided. Finally, in Section 3.3, we present a concise study on manual landmark detection on coronal US frames.

3.1 Manual Radiographic Measurements Study

The author approached the task of hip displacement analysis without any previous experience. To acquire the necessary knowledge and develop a comprehensive understanding of the problem, a diligent study was taken on radiographs. As part of this study, the author personally conducted MP measurements on the radiographs, meticulously assessing the intra- and inter-reliability of these measurements.

3.1.1 MP Measurement Process on Radiographs

Figure 3.1 illustrates a sample hip radiograph along with the indated points, lines, and measured distances. The MP measurement process involved several steps. Firstly, the Hilgenreiner line was drawn, which connects the lower parts of both triradiate cartilages (A1, A2). Secondly, Perkin line was drawn, defined as the line perpendicular to the Hilgenreiner line and passing through the lateralmost point of the acetabulum (B1, B2). The next step involved measuring the width of the femur head. To accomplish this, the two marginal points of the femur head were identified (C1, D1 and C2, D2), and lines were then drawn from these marginal points, perpendicular to the Hilgenreiner line. The distance between the lines passing through (C1, D1) or (C2, D2) was measured as the total width of the femur head ("B"). Finally, the distance between the Perkin line and the closest lateral point of the acetabulum to the Perkin line (C1 or C2) was determined as "A." In cases where the femur head was entirely located below the acetabulum, the value of "A" was set to zero. Equations (3.1) to (3.7) provide the mathematical equations used to calculate MP based on the coordinates of the reference points.

Assuming the explicit equation of a line, it can be represented as:

$$y = m_0 \cdot x + h_0, \tag{3.1}$$

with m_0 being the slope of the line and h_0 denoting the y-intercept value. The slope of the line passing through the points A1 and A2 (Hilgenreiner line) can be calculated as:

$$m' = \frac{y_{A1} - y_{A2}}{x_{A1} - x_{A2}}. \tag{3.2}$$

Furthermore, the slope of the line perpendicular to the Hilgenreiner line can be determined as $m = -1/m'$, where m' represents the slope of the Hilgenreiner line. As a result, we can calculate the y-intercept of the lines perpendicular to Hilgenreiner line that pass through points B1, C1, and D1 as:

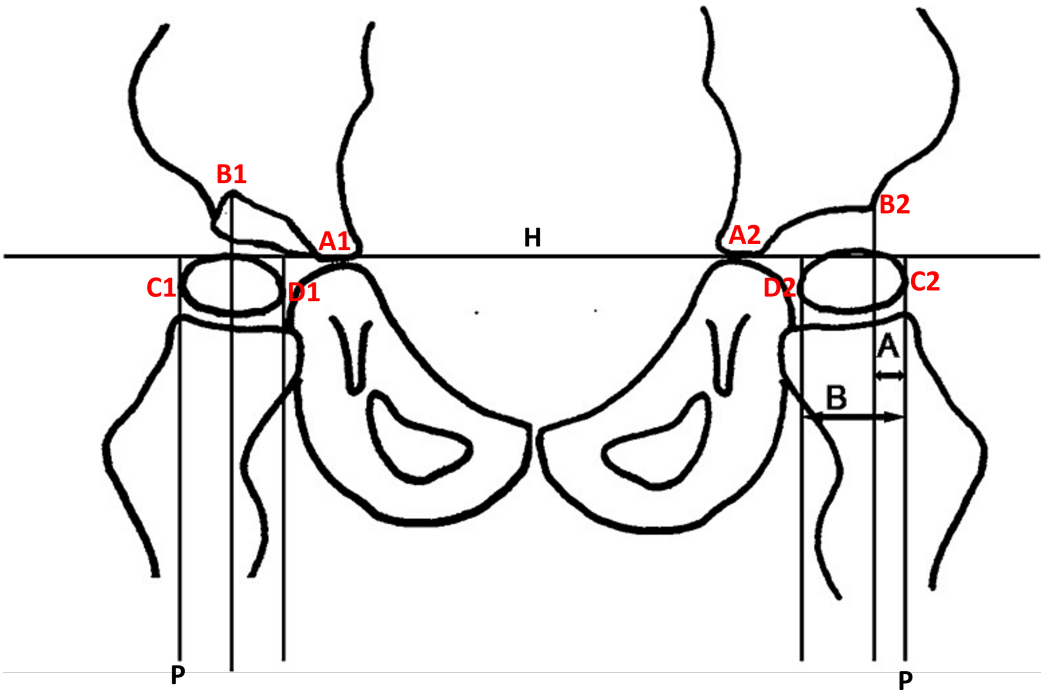


Figure 3.1: Sample hip radiograph illustrating the Hilgenreiner line (H), Perkin line (P), eight reference points for "A", "B", and MP measurements.

$$h_{B1} = y_{B1} - m \cdot x_{B1} \tag{3.3}$$

$$h_{C1} = y_{C1} - m \cdot x_{C1} \tag{3.4}$$

$$h_{D1} = y_{D1} - m \cdot x_{D1} \tag{3.5}$$

Therefore, the variables "A" and "B" correspond to the distances between parallel lines passing through B1 and C1, and the parallel lines passing through D1 and C1, respectively. These values can be computed as follows:

$$A = \max\left(sgn(m') \cdot \frac{h_{B1} - h_{C1}}{\sqrt{1+m^2}}, 0\right), \tag{3.6}$$

$$B = \frac{|h_{D1} - h_{C1}|}{\sqrt{1+m^2}}. \tag{3.7}$$

The variables "A" and "B" can be computed for the left hip by following a similar approach.

3.1.2 Materials and Methods

Anteroposterior pelvis radiographs from 50 children (aged 4-10 years) who had been diagnosed with cerebral palsy (GMFCS II-V) from a previous study[55] were used. The inclusion criteria were: (1) individuals with no prior history of pelvis or femoral surgeries, and (2) radiographic images exhibiting distinct visibility of both the femoral head and acetabular structures.

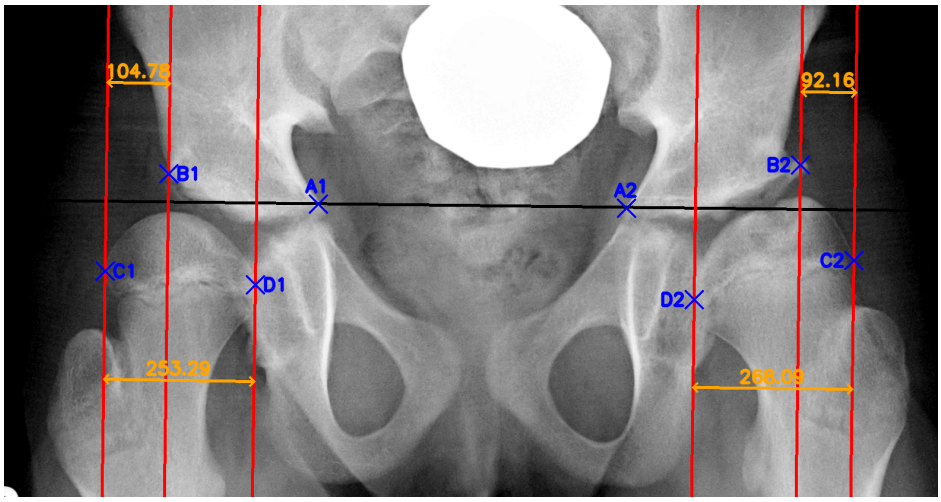
The MP measurements in this study were performed by the author, who will be referred to as R_1, on two separate occasions with a one-week interval between them to minimize memory bias. Prior to the measuring sessions, R_1 received explicit instructions regarding the measurement procedure during a training session with an experienced rater. Additionally, R_1 received further feedback after the initial measuring session to improve the accuracy and consistency of their measurements. To mitigate potential bias, R_1 measured the radiographs in a random order during each session. This approach prevented reliance on memorization or any predetermined sequence. Subsequently, the R_1's measurements were compared with those acquired by two experienced raters (R_2 and R_3) and a previously validated and published AI method [44].

The reliability of the R_1's MP measurements was assessed by calculating the mean absolute difference (MAD) $\pm$ standard deviation (SD) between the R_1's measurements and those of R_2, R_3, and AI. The intra- and inter-method intraclass correlation coefficient (ICC) of the MP measurements were calculated assuming a two-way mixed effects, absolute agreement, single rater/measurement model (ICC(2,1)). The ICC was assessed qualitatively according to the definitions provided by Koo [56]: poor (ICC $<$0.50), moderate (0.50-0.75), good (0.75-0.90), and excellent (ICC$>$0.90). Furthermore, the percentage of manual measurements falling within the clinically accepted range (MAD$<$10%) was determined, assuming the measurements from R_2, R_3, and AI as the reference values. All statistical analyses were performed using IBM

SPSS Statistics 27 software.

To streamline the MP measurement process, a semi-automatic application was developed by the author using tkinter library of Python. This application allows the user to input a set of images, which are then presented in random order. The user identifies the 8 required points in the specified order. After confirmation, the application automatically performs the entire measurement process based on Equations (3.2) to (3.7). Each radiograph is then annotated and saved for later review. Figure 3.2 shows a screenshot of the application alongside a sample annotated image.

(a)



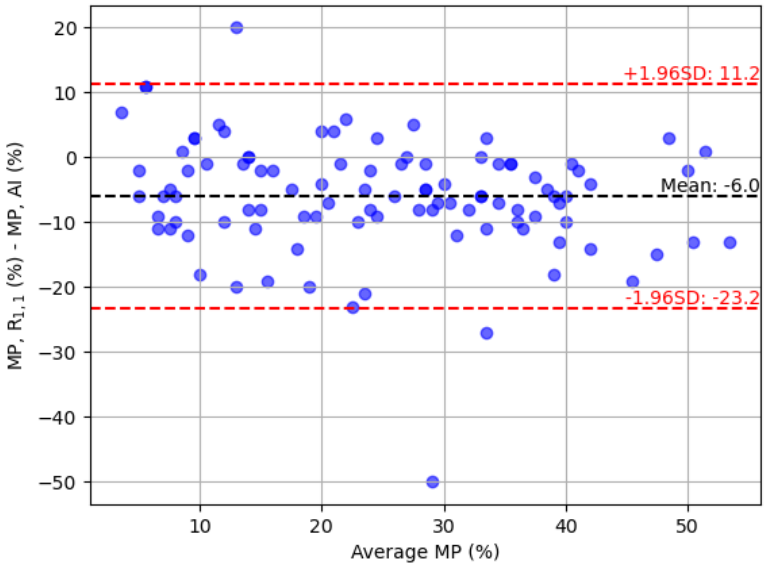
(b)

Figure 3.2: (a) A screenshot of the developed application for the semi-automatic measurement of MP on hip radiographs. (b) The generated annotated image for later review.

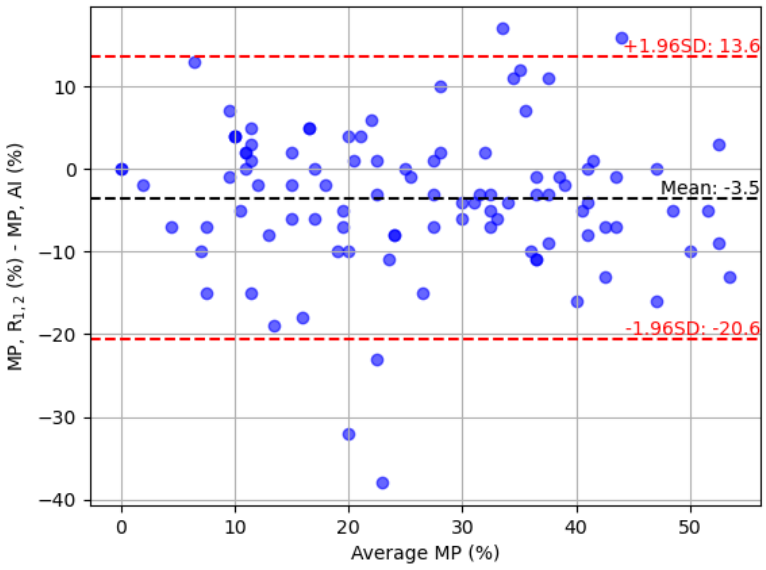
3.1.3 Intra and Inter-rater Reliabilities

The ICC(2,1) between the R_1's first and second measurements was 0.776, indicating a good level of reliability. However, the MAD ± SD between the two sessions was 5.32% ± 7.32%, higher than the reported intra-rater MAD ± SD of 3.2% ± 2.9% in [18]. This greater absolute difference likely was due to the R_1's inexperience, specifically errors made during the first session. These errors included incorrectly identifying the lateral point of the acetabulum and failing to report the correct measurement for some radiographs.

For assessing inter-rater reliability, measurements by the R_1 were compared to those of R_2, R_3, and an AI-based method. Table 3.1 summarizes the comparison results. The ICC(2,1) for the first set of measurements was lower than the second set, due to the feedback provided after the initial measurements. This feedback allowed the first rater to correct previously identified mistakes. The ICC(2,1) values for the second set of measurements indicated good agreement with R_2 (0.808), R_3 (0.798), and the AI (0.789). Furthermore, the average error decreased between the first and second measurements, demonstrating the positive impact of feedback and error correction on measurement accuracy. The Bland-Altman plot in Figure 3.3 compares the first and second measurements by R_1 against the AI measurements. The first measurements exhibited a bias of -6.04 with limits of agreement (LoA) of -23.28 to 11.20. The second measurements showed a reduced bias of -3.52 and narrower LoA (-20.56 to 13.52).



(a)



(b)

Figure 3.3: The Bland-Altman plot of the first MP measurements (a) and the second MP measurements on X-ray images, with bias (black line) and LoA (red lines).

Table 3.1: Summary of inter-rater reliability of MP measurements on radiographs.

		R_2	R_3	AI
$R_{1,1}$	ICC(2,1)	0.731	0.705	0.722
	MAD±SD	7.81% ± 7.36%	7.73% ± 7.98%	7.92% ± 7.15%
	CA percentage	70%	74%	69%
$R_{1,2}$	ICC(2,1)	0.808	0.798	0.789
	MAD±SD	6.39% ± 6.57%	6.07% ± 7.03%	6.84% ± 6.42%
	CA percentage	73%	78%	73%

$R_{1,j}$, j-the measurement of first rater; R_2,R_3, experienced raters; AI, measurements by AI-based method; CA percentage, percentage of the measurements within the clinical acceptance error.

3.1.4 Summary

This study provided R_1 with essential background knowledge on hip displacement and MP measurement using X-ray images. Throughout the study, R_1 received constructive feedback, enabling them to achieve high accuracy in MP measurements by its conclusion.

3.2 Manual Ultrasound Frame Selection Study

Each US scan typically comprises 100-1000 frames. However, only a small subset (25-100 frames) contains the necessary features for measurement. Accurate measurement depends heavily on identifying these frames appropriately. Our developed algorithm addresses this challenge by using convolutional neural networks (CNNs) for frame selection. These CNNs are trained using manually labeled frames. To ensure the reliability of the manual frame selections, we conducted a brief study comparing the author's labeling against those of two experienced raters (R_2 and R_3).

3.2.1 Materials and Methods

At the time of conducting the manual frame selection There were a total of 54 hip US scans from 15 children available, with 27 scans in the coronal view and 27 scans in

the transverse view. To determine the appropriate range of frames within each view, each rater independently identified the beginning and ending frames. This process was repeated by each rater, with a one-week interval between the measurement sessions. Prior to the second rating, constructive feedback was provided to the R_1 with the intention of achieving a unified definition of the measurement features. In order to assess the agreement between R_1 vs. R_2 and R_3, we utilized the ICC(2,1) to calculate the level of agreement for the beginning and ending frames in both the coronal and transverse view scans.

3.2.2 Results of Intra- and Inter- Rater Reliabilities

The results for coronal view frame selection are presented in Tables 3.2 and 3.3. For selection of the beginning frame, the average ICC(2,1) was 0.764, with 6 out of 8 comparisons exceeding 0.75. This indicates an overall good level of agreement with R_2 and R_3. The selection of ending frame showed even higher ICC values, averaging 0.835, with all comparisons exceeding 0.75. This suggests an excellent level of agreement with R_2 and R_3 in identifying ending frames. Additionally, the R_1's intra-rater reliability in coronal view frame selection was assessed. The ICC(2,1) between the first and second measurements was determined to be 0.881 for selection of beginning frames and 0.890 for selection of ending frames, indicating good reliability and consistency in frame selection by the first rater.

Table 3.2: Summary of manual beginning frame selection in coronal view.

	$R_{2,1}$	$R_{2,2}$	$R_{3,1}$	$R_{3,2}$
$R_{1,1}$	0.787	0.756	0.721	0.764
$R_{1,2}$	0.79	0.817	0.726	0.751

$R_{i,j}$, j-th measurement of rater i; R_1, The first rater; R_2,R_3, The experienced raters; Average ICC: 0.764.

Table 3.3: Summary of manual ending frame selection in coronal view.

	$R_{2,1}$	$R_{2,2}$	$R_{3,1}$	$R_{3,2}$
$R_{1,1}$	0.842	0.823	0.806	0.844
$R_{1,2}$	0.85	0.884	0.827	0.811

$R_{i,j}$, j-th measurement of rater i; R_1, The first rater; R_2,R_3, The experienced raters; Average ICC: 0.835.

The results for the transverse view frame selection are presented in Tables 3.4 and 3.5. The R_1 demonstrated excellent intra-rater reliability in transverse view frame selection. ICC(3,1) values were 0.975 for beginning frames and 0.979 for ending frames, indicating highly consistent frame selection. For the inter-rater reliability in selection of beginning frames, the average ICC was calculated to be 0.888. Among the 8 comparisons with R_2 and R_3, half showed good agreement (ICC >0.75) and half showed excellent agreement (ICC >0.9) in selecting beginning frames for the transverse US scans. The average inter-rater ICC for selection of ending frames comparisons was 0.903, indicating strong agreement with R_2 and R_3. Similarly, half of the comparisons demonstrated good agreement (0.75 <ICC <0.9) and half demonstrated excellent agreement (ICC >0.9).

Table 3.4: Summary of manual beginning frame selection in transverse view.

	$R_{2,1}$	$R_{2,2}$	$R_{3,1}$	$R_{3,2}$
$R_{1,1}$	0.852	0.974	0.786	0.969
$R_{1,2}$	0.85	0.934	0.785	0.958

$R_{i,j}$, j-th measurement of rater i; R_1, The first rater; R_2,R_3, The experienced raters; Average ICC: 0.888.

Table 3.5: Summary of manual ending frame selection in transverse view.

	$R_{2,1}$	$R_{2,2}$	$R_{3,1}$	$R_{3,2}$
$R_{1,1}$	0.887	0.965	0.780	0.950
$R_{1,2}$	0.897	0.966	0.818	0.964

$R_{i,j}$, j-th measurement of rater i; R_1, The first rater; R_2,R_3, The experienced raters; Average ICC: 0.903.

3.2.3 Summary

This study established R_1's capability in consistent and accurate US frame selection across both transverse and coronal views, as evidenced by high intra- and inter-rater reliability scores. Notably, ICC values were higher for transverse frame selection, likely due to the clearer definition of anatomical features within those views.

3.3 US Landmark Detection Study

The accurate detection of acetabulum and femur head landmarks on coronal frames is crucial for measuring variable "A" and, thereby, MP. To train corresponding neural networks (which will be explained in Chapter 4), these landmarks were manually labeled. We conducted a comparative study to ensure the reliability of these labels, comparing the R_1's labeling against that of an experienced rater (R_3). This study aimed to validate the consistency and accuracy of the R_1's labeling, ensuring a reliable foundation for the landmark detection process.

3.3.1 Materials and Methods

For the landmark detection study, 10 hip scans were randomly selected from a pool of 27 coronal US scans (taken from 15 children). Five frames from each scan were randomly chosen, ensuring they were among the frames commonly selected by R_1 and R_3 in the previous study (see Section 3.2). To assess intra-rater reliability, the R_1 labeled these frames twice, with a one-month interval between labeling sessions.

Feedback was provided between the measurement sessions to eliminate unclarities in landmark definitions. For inter-rater reliability, the same frames were independently labeled by R_3, and their results were compared with those of the first rater.

To facilitate the labeling process, a Python application was developed. This application allowed the user to select the desired frames and loads them for labeling. The frames were displayed to the user in a random order, without revealing the filenames to prevent memorization. The user could click on the frames to determine the landmarks, zoom in by dragging the mouse on a specific area, and clear any selected dots if necessary. Once the landmarks were specified, the application generated a CSV file containing the coordinates of the landmarks and their corresponding vertical distances, which corresponded to variable "A". After the specified points were confirmed, each frame was automatically labeled and saved for future review. Figure 3.4 depicts a flowchart of the developed application, and Figure 3.5 displays a screenshot of the application developed for US frame labeling and the resulting output frame.

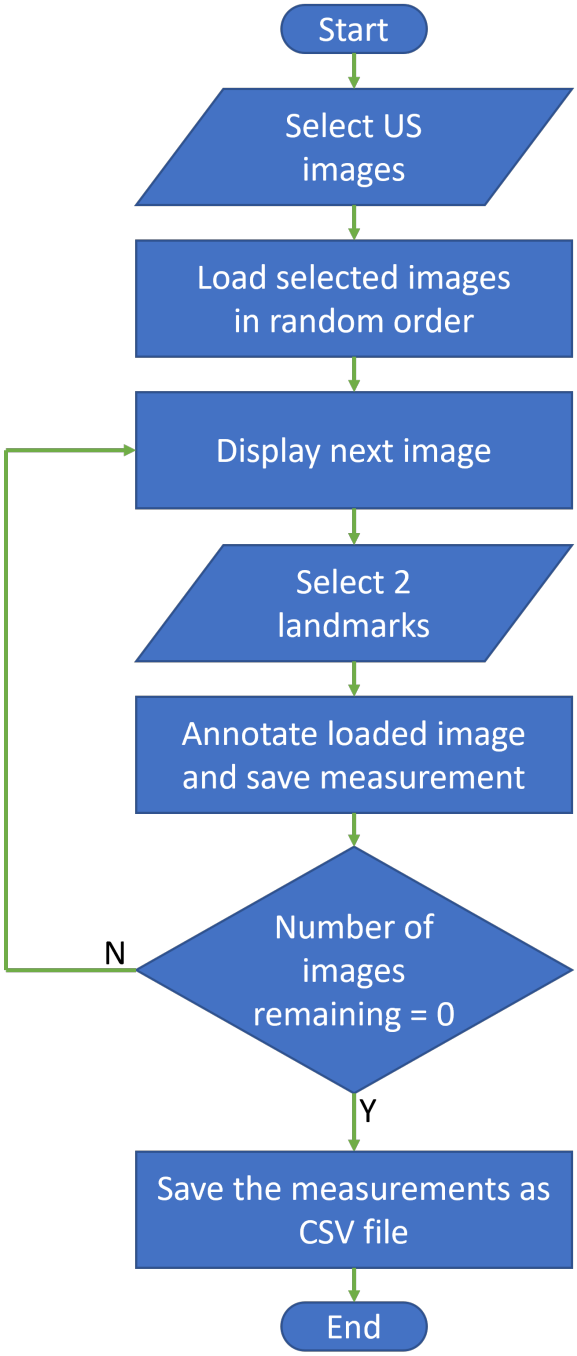
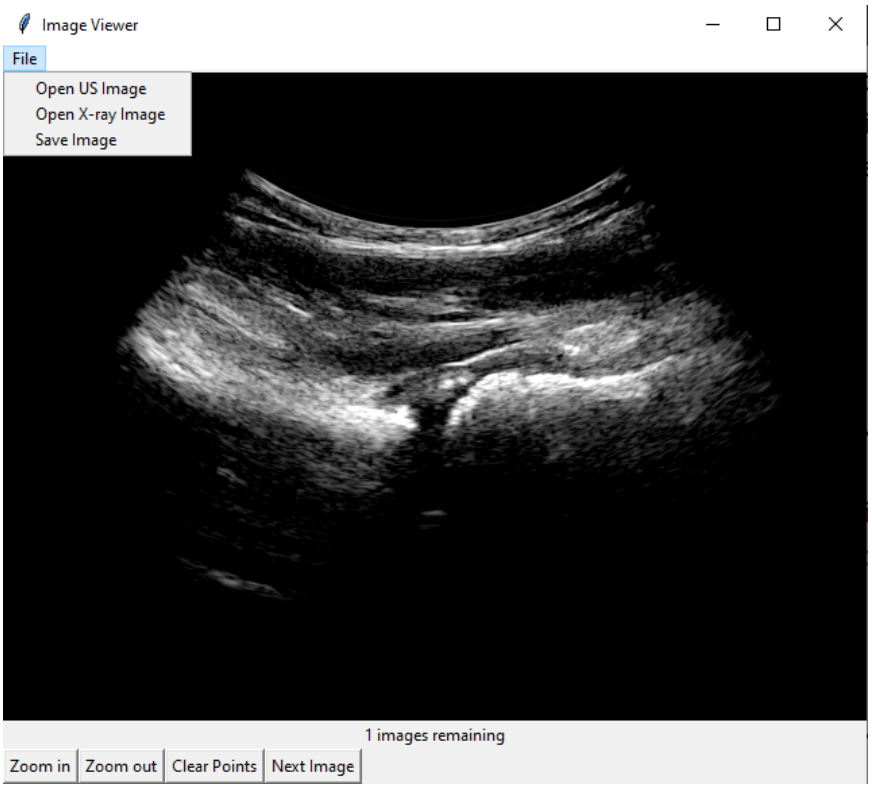
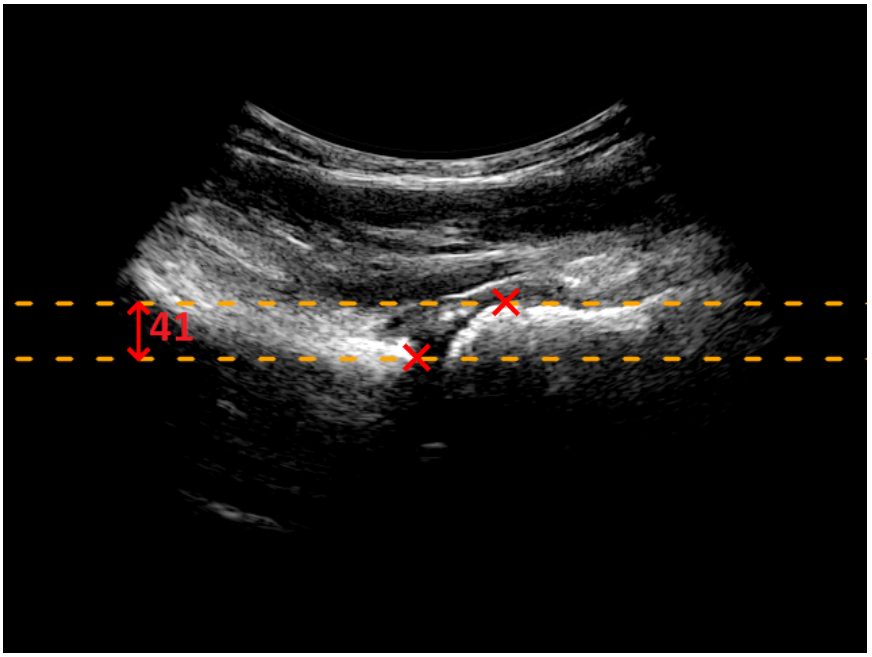


Figure 3.4: Flowchart of the developed semi-automatic application for "A" measurement.



(a)



(b)

Figure 3.5: A screenshot (a) of the dedicated application designed for evaluating the reliability of "A" measurements on hip scans. (b) The generated annotated image for later review.

3.3.2 Results of Accuracy and Intra- and Inter- Rater Reliabilities

Table 3.6 summarizes the pair-wise comparisons between the R_1's first and second measurements and the R_3's measurements. The average difference between the first and second R_1 measurements was -0.2mm $\pm$ 0.8mm, with LoA of (-1.8mm, 1.4mm), deviating from the previously reported LoA of $\pm$1.0mm in [39]. This discrepancy likely stemmed from the R_1's initial inexperience and some ambiguity in landmark definition during the first measurement session. Additionally, the greater difference value can be attributed to our study's higher average LHD (6.9mm vs. 3.7mm in [39])), a consequence of higher average participant age. Despite these factors, the excellent intra-rater ICC value demonstrated the R_1's overall reliability. The Bland-Altman analysis of the first and second measurements by R_1 is presented in Figure 3.6, displaying the bias, LoA, and the measurements falling in the LoA.

Overall, the mutual ICCs between R_1 and R_3 were above 0.9 in both sessions, indicating excellent agreement between R_1 and R_3 measurements, and higher than the reported ICC of 0.84 for inter-rater agreement in literature [39]. The average difference between the measurements of R_1 and R_3 decreased in the second measurement comparing to the first measurements. In addition, the LoA in the second measurements narrowed down. All demonstrating the improvement in R_1's measurement accuracy from the first measurement session to second measurement session. Figure 3.7 presents Bland-Altman plots comparing the first rater's first and second measurements to those of R_3.

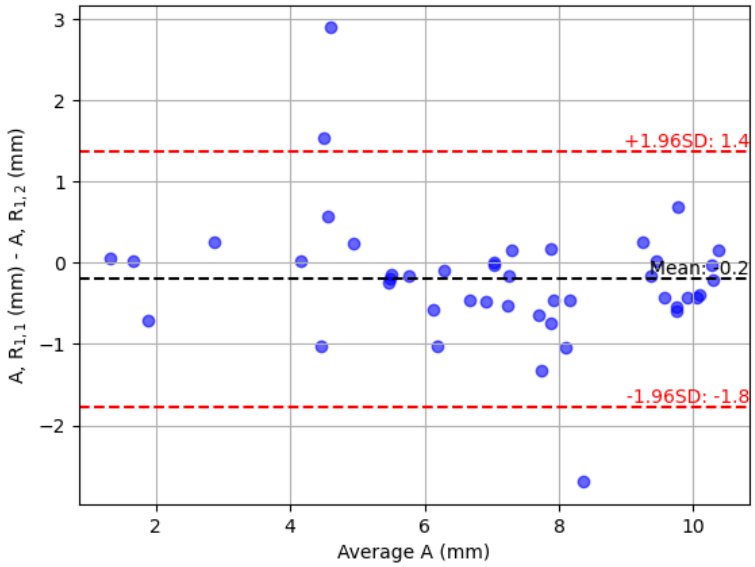
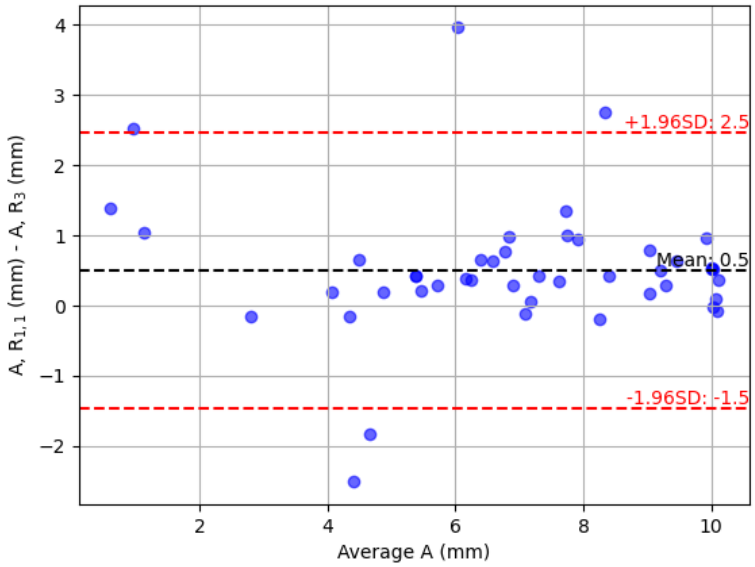


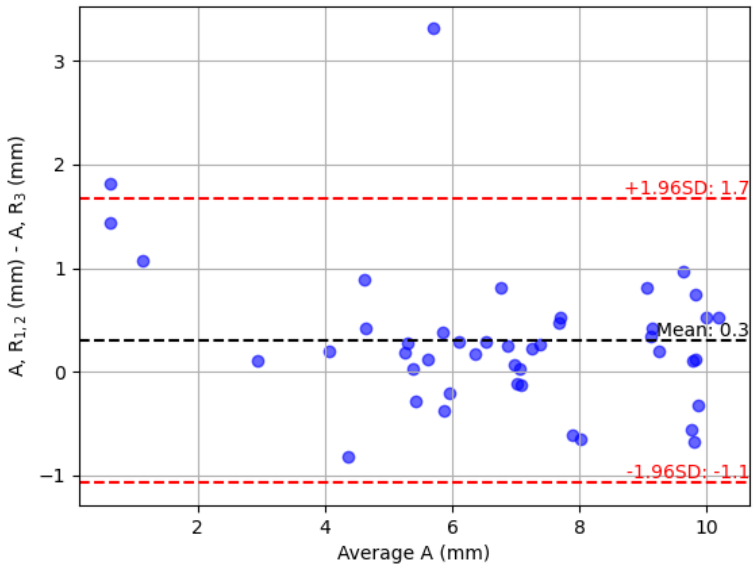
Figure 3.6: Bland-Altman plot of the first and second LHD measurements on US images by the author, including average difference and LoA.

Table 3.6: Summary of the intra- and inter- reliability of the LHD measurements by the first rater (author).

	ICC(3,1)	Average Difference (mm)	LoA (mm)
$R_{1,1}$ vs $R_{1,2}$	0.949	-0.2 ± 0.8	(-1.8, 1.4)
$R_{1,1}$ vs $R_{3,1}$	0.913	0.5 ± 1.0	(-1.5, 2.5)
$R_{1,2}$ vs $R_{3,1}$	0.949	0.3 ± 0.7	(-1.1, 1.7)



(a)



(b)

Figure 3.7: Bland-Altman plots of the author's first (a) and second (b) measurements vs. the experienced rater (R_3), including average differences and LoA.

3.3.3 Summary

In the course of this study, the author focused on learning to accurately identify landmarks on coronal frames. Upon the study's completion, the author's ability to identify landmarks was deemed both accurate and reliable.

3.4 Chapter Summary

To validate key aspects of our measurement process, we conducted several studies. First, a study on X-ray images demonstrated excellent intra- and inter-rater reliability for the author's MP measurements. We then developed an application to streamline this process, significantly reducing manual measurement time. This study established the author's understanding of hip displacement and MP measurement, as well as their ability to measure MP from radiographs.

Additionally, we validated the author's manual US frame selection, a crucial step for our proposed CNN-based algorithm. This study showed good to excellent agreement between the author's selections and those of two experienced raters, confirming their ability to correctly select and label US frames according to the selection criteria. Finally, we authenticated the author's acetabulum and femur head landmark labels, finding excellent intra-rater reliability and agreement with an experienced rater. This validates the author's ability to accurately identify landmarks on coronal frames. These labelings are used later for model training.

Chapter 4

Algorithms for Automated Measurements

In this chapter, we provide a comprehensive explanation of the development and validation procedure of the developed method for migration percentage (MP) measurement from ultrasound (US) images. The chapter structure will unfold as follows:

Firstly, in Section 4.1, we present an overview of the training and testing workflow, outlining the key steps involved in the model's development and evaluation. In Section 4.2, we describe the procedure of data acquisition for this study. Subsequently in Sections 4.3 and 4.4 the preprocessing steps and overall architecture models that are used are presented. In Section 4.5, the details of dataset preparation and training procedure for measurement of both variables "A" and "B" are presented. Sections 4.6 and 4.7 describe the algorithm for using the trained models and making new measurements. we discuss how the model is validated and tested in Section 4.8. Finally, a summary of the chapter is given in Section 4.9. Portions of this chapter were submitted in the paper: Yousefvand, R., Pham, T-T., Le, L.H., Andersen, J., Lou, E. H, "Applying Deep Learning for Automatic Measurement of Migration Percentage from Ultrasound Images in Children with Cerebral Palsy," Medical & Biological Engineering & Computing, 2024, (Submitted).

4.1 Overview of the Developed Method

Figure 4.1 illustrates the general data flow in this study. As it can be seen, initially all training, validation and test data are preprocessed. The train and validation sets are used to develop and optimize the networks. Some of the optimized networks are used in the training procedure of the other networks. Finally, the test dataset is used to evaluate the trained model.

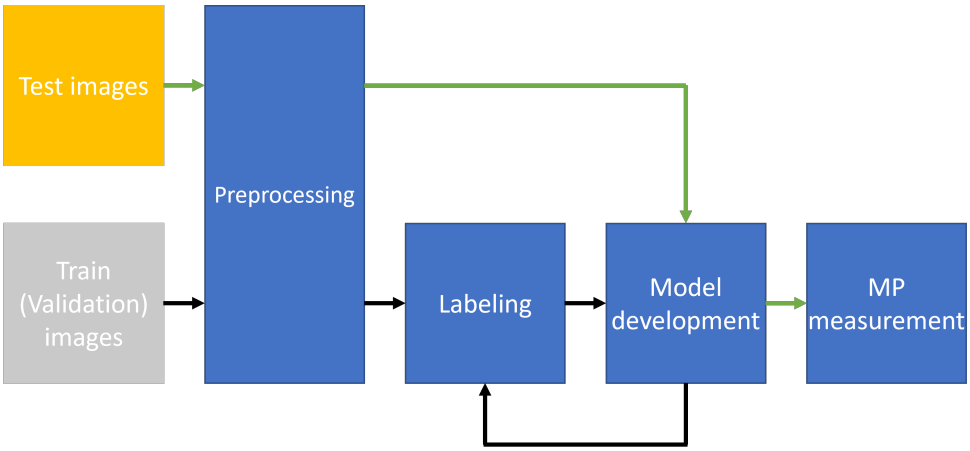
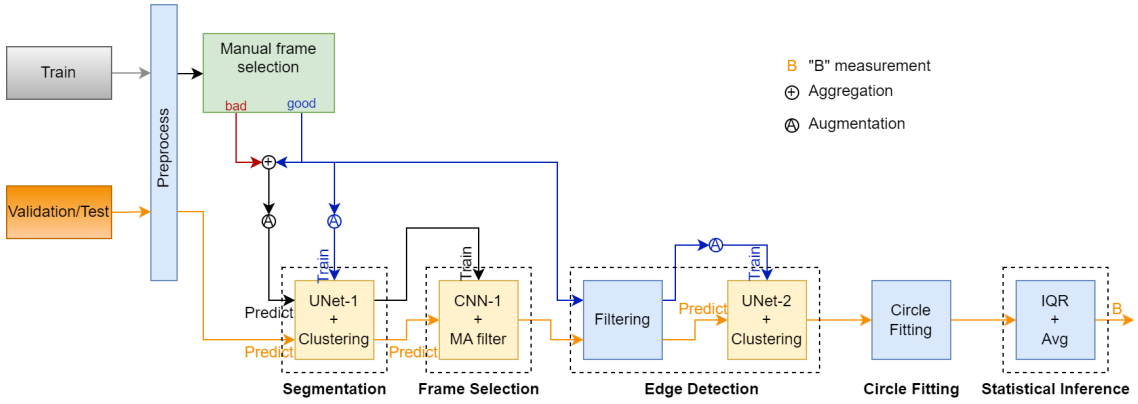


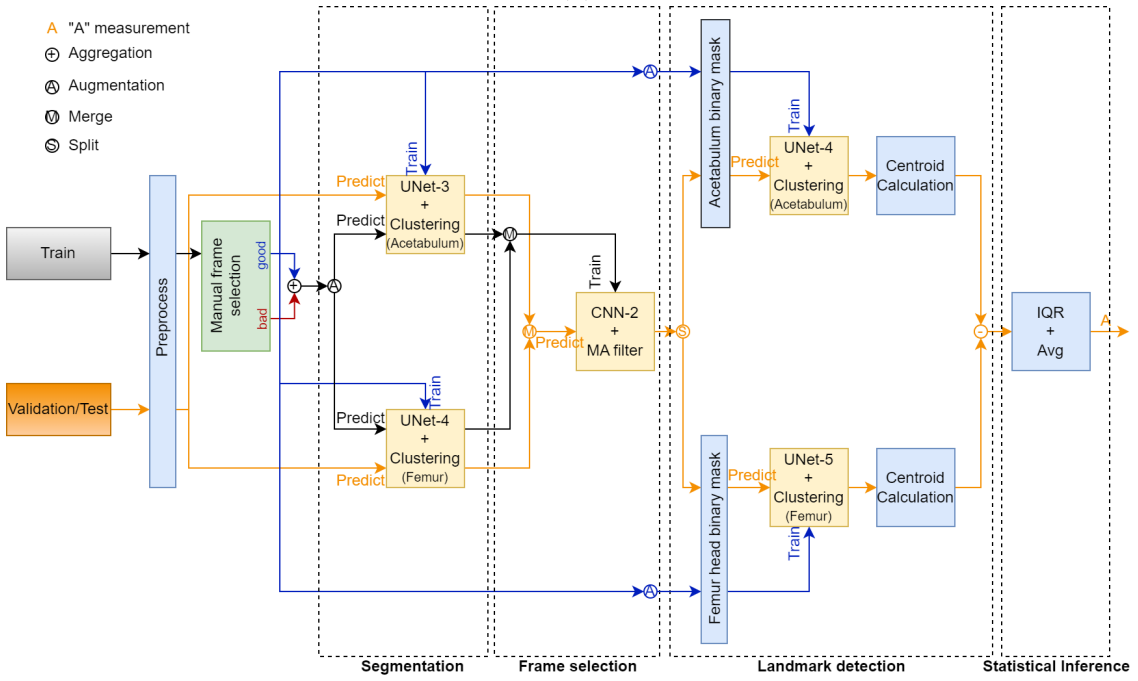
Figure 4.1: General data flow in model development and testing.

The model training process for variable "B" measurement consists of five stages: a) image preprocessing, b) manual frame selection and labeling, c) segmentation model training, d) automatic frame selection, and e) edge detection model development. This process closely mirrors the steps for variable "A", with the key difference being the landmark detection model development in step e.

Validation/testing follows a similar process: a) image preprocessing, b) segmentation, c) automatic frame selection, and d) either edge detection (for "B") or landmark detection (for "A"). Finally, e) statistical inference is performed. Figure 4.2 (referencing the figure) illustrates these procedures in detail. Note: orange arrows represent test/validation data, while blue, red, and black arrows indicate "good", "bad", and "aggregated" frames within the training data.



(a)



(b)

Figure 4.2: Diagram illustrating data flow for training and testing/validation of variables “B” (a) and “A” (b). Blue, red, and black arrows indicate good, bad, and mixed frames during training; green arrows depict test/validation flow.

4.2 Data Acquisition

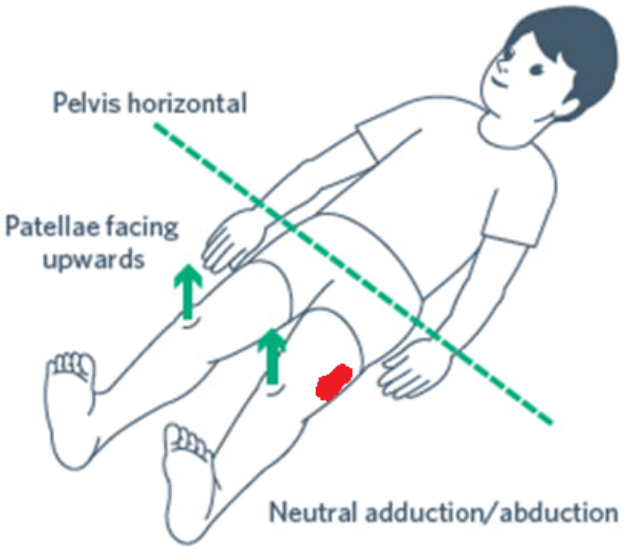
A total of 38 children (26 male, 12 female) with an average age of 9 years were recruited from a local rehabilitation hospital. The inclusion criteria for participation in the study were as follows:

- Diagnosed with cerebral palsy (CP)
- Age between 4 and 16 years
- No history of intervention surgeries
- Participation in a surveillance program
- An anteroposterior radiography image taken within two months

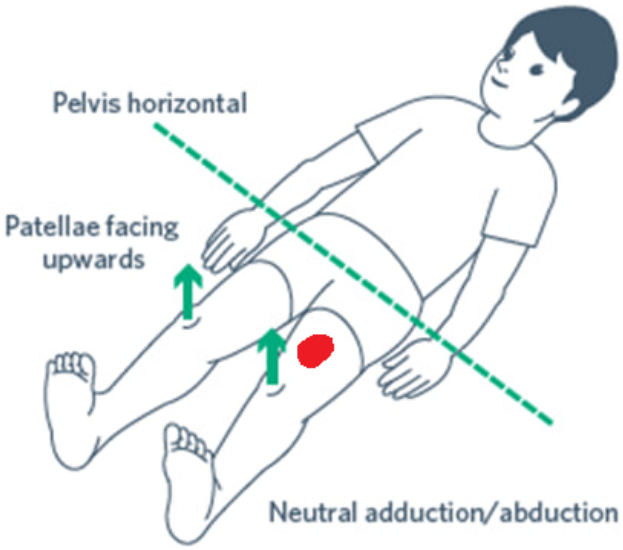
Permission from the institutional research ethics board was obtained prior to conducting the study. Written parental consent and participant assent was obtained from both the participants and their parents before the scanning procedure. For the US scanning, a hand-held Clarius C3 convex scanner (Figure 4.3) with an operating frequency of 2 to 6 MHz was used. The scanning parameters included the musculoskeletal mode, 4 MHz frequency, and 6 cm imaging depth. During the scanning process, the patients were positioned in a supine position on a bed, and their lower limbs were adjusted to be as parallel as possible. Prior to the scanning, ultrasonic gel was applied to the scanning regions of the hips to prepare for the US. The ultrasonic gel was already warmed to prevent startling reactions in children caused by a cold substance, making the scanning procedure smoother. The coronal view scans were obtained by scanning the lateral side with the transducer along the superior-inferior axis. For transverse view scans, the frontal view of the hip was scanned with the transducer positioned horizontally to the superior-inferior axis. Figure 4.4 depicts the patient positioning for both coronal and transverse scanning. A total of 104 scans

were obtained, with one coronal and one transverse scan acquired for each hip. Figure 4.5 shows the whole setup for scanning in a clinical room, including the Clarius scanner, the bed, the pillow, the laptop used to display the scanned frames and save the data.

Figure 4.3: A picture of the hand-held Clarius C3 convex scanner used for obtaining hip sonograms in this study.



(a)



(b)

Figure 4.4: Correct positioning for US hip scanning. The red color shows the scanning area for coronal (a) and transverse (b) view scans.

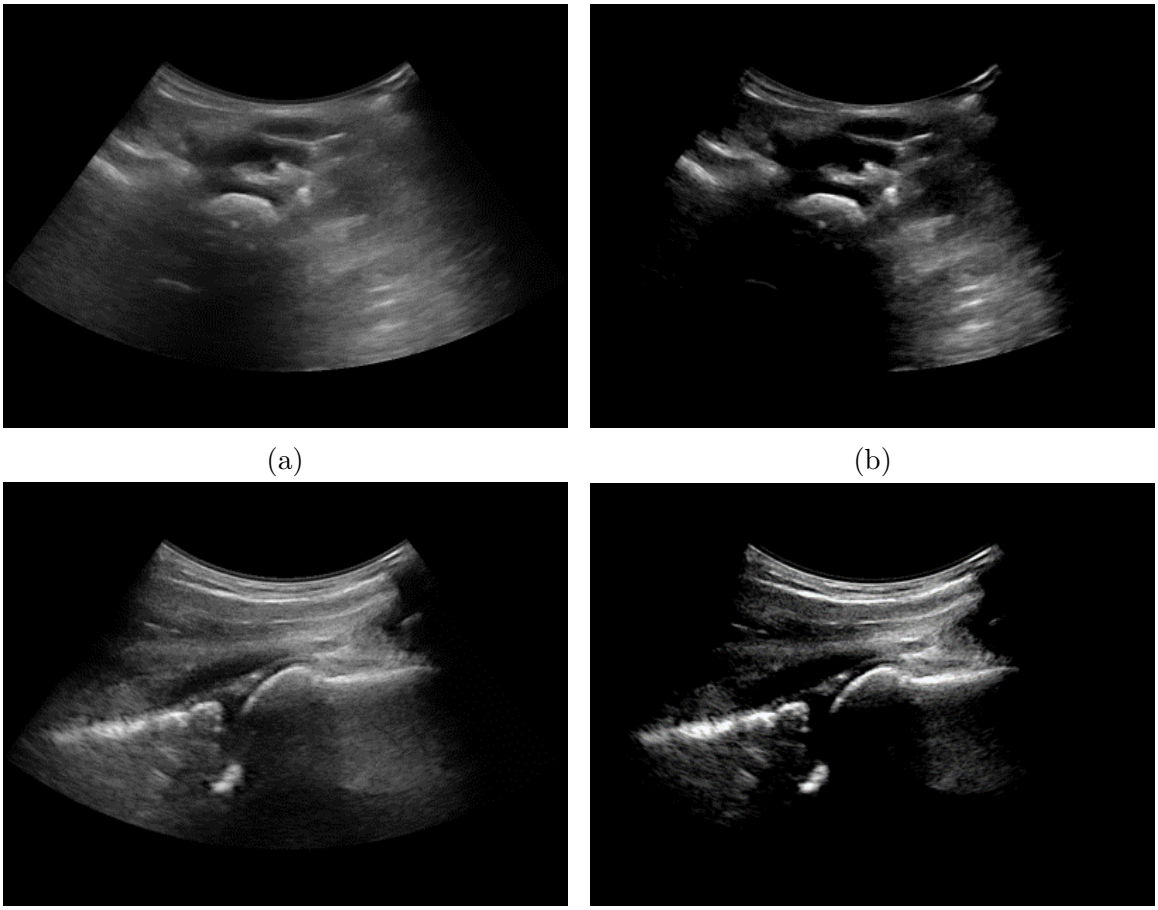
Figure 4.5: Hip scanning setup with the bed for supine positioning, pillow for support, US scanner, laptop for data collection, and tissues for post-scan cleanup.

4.3 Preprocessing

As outlined in Section 4.1, our method for measuring "A" and "B" begins with a preprocessing step. This step is essential to address the inherent noise and fuzziness of US images, which can hinder image quality. We employ the windowing technique to reduce noise and enhance contrast. Windowing works by restricting pixel values to a specific range and rescaling them to optimize the image's dynamic range. The following equation demonstrates how windowing modifies pixel values:

$$p' = \begin{cases} 0 & \text{p<a,} \\ \frac{p-a}{b-a} * m & \text{a<p<b,} \\ m & \text{p>b,} \end{cases} \tag{4.1}$$

where the dynamic range of the image is denoted as [0, m], with a and b being arbitrary pixel values within that range. The original and modified pixel values are represented by p and p', respectively. Figure 4.6 provides a visual representation of preprocessing effects on a transverse US frame, illustrating the image prior to and



(a) (b)

(c) (d)

Figure 4.6: Transverse US frame: (a) original, (b) after windowing preprocessing. Coronal US frame: (c) original, (d) after windowing preprocessing

following the application of windowing.

4.4 Neural Network Architecture and Training

4.4.1 CNN Architecture and Training

Convolutional neural networks (CNNs) are multilayered neural networks based on convolution and are designed to extract patterns and features from images. CNNs have revolutionized fields like image classification, object detection, and etc. In this study, we use four widely used CNNs: ResNet152, ResNet50, VGG19, and DenseNet121 for frame classification, and as part of the UNet architecture which will be explained in Section 4.4.2. These architectures were selected due to their

proven performance in image-related tasks. To leverage transfer-learning techniques for training CNNs, pre-trained CNNs were utilized. The original output layer of each CNN was replaced with a single-output layer to enable binary classification. During the training, the weights of all layers except the last layer were kept frozen. Binary cross-entropy (BCE) loss function was used for training.:

$$L = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p(y_i)) + (1 - y_i)\log(1 - p(y_i)),\ y_i \in 0, 1, \tag{4.2}$$

Where $y(i)$ is the label of i-th sample and $p(y_i)$ is the predicted probability for that label.

4.4.2 UNet Architecture

UNet is a deep learning architecture commonly used for image segmentation tasks. It consists of an encoder-decoder structure, where the encoder captures contextual information from the input image and the decoder generates a segmented output. Skip connections are utilized to combine features from different resolutions, enabling precise localization of objects in the image [57]. To leverage transfer learning, we used a pre-trained CNN as the encoder component in the UNet architecture. The decoder part of the UNet was constructed by appending four upsampling layers to the outputs of four CNN layers and incorporating skip connections. The overall UNet architecture is illustrated in Figure 4.7. During training, the weights of the CNN were kept frozen to retain the knowledge learned from its previous task. The network was trained using the sparse categorical cross-entropy (SCCE) loss function, which quantifies the dissimilarity between the predicted and the actual probability distribution of the target variable. The mathematical equation for the SCCE loss function is presented in Equation (4.3).

$$L = \frac{1}{N}\sum_{i=1}^{N} -\log(p(y_i)), \tag{4.3}$$

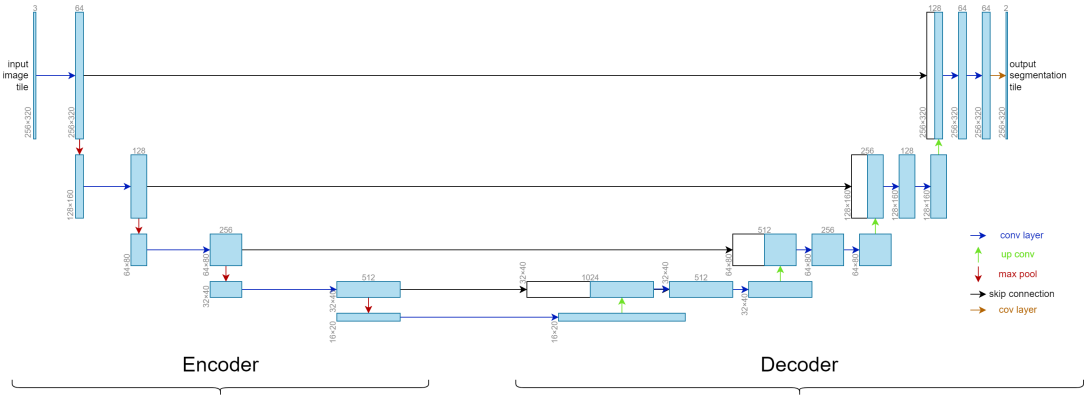


Figure 4.7: A sample UNet architecture with the encoder and decoder blocks.

where $y(i)$ is the label of i-th sample and $p(y_i)$ is the predicted probability for that label. In this study, we used UNets for segmentation, edge detection, and landmark detection which will be explained further.

4.4.3 Residual Blocks

Training very deep CNNs faces a challenge called vanishing gradients, that is, the gradients, which are used to update the weights of the network during backpropagation, become very small as they propagate backward from the deeper layers to the earlier layers. The residual block, as depicted in Figure 4.8, introduced skip connections that allowed the network to learn residual mappings. By adding the original input to the output of the block, the gradient can flow directly from the output to the input, bypassing intermediate layers. This helped alleviate the vanishing gradient problem by providing a direct path for gradients to reach earlier layers, enabling effective training of deep CNNs [58].

ResNet architectures, such as ResNet-50 and ResNet-152, were groundbreaking in their use of residual blocks to enable the training of significantly deeper networks [59]. Their success in image recognition tasks demonstrated the power of this approach in overcoming the challenges of training very deep CNNs.

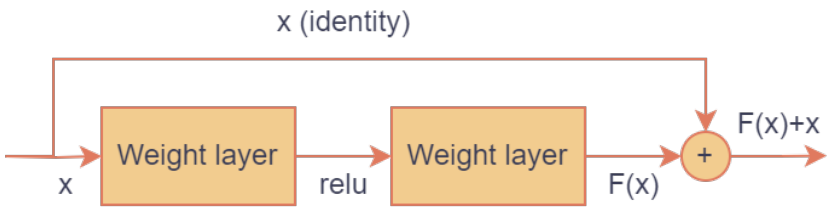


Figure 4.8: Residual block in the ResNet architecture. The skip connection adds the output from layer l to layer l+2, enabling the gradient flow and improving training efficiency.

4.4.4 VGG19

VGG19 is a highly influential CNN architecture known for its simplicity and striking performance. VGG19 features a 19-layer architecture consisting primarily of convolutional layers with small 3x3 filters. This design emphasizes depth, enabling the extraction of increasingly complex features from input images. The network's key contribution lies in demonstrating that a substantial increase in depth leads to improved performance in large-scale image recognition tasks [60]. An overal view of VGG19 architecture is displayed in Figure 4.9.

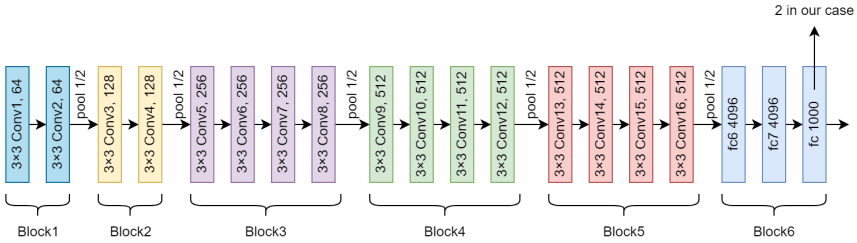


Figure 4.9: VGG19 overall architecture.

4.4.5 DensNet121

Densely Connected Convolutional Networks (DenseNets) are a compelling development in deep learning architectures for image classification tasks. At their core, DenseNets address the vanishing gradient problem that can hinder the training of very deep networks. They achieve this by establishing direct connections between all pairs of layers with the same feature map size. A sample block of DenseNet architec-

ture is displayed at Figure 4.10. DenseNet121 is a specific and popular configuration of the DenseNet architecture, offering a balance of accuracy and computational efficiency. Its innovative structure has demonstrated remarkable performance across various image recognition benchmarks [61].

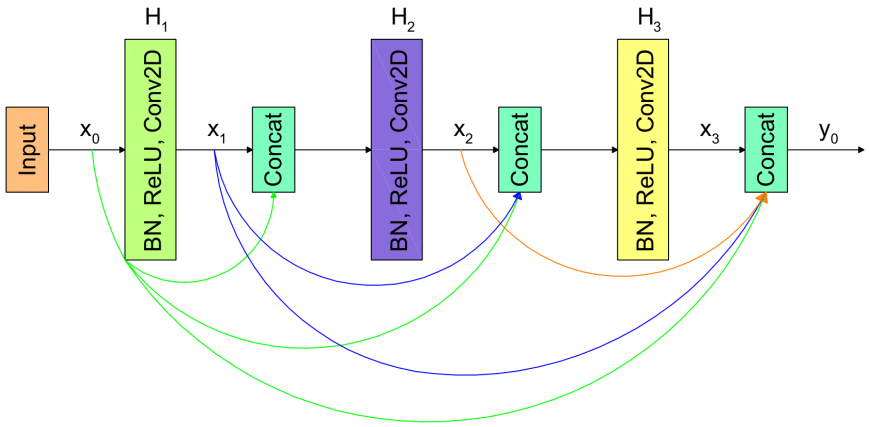
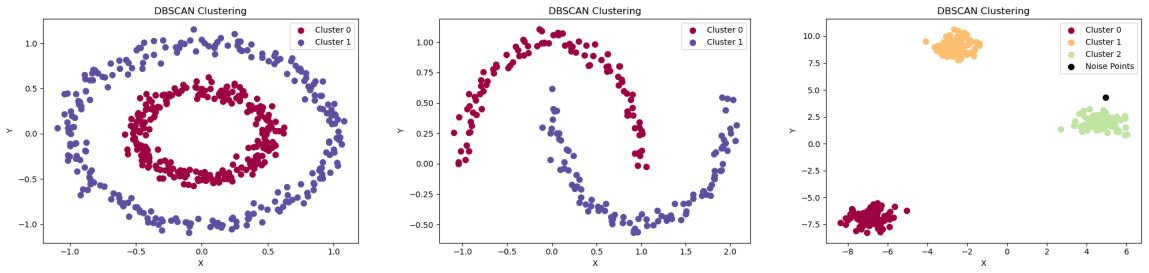


Figure 4.10: A 3-layer DenseNet block, showing how all pairs of layers with the same feature map size are connected. Each layer consists of a Batch Normalization (BN) layer, a ReLU activation layer, and a Convolution2D (Conv2D) layer. The outputs of all previous layers are concatenated before feeding into the next layer.

4.4.6 DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm distinguished by its capacity to identify clusters of arbitrary shapes while relying solely on data point density. DBSCAN requires two parameters: epsilon (ϵ), defining the radius of a point's neighborhood, and minimum samples (m), specifying the threshold number of samples needed to constitute a cluster. Three data point classifications exist within DBSCAN: core points, border points, and noise points. A core point possesses at least m other points within its ϵ-neighborhood. A border point lies within a core point's ϵ-neighborhood but lacks sufficient neighbors to be classified as a core point. Noise points fall outside the influence of any core points. DBSCAN initiates by randomly selecting an unvisited data point. If this point fulfills



(a) (b) (c)

Figure 4.11: Visual examples of DBSCAN performance on different cluster configurations.

the criteria of a core point, a cluster is formed, and neighboring data points are incorporated. Newly identified core points trigger further cluster expansion; this process iterates until the cluster reaches its maximum extent. Upon completion, data points unassigned to a cluster are designated as noise points. Figure 4.11 depicts the application of DBSCAN clustering to randomly generated data exhibiting diverse cluster shapes [62].

We utilized DBSCAN clustering to refine the outputs of UNets and mitigate potential noise. DBSCAN was applied to cluster the labeled pixels, facilitating noise removal. The resulting clusters were then sorted by total area, and only the largest cluster was retained. This means that pixel values in the smaller clusters were set to '0'.

4.5 Datasets Development

To calculate MP, Equation (4.4) is used. We will explain the model development steps for each parameter in the following.

$$MP = \frac{A}{B} * 100 \tag{4.4}$$

4.5.1 Augmentation

Training deep networks require a substantial amount of data, which may not always be feasible to obtain in certain scenarios. To address this problem, data augmentation is used to increase the size and diversity of the training dataset through different image transformations. Applying data augmentation helps to increase the generalizability and robustness of the model and avoid overfitting problems [63].

Given the limited availability of data in this study, we used image augmentation to increase the size of the training datasets. The image transformations used for augmentation are affine transformations including scaling, translation, and rotation, as well as cropping, noise injection, multiplying all pixel values by a random number, and elastic transformations. In subsequent sections, where augmentation is discussed, it should be noted that the aforementioned transformations were applied to the original datasets in a randomized order.

4.5.2 B-Measurement Datasets Development

Figure 4.12 depicts the general steps for development of B-measurement datasets. Among the 62 transverse US hip scans, 36 were designated for training and 8 for validation. Each scan comprised 100-1000 frames, yet only 25-200 of these frames displayed characteristics suitable for B-measurement. For each scan within the training and validation sets, a sequence of frames showcasing a visible femur head was identified as the target range. From each specified range, a maximum of 25 frames were selected and categorized as "good" frames, resulting in a total of 872 "good" frames. Excluding these, along with an additional buffer of 10% of the total frame number on both ends of the target range, 25 evenly distributed frames were chosen from the remaining frames and labeled as "bad" frames, resulting in a total of 900 "bad" frames. Figure 4.13 illustrates the selection process of "good" and "bad" frames from a hip scan.

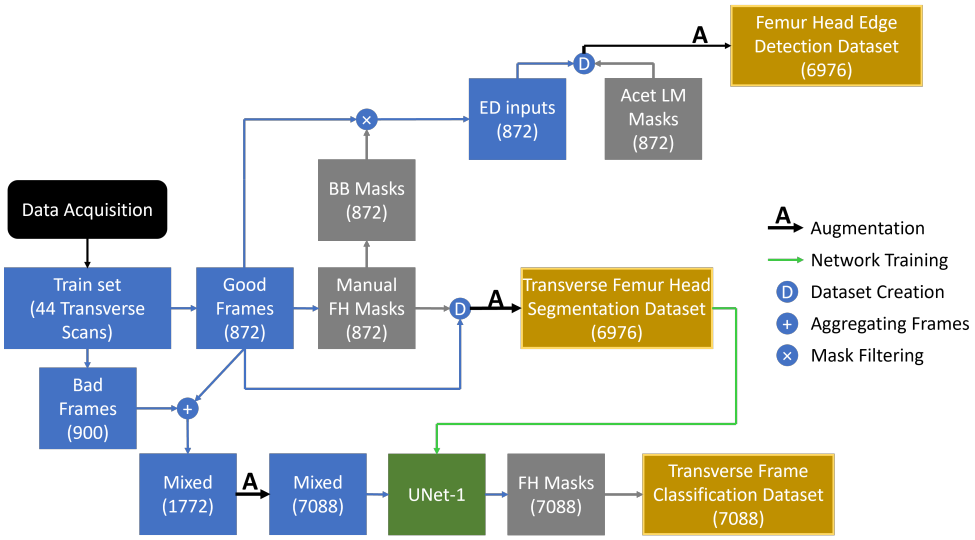


Figure 4.12: B-measurement dataset development process

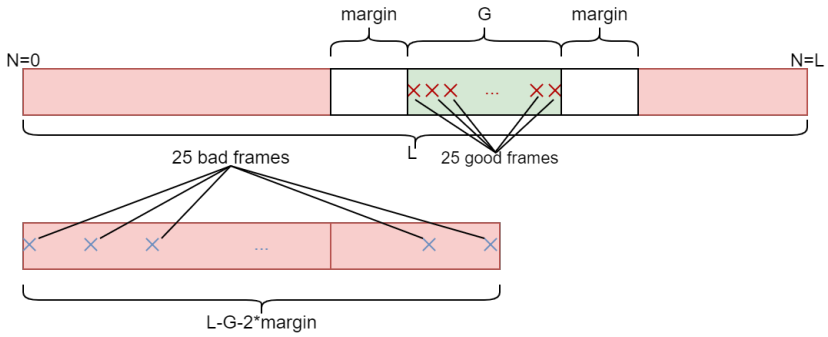


Figure 4.13: The manual procedure of "good" and "bad" frame sampling.

4.5.2.1 Femur Head Segmentation Dataset

For all selected "good" transverse frames, the femur head masks were manually drawn and generated. Using augmentation, the frames and corresponding segmentation masks number was increased by a factor of 8, yielding a final dataset of 6968 images and 6968 segmentation masks. The data set was named as "transverse femur head segmentation dataset" and used to train UNet-1 for femur head segmentation. Figure 4.14 displays a sample transverse frame with the manually delineated femur head and the generated segmentation mask.

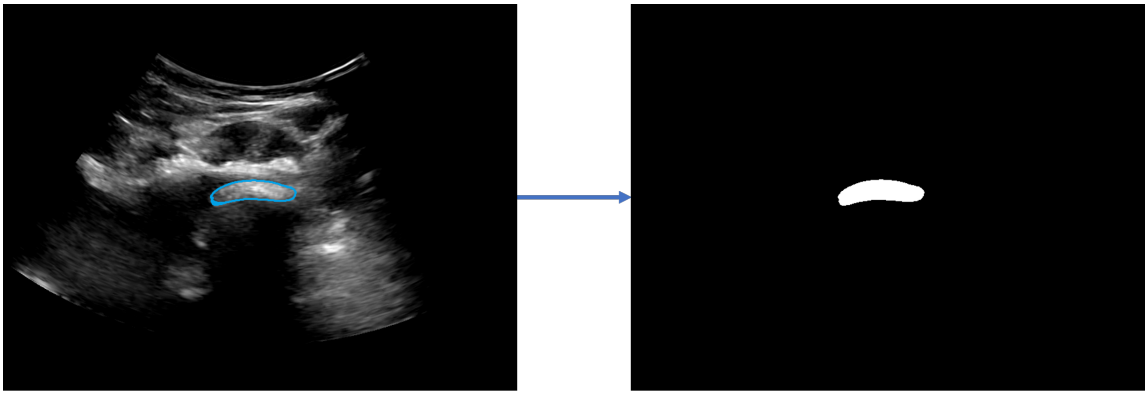


Figure 4.14: A sample transverse US frame with the manually drawn femur head and its corresponding generated segmentation mask.

4.5.2.2 Femur Head Edge Detection Dataset

For all "good" frames, a bounding box with a margin was created around the femur head using the segmentation masks from Section 4.5.2.1, according to Equations (4.5) and (4.6). The bounding box was used to filter out the femur head, and new frames containing only the area within the bounding box were generated. The margin is used to ensure the entire femur head was in the filtered frames. Figure 4.15 depicts a sample US frame, the segmentation mask with the corresponding bounding box, and the created filtered frame by applying the bounding box as a filter. In addition, for all "good" frames the upper edge of the femur head was manually drawn and used to create edge masks. Figure 4.16 depicts a sample frame, the manually drawn femur head edge, and the generated edge mask. The number of filtered frames and the corresponding edge masks was increased by a factor of 8. The 6976 augmented filtered frames were used as input and the 6976 augmented edge masks were used as target to train UNet-2 for femur head edge detection and was named as "femur head edge detection" dataset.

$$x_1 = \min\{A_x\} - m_x,\ x_2 = \max\{A_y\} + m_x, \tag{4.5}$$

$$y_1 = \min\{A_y\} - m_y,\ y_2 = \max\{A_y\} + m_y. \tag{4.6}$$

Here (x_1, y_1) and (x_2, y_2) represent the coordinates of the top-left and bottom-right corners of the cropping box, respectively. A_x and A_y denote the sets of x and y coordinates of the segmented pixels, while m_x and m_y indicate the cropping margins in x and y directions.

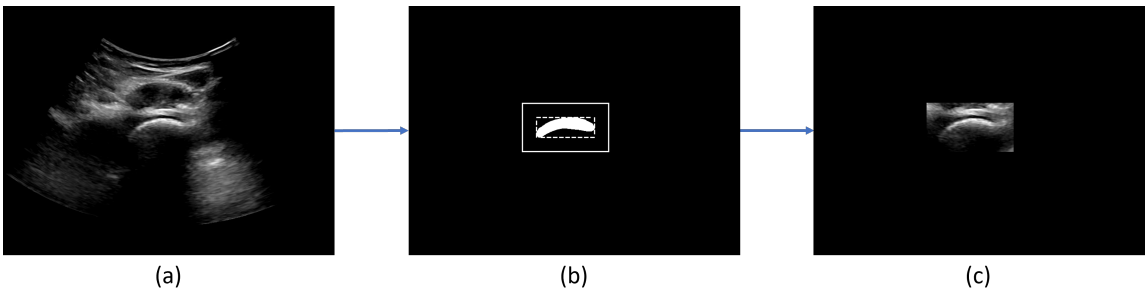


Figure 4.15: A sample transverse US frame (a) with its corresponding segmentation mask (b). Dashed and solid rectangles indicate bounding boxes (no margin and with margin, respectively). Filtered frame (c) created using the bounding box with margin as a filter.

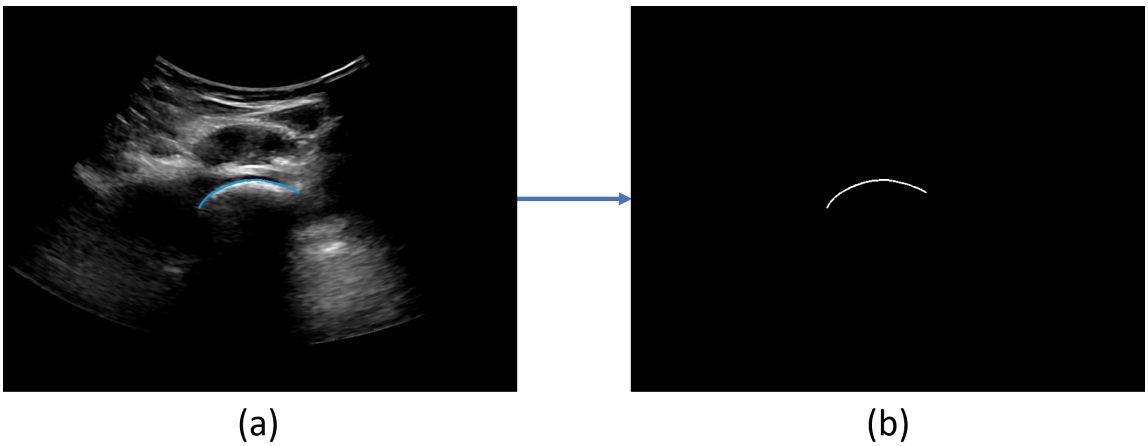


Figure 4.16: A sample transverse US frame with the manually delineated femur head edge and its corresponding generated mask.

4.5.2.3 Transverse Frame Classification Dataset

The selected "good" and "bad" transverse frames were augmented by a factor of 4. The augmented frames were then given to UNet-1, and its outputs were named as "transverse frame classification" dataset with a total size of 7088 images and were used to train CNN-1 for classification of transverse frames into good and bad frames. Figures 4.17 and 4.18 illustrate the UNet-1 output for a sample "good" transverse frame and a sample transverse "bad" frame, respectively.

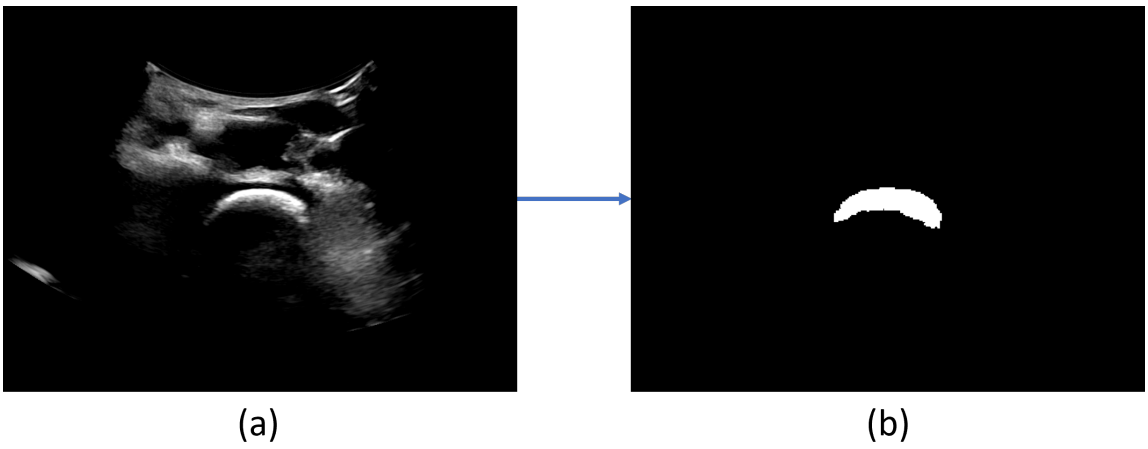


Figure 4.17: A sample good transverse US frame (a) and the femur head (b) segmentation mask obtained using UNet-1.

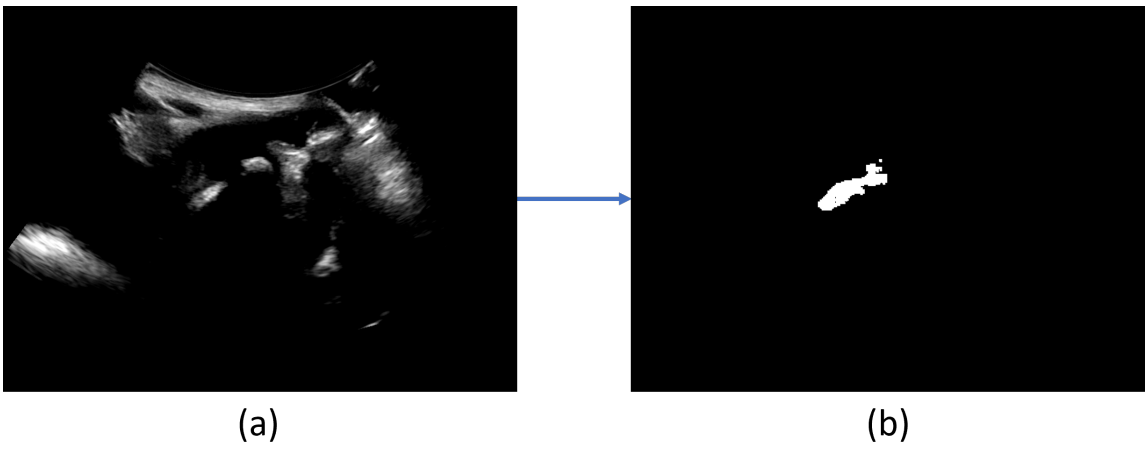


Figure 4.18: A sample bad transverse US frame (a) and the femur head (b) segmentation mask obtained using UNet-1.

4.5.3 A-Measurement Datasets Development

Figure 4.19 depicts the general steps for development of A-measurement datasets. Among the 62 coronal US hip scans, 36 were used for training and 8 for validation. For each scan within the training and validation sets, a sequence of frames showcasing a visible femur head and acetabulum was identified as the target range. Similar to our approach in Section 4.5.2, up to 25 frames within the specified range were selected as "good" frames, and 25 frames outside the target range were selected as "bad" frames, resulting in a total of 871 "good" frames and "900" bad frames.

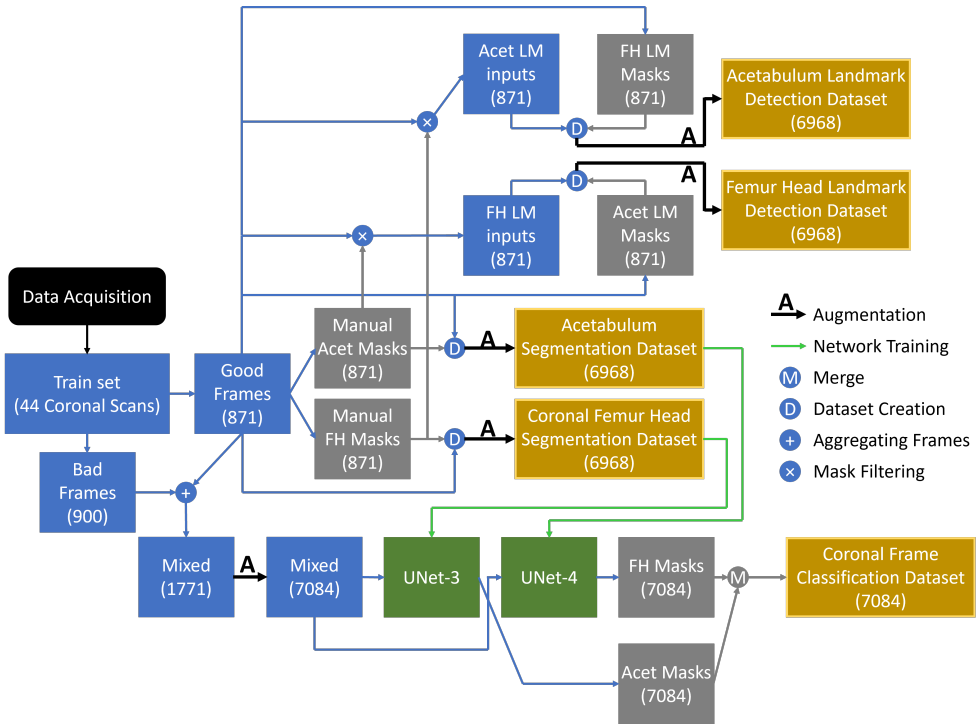


Figure 4.19: A-measurement dataset development process

4.5.3.1 Acetabulum and Femur Head Segmentation Datasets

For every "good" frame selected, both the femur head and the acetabulum were manually delineated, resulting in the creation of separate masks for each structure in each frame (871 acetabulum masks and 871 femur head masks). To expand the dataset size 8 times, augmentation techniques were applied independently to the

masks of the acetabulum and femur head. The augmented masks for the acetabulum, along with their corresponding frames, served as the training data for UNet-3, and named as "acetabulum segmentation" dataset. Similarly, the training data for UNet-4 consisted of the augmented femur head masks and their corresponding frames, named as "coronal femur head segmentation" dataset. Each dataset contained 6968 images and their corresponding segmentation masks. Figure 4.20 shows a sample coronal frame, manually delineated boundaries and created acetabulum and femur head segmentation masks.

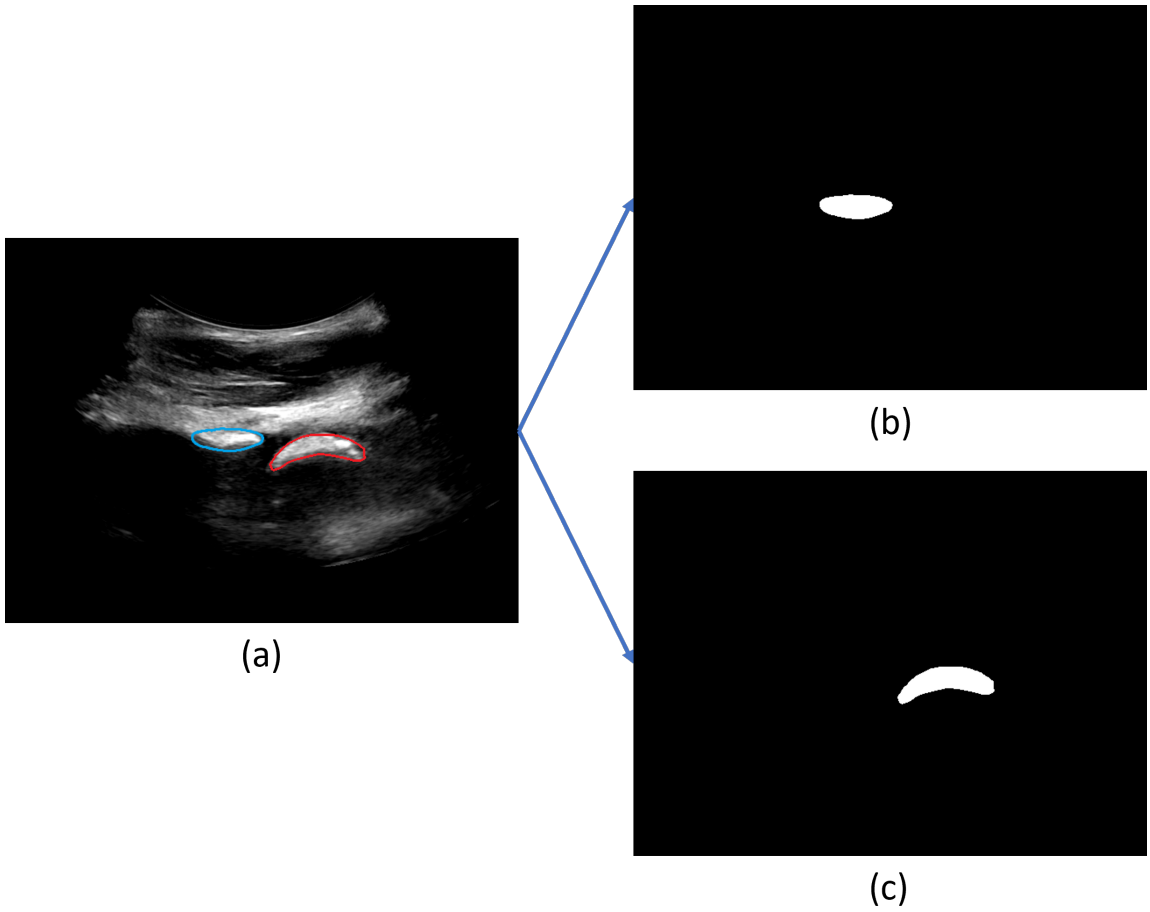


Figure 4.20: A coronal US frame with manually delineated acetabulum (blue), femur head (red), and corresponding generated segmentation masks.

4.5.3.2 Acetabulum and Femur Head Landmark Datasets

For each "good" frame selected, two landmarks, L_1 and L_2, were manually marked as a single pixel: L_1 representing the lateral margin of the femur head, and L_2 indicating the lateral border of the acetabulum. Separate masks for L_1 and L_2 were then generated, yielding 871 acetabulum landmark masks and 871 femur head landmark masks. The size of each landmark within its mask was expanded to a 7×7 square. Figure 4.22 displays a sample frame with manually delineated landmarks and generated masks for each landmark. Furthermore, for every frame, the corresponding segmentation masks from Section 4.5.3.1 were dilated and applied as binary masks to create two types of filtered frames: one containing the acetabulum and the other containing the femur head. The dilation was to make sure the boundaries of the femur head and acetabulum are included in the filtered frames. Figure 4.21 displays a sample coronal frame, the acetabulum and femur head segmentation masks, and the corresponding filtered frames. The filtered frames served as the input, and the landmark masks as the targets, for training UNets dedicated to landmark detection. The dataset was expanded eightfold by separately augmenting the input images and corresponding ground truth segmentation masks for the femur head and the acetabulum. These augmented datasets were named "acetabulum landmark detection" and "femur head landmark detection", respectively, and each contained 6968 images and 6968 target masks. Utilizing these augmented filtered frames and landmark masks, UNet-5 was specifically trained for detecting the femur head landmark, while UNet-6 focused on the acetabulum landmark.

4.5.3.3 Coronal Frame Classification Dataset

The selected "good" and "bad" frames were augmented by a factor of 4. The augmented frames were given to UNet-5 and UNet-6, and their outputs were merged into a single mask with the pixel values of 127 and 255 for UNet-5 and UNet-6 outputs, respectively. The merged outputs were used to train CNN-2 for classifying "good" and

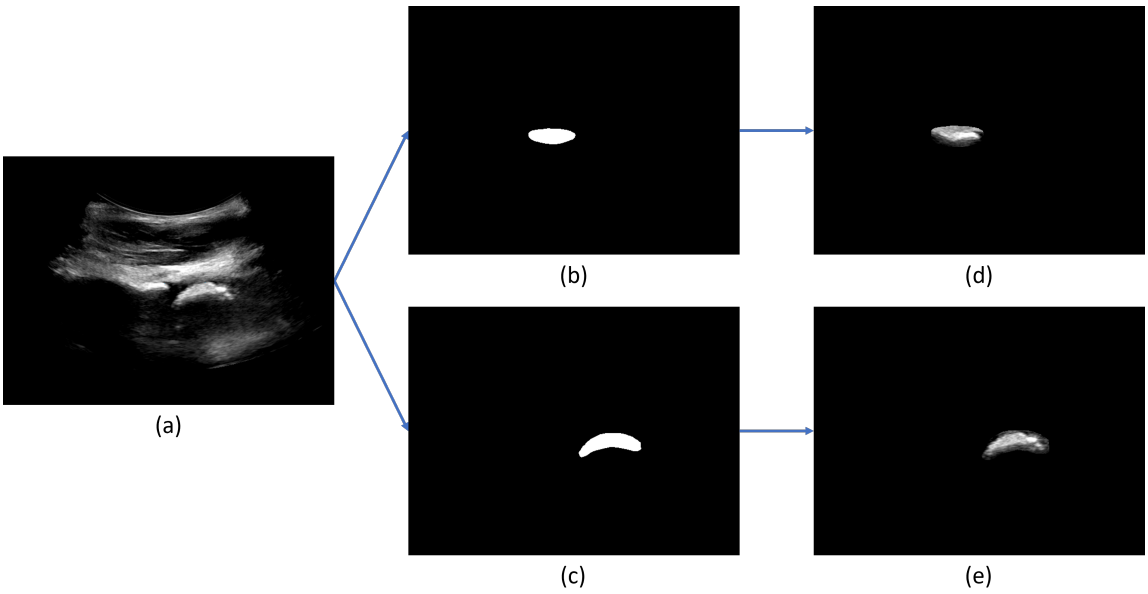


Figure 4.21: A sample coronal frame (a) with segmentation masks (b, c) applied as filters, resulting in filtered frames (d, e).

"bad" frames in coronal US images, named as "coronal frame classification" dataset, containing 7084 images. Figures 4.23 and 4.24 display sample good and bad coronal frames respectively, along with the UNet-3 and UNet-4 output and the created mergerd mask for each frame.

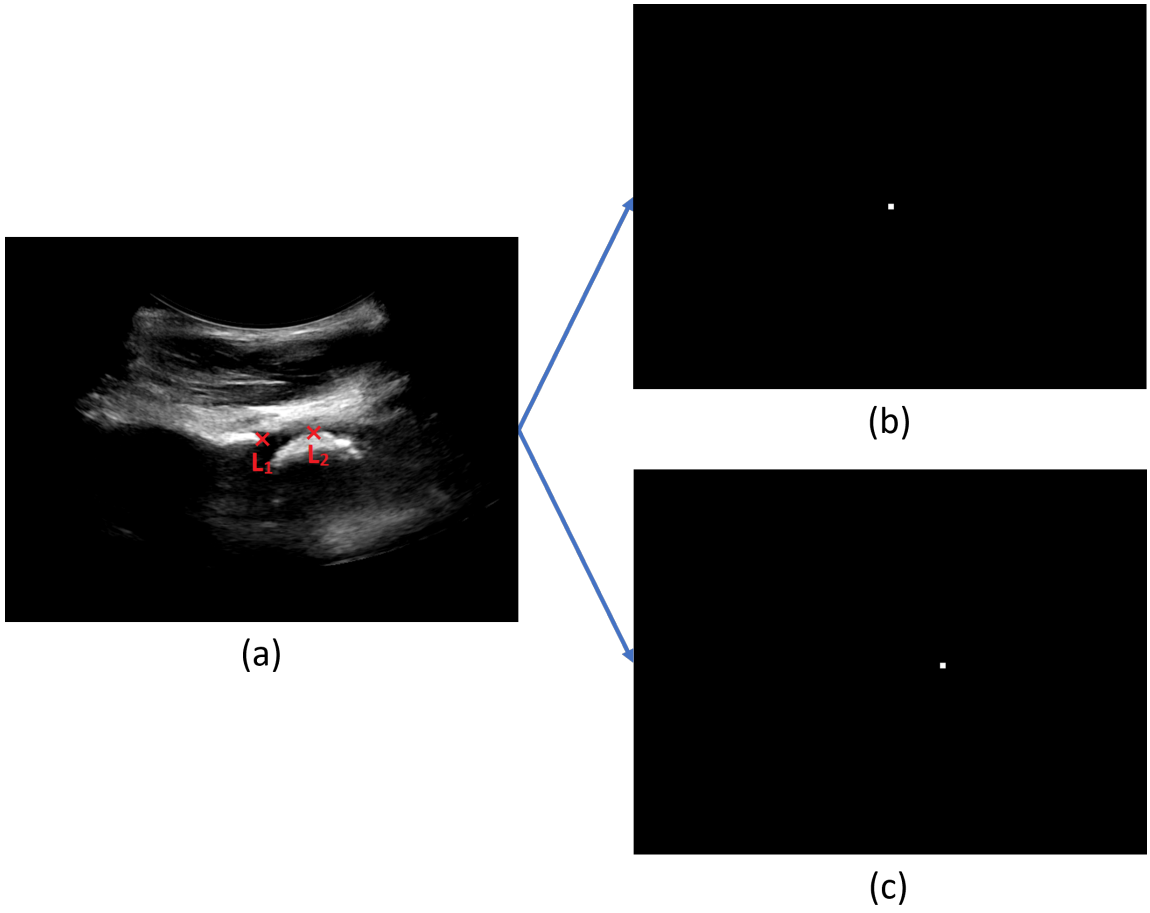


Figure 4.22: A sample US frame with manually delineated landmarks for the acetabulum (L_1) and femur head (L_2), and their corresponding generated masks.

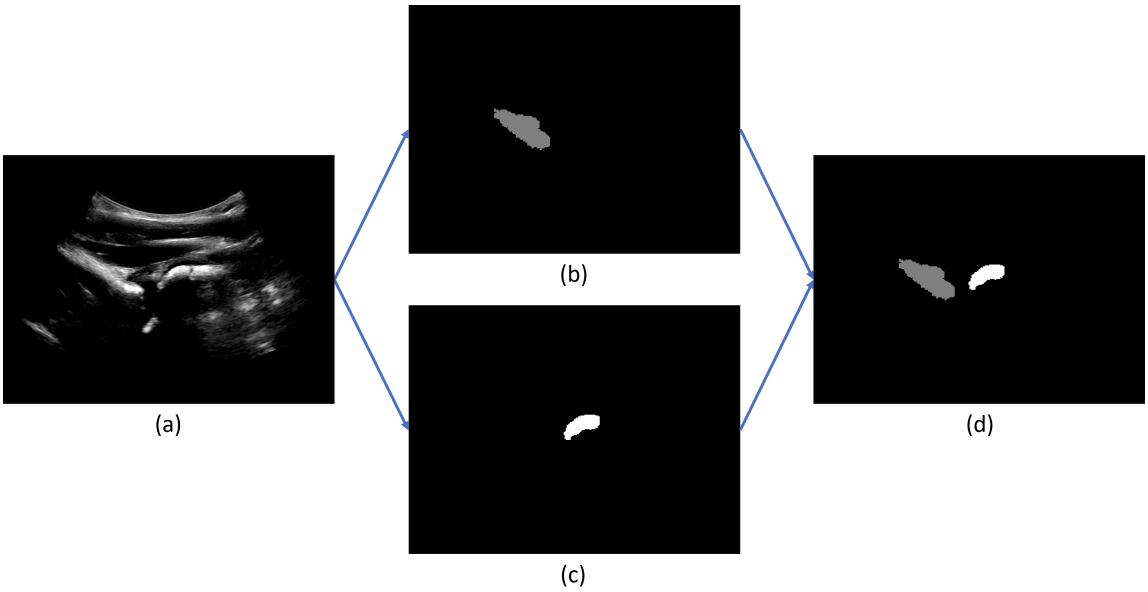


Figure 4.23: A sample good coronal US frame (a), acetabulum (b) and femur head (c) segmentation masks obtained using UNet-3 and UNet-4, and the resulting merged mask (d).

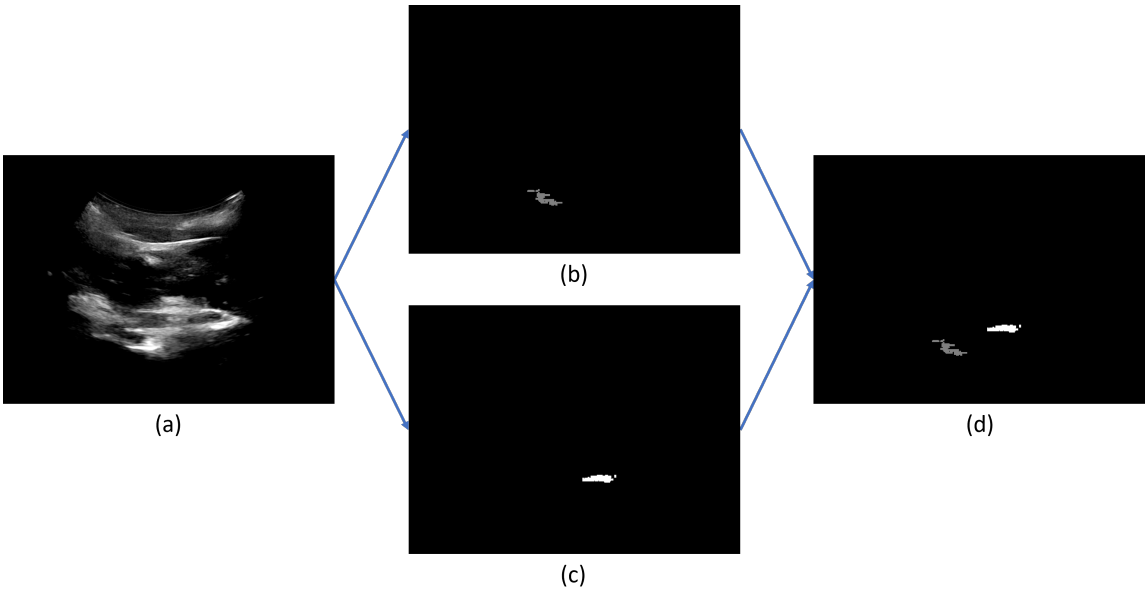


Figure 4.24: A sample bad coronal US frame (a), acetabulum (b) and femur head (c) segmentation masks obtained using UNet-3 and UNet-4, and the resulting merged mask (d).

4.6 Obtaining "B" Measurements

In this section we will explain the flow diagram of making new measurements for new hip transverse view songorams, as depicted in Figure 4.2a with green arrows. New measurements are made in six steps as described below:

1. Preprocessing: All US frames are preprocessed using the windowing function described in Section 4.3.

2. Segmentation: The preprocessed US frames are given to UNet-1 for segmentation. The outputs of UNet-1 are filtered by DBSCAN clustering.

3. Frame Selection: The filtered outputs from UNet-1 are fed into CNN-1. The output from the softmax layer at the end of CNN-1 is used to estimate the probability of detecting the femur head in each frame, referred to as the frame score. A range of frames with the highest probability for femur head detection is selected by applying a discrete filter to these estimated probabilities. Details on applying the discrete filter are provided in Section 4.6.1.

4. Edge detection: The ROI is determined and filtered (as described in Section 4.5.2.2) for all selected using the corresponding filtered UNet-1 of that frame. Afterwards, the selected frames with the determined ROI are given to Unet-2 for edge detection. The outputs of UNet-2 are filtered by DBSCAN clustering.

5. Circle fitting: A circle is fitted to filtered outputs of UNet-2 using the Taubin method. The Taubin method will be explained in Section 4.6.2. The diameter of the fitted circle is considered as an estimation of variable "B".

6. Statistical Inference: In this step, a final estimated value for variable "B" is estimated. First, outlier measurements are omitted by IQR method as described

in Section 4.6.3. Finally, the mean and standard deviation of the measurements are calculated to have an estimation of "B" and possible measurement error.

4.6.1 Discrete Signal Filtering for Frame Selection

After obtaining the frame scores for all frames (L) in the hip sonogram, a filter is convolved with the scores to obtain a smoother function. The maximum point of this signal is determined, and a ratio of it, is used as the threshold for frame selection. Finally, all frames near the maximum point that have a filtered probability higher than the threshold are selected. Equations (4.7) to (4.10) show the exact mathematical equations used to select the desired frames.

$$y_f[n] = y_r[n] * f[n] \tag{4.7}$$

$$n_0 = \arg\max_n \ y_f[n] \tag{4.8}$$

$$N_i = \arg\min_n \{y_f[n] \mid \forall x,\ n \leq x < n_0 \ :\ y_f[x] > r \cdot f[n_0]\} \tag{4.9}$$

$$N_f = \arg\max_n \{y_f[n] \mid \forall x,\ n_0 \leq x < n \ :\ y_f[x] > r \cdot f[n_0]\} \tag{4.10}$$

where $f[n]$ is the applied filter, $y_r[n]$ is the raw predicted scores, $y_f[n]$ is the scores convolved with filter, n_0 is the number of the frame with highest score in $y_f[n]$, r is the desired ratio for determining the threshold, and N_i, N_f are the first and last frame in the selected range.

In this study we used two filters for achieving the filtered scores: the moving average filter and the Gaussian filter. Equations (4.11) and (4.12) display the mathematical equations for the moving average and the Gaussian filters, respectively.

$$f[n] = \begin{cases} 1, & 0 \leq n \leq N, \\ 0, & o.w., \end{cases} \tag{4.11}$$

where N is the length of the moving average filter.

$$f[n] = \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{n^2}{2\sigma^2}, \tag{4.12}$$

where σ is the standard deviation of the Gaussian filter.

A sample of the predicted probabilities for a given sonogram is depicted in Figure 4.25. Figures 4.26 and 4.27 display the filtered probabilities by the moving average and Gaussian filters, and the selected frames by each filter, respectively.

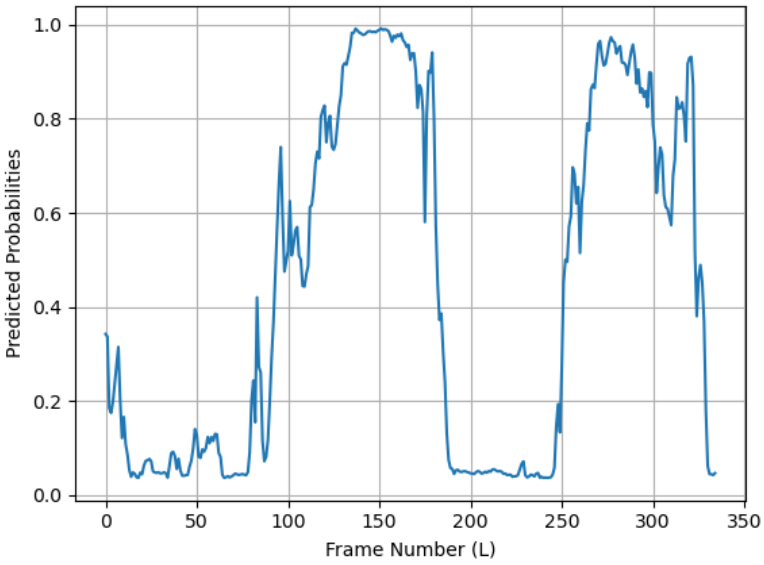


Figure 4.25: The calculated probabilities by CNN-1/2 for a sample US hip scan.

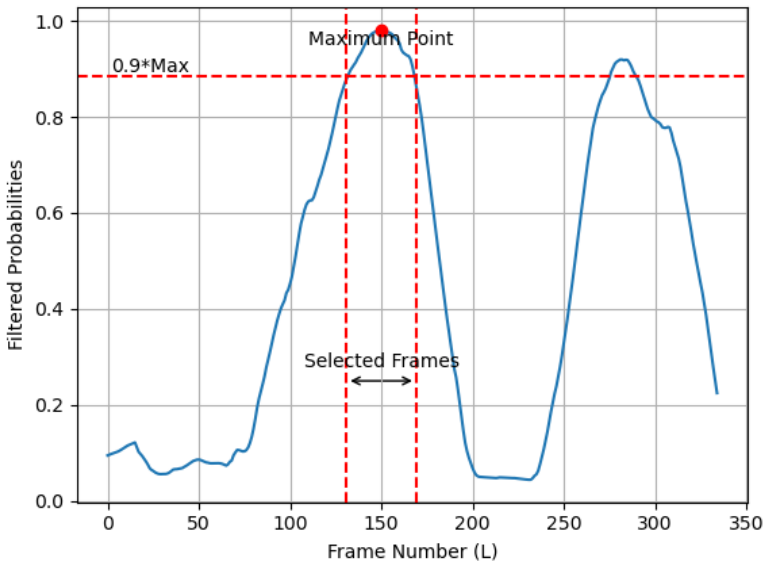


Figure 4.26: Filtered probabilities with **moving average filter**, highlighting maximum point, selection threshold, and resulting frame selection in a sample US sonogram.

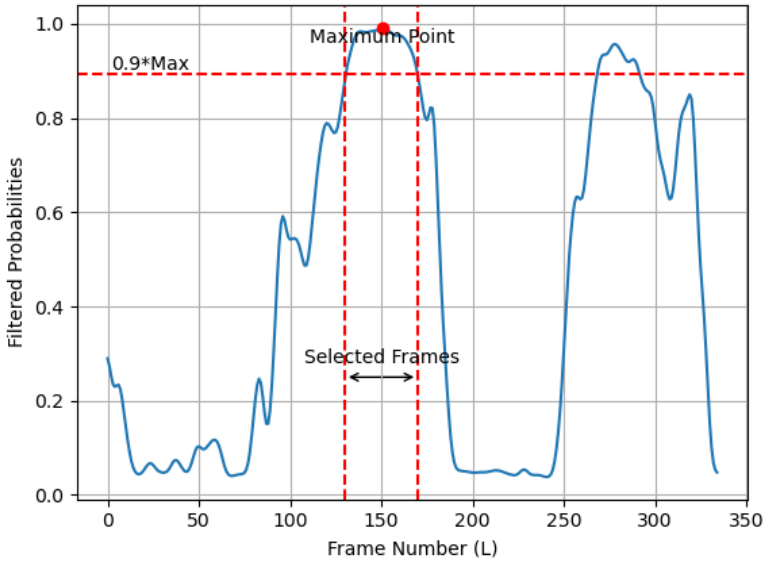


Figure 4.27: Filtered probabilities with **Gaussian filter**, highlighting maximum point, selection threshold, and resulting frame selection in a sample US sonogram.

4.6.2 Taubin Method

Given a set of data points, fitting a circle is equivalent to optimizing Equation (4.13):

$$\mathcal{F}(a,b,R) = \sum_{i=1}^{N}[(x_i-a)^2+(y_i-b)^2-R^2]^2, \tag{4.13}$$

where a, b, R are the coordinates and radius of the circle and (x_i, y_i) are the coordinates of the data points. Equation (4.13) can be written as:

$$\mathcal{F} = \sum_{i=1}^{N}[(r_i-R)(r_i+R)]^2 = \sum_{i=1}^{N} d_i^2 D_i^2, \tag{4.14}$$

where $r_i = \sqrt{(x_i-a)^2+(y_i-b)^2}$, $d_i = r_i - R$, and $D_i = r_i + R$. Observe that $D_i = d_i + 2R$, hence with the assumption of $|d_i| \ll R$ we will have:

$$\mathcal{F} \approx 4R^2 \sum_{i=1}^{N} d_i^2. \tag{4.15}$$

It can be shown that minimizing Equation (4.15) is equivalent to minimization of:

$$\mathcal{F}(A,B,C,D) \approx \sum_{i=1}^{N}(Az_i + Bx_i + Cy_i + D)^2, \tag{4.16}$$

subject to the constraint

$$4A^2\overline{z} + 4AB\overline{x} + 4AC\overline{y} + B^2 + C^2 = 1, \tag{4.17}$$

where

$$z = \sum_{i=1}^{N} x_i^2 + y_i^2, \overline{z} = (\sum_{i=1}^{N} z_i)/n, \overline{x} = (\sum_{i=1}^{N} x_i)/n, \overline{y} = (\sum_{i=1}^{N} y_i)/n.$$

and

$$a = -\frac{B}{2A}, b = -\frac{C}{2A}, R^2 = \frac{B^2+C^2-4AD}{4A^2}.$$

Equation (4.16) is a quadratic function of D, therefore by keeping A, B, C as constant values we can obtain the minimum of $\mathcal{F}$ with respect to D as:

$$D = -A\overline{z} - B\overline{x} - C\overline{y}. \tag{4.18}$$

Substituting Equation (4.18) into Equations (4.16) and (4.17), we will have to minimize:

$$\mathcal{F}(A, B, C, D) \approx \sum_{i=1}^{N} [A(z_i - \overline{z}) + Bx_i + Cy_i]^2, \tag{4.19}$$

subject to:

$$4\overline{z}A^2 + B^2 + C^2 = 1. \tag{4.20}$$

Introducing a new variable, $A_0 = 2\overline{z}^{1/2}A$, we have to minimize:

$$\mathcal{F}(A, B, C, D) \approx \sum_{i=1}^{N} [A_0 \frac{z_i - \overline{z}}{2\overline{z}^{1/2}} + Bx_i + Cy_i]^2, \tag{4.21}$$

subject to

$$A_0^2 + B^2 + C^2 = 1. \tag{4.22}$$

We can solve the optimization problem by reducing it to an eigenvalue problem. The Equation (4.21) can be written in matrix form as

$$\mathcal{F}(A, B, C, D) = \|\boldsymbol{X_0 A_0}\|^2 = \boldsymbol{A_0}^T(\boldsymbol{X_0^T X_0})\boldsymbol{A_0}, \tag{4.23}$$

where $\boldsymbol{A_0} = (A_0, B, C)^T$ and

$$T = \begin{bmatrix} (z1 - \overline{z_1})/(2\overline{z_1}^{1/2}) & x_1 & y_1 \\ \vdots & \vdots & \vdots \\ (z_n - \overline{z_n})/(2\overline{z_n}^{1/2}) & x_n & y_n \end{bmatrix}. \tag{4.24}$$

The constraint in Equation (4.22) means $\|\boldsymbol{A_0}\| = 1$, i.e. $\boldsymbol{A_0}$ must be a unit vector. Therefore, the minimum of Equation (4.24) is attained on the unit eigenvector of the matrix $\boldsymbol{X_0}^T\boldsymbol{X_0}$ corresponding to its smallest eigenvalue.

4.6.3 Interquartile Range Method

The Interquartile Range (IQR) is a measure of variability in a dataset that is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). The quartiles divide a dataset into four equal parts, with Q1 representing the 25^{th} percentile and Q3 representing the 75^{th} percentile. Any data point that falls below

$Q1 - 1.5IQR$ or above $Q3 + 1.5IQR$ is considered an outlier. The IQR method is a relatively simple and robust method for identifying outliers, but it may not detect all outliers. Therefore, we use it iteratively until no outlier is detected. Figure 4.28. depicts how the IQR method works in a box plot.

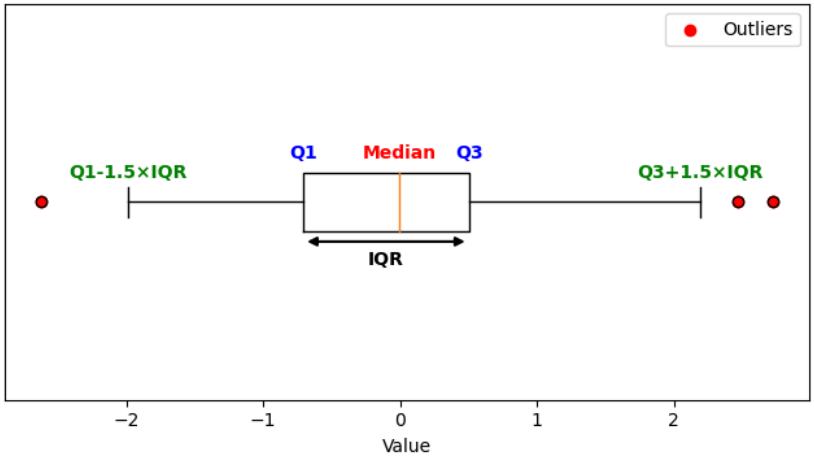


Figure 4.28: The IQR method for detecting outliers

4.7 Obtaining "A" Measurements

This section presents the flow diagram for making new measurements on new hip coronal view sonograms, as shown in Figure 4.2b with green arrows. The process of obtaining new "A" measurements involves five steps, which are described below:

1. Preprocessing: All US frames are preprocessed using the windowing function described in Section 4.3.

2. Segmentation: The preprocessed frames are fed to UNet-3 and UNet-4 for segmentation. UNet-3 identifies the femur head, while UNet-4 segments the acetabulum. The outputs of both networks are filtered using DBSCAN clustering.

3. Frame Selection: The filtered outputs of UNet-3 and UNet-4 for each frame are merged into a single mask as explained in Section 4.5.3.3. The masks are then

fed into CNN-2, and the probability of detecting the femur head and acetabulum in each frame, referred to as the frame score, is estimated by the output from the softmax layer at the end of CNN-2. A range of frames with the highest detection probability is selected using a discrete filter, as described in Section 4.6.1.

4. Landmark detection: The filtered outputs of UNet-3 and UNet-4 are used to determine and filter the ROI in selected frames and create separate images containing acetabulum and femur head ROIs, as described in Section 4.5.3.2. The image with acetabulum ROI is given to UNet-5 and the image with femur head ROI is given to UNet-6. UNet-5 processes the femur head frame to detect the femur head landmark area, while UNet-6 processes the acetabulum frame to detect the acetabulum landmark area. The outputs of UNet-5 and UNet-6 are filtered using DBSCAN clustering.

5. Vertical Distance Calculation: The centroids of the segmented area around the acetabulum and femur head landmarks are calculated. These centroids represent the detected locations of the landmarks. As the US images in the coronal view correspond to 90 °rotated radiographs (refer to Figure 2.4), the vertical distance between L_1 and L_2 is measured as the "A" value.

6. Statistical Inference: Outlier measurements are identified and removed using the iterative IQR method (see Section 4.6.3). Finally, the mean and standard deviation of the measurements are calculated to estimate the "A" value and assess possible measurement errors.

4.8 Model Training and Evaluation

All 62 hip scans were split into three parts: train (36), validation (8), and test (18) datasets. The training dataset is used for developing the algorithm and training the deep networks of the model. While training, the validation set was used to find the

optimums number of epochs. In addition, the performance of the trained models using various architectures and parameters on validation set was used to find the best models. Finally, the test dataset was used for the final evaluation of the developed model.

4.9 Chapter Summary

We gathered US scan data of CP children at a local hospital, adhering to ethical approvals and a standard protocol. A total of 54 hip scans were obtained. We developed an automated method to measure MP from these US images. This method involved two algorithms: one to quantify the femoral head width ("B"), and the other to measure the femoral head displacement from the acetabulum ("A"). To determine "B", we segmented the femoral head from transverse US frames using a UNet. A CNN and a discrete signal filter assisted in frame selection. We then detected the femoral head edge (using another UNet), fit a circle to the edge with Taubin's method, and aggregated measurements for the final "B" value. Similarly, for "A", we segmented the femoral head and acetabulum from coronal US frames. A CNN and discrete signal filter aided in frame range selection. We then detected landmarks on the acetabulum and femoral head, ultimately deducing the "A" value from these measurements. Finally, we outlined our strategy for training, optimizing, and testing the developed method.

Chapter 5

Training and Evaluation of AI Model

This chapter outlines the training, validation, and testing stages of the developed method. Section 5.1 details the training, validation, and optimization procedure of the developed method. Section 5.2 presents the model's evaluation results on the test sets. A comprehensive discussion of these results is provided in Section 5.3, followed by a summary of the chapter's findings in Section 5.4. Portions of this chapter were submitted in the paper: Yousefvand, R., Pham, T-T., Le, L.H., Andersen, J., Lou, E. H, "Applying Deep Learning for Automatic Measurement of Migration Percentage from Ultrasound Images in Children with Cerebral Palsy," Medical & Biological Engineering & Computing, 2024, (Submitted).

5.1 Training Deep Learning Models

In this section we present the training plots for UNets and convolutional neural networks (CNNs) with different architectures in the developed algorithm. The best architecture for UNets and CNNs was selected according to the performance on validation set.

5.1.1 UNet training

UNets were used in the developed algorithm for segmentation, edge detection, and landmark detection. The details about the architecture of the UNet models can be found in Section 4.4.2. Each UNet model was trained using the SCCE loss function, and Adam optimizer was used to update the weights of the models with a learning rate of 0.0001. To assess the performance of the models on both training and validation datasets, we utilized the dice coefficient as the evaluation metric. Equation (5.1) gives the formula for calculation of dice coefficient. Throughout the training process, we monitored the average dice coefficient of validation set, and if it failed to improve for ten consecutive epochs, we considered it as an indication to stop training.

$$y = \frac{2|X \cap Y|}{|X \cup Y|}, \tag{5.1}$$

where X and Y are predicted and ground truth masks.

For each UNet, four CNN models, DenseNet121, ResNet152, ResNet50, and VGG19 were tested as the encoder to obtain the best UNet structure. After training each UNet with different encoders, the architecture that achieved the highest average dice coefficient on validation set was selected.

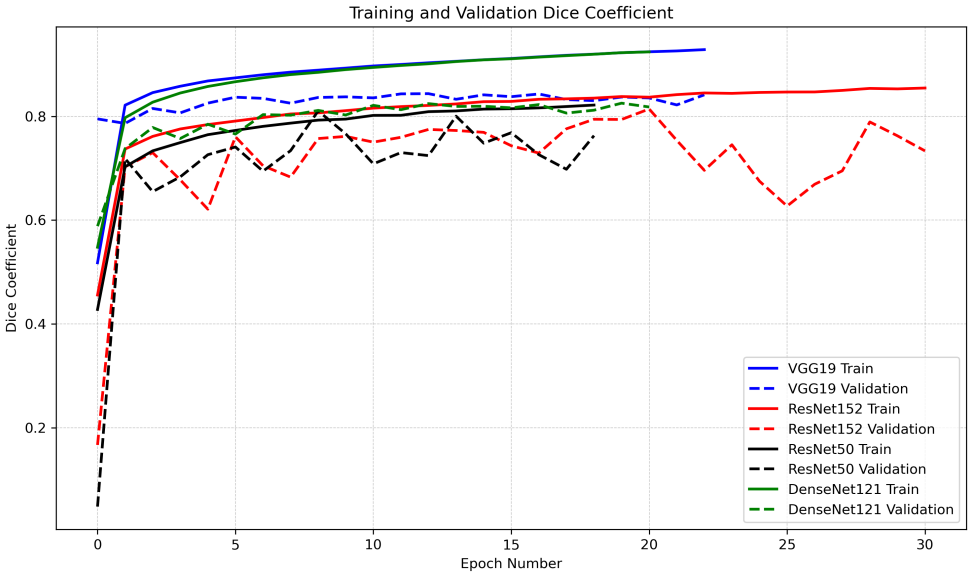
5.1.1.1 Segmentation Networks

UNet-1 was used to segment the femur head from transverse ultrasound (US) frames and UNet-3 and UNet-4 were used to segment the femur head and acetabulum from coronal US frames. "transverse femur head segmentation" dataset (Section 4.5.2.1) was used for training the UNet-1 and "acetabulum segmentation" and "coronal femur head segmentation" datasets (Section 4.5.3.1) were utilized for training UNet-3 and UNet-4, respectively.

Figures 5.1 to 5.3 display the training and validation losses as well as the average dice coefficient values on the training and validation sets during the training process of UNet-1, UNet-3, and UNet-4, respectively. Furthermore, Table 5.1 presents the final

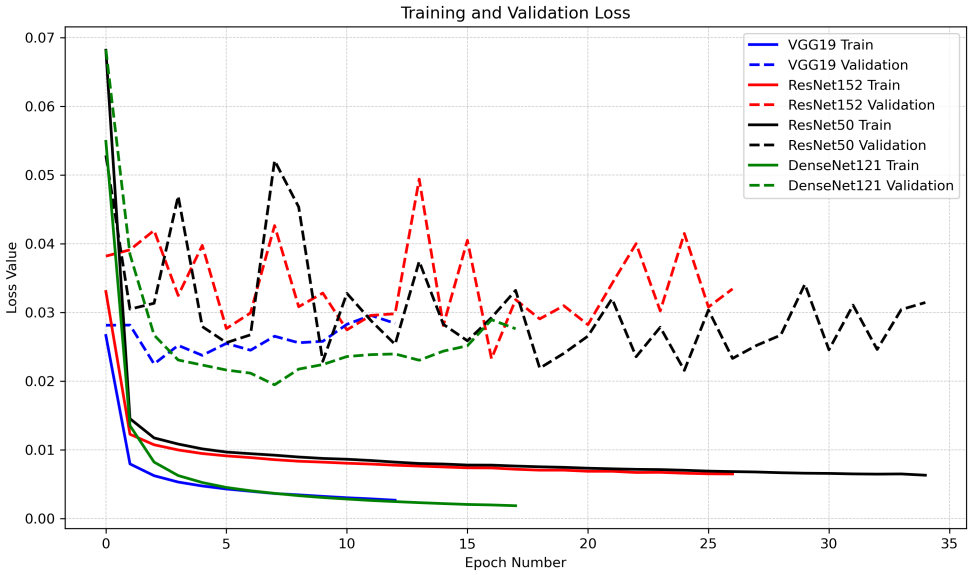
dice coefficient values achieved on the validation set for UNet-1, UNet-3, and UNet-4, respectively. From these results, it is evident that overall, VGG19 outperformed other encoders, obtaining a higher average dice coefficient for UNet-1, UNet-3, and UNet-4. Therefore, we employed VGG19 as the encoder for these models.

(a)

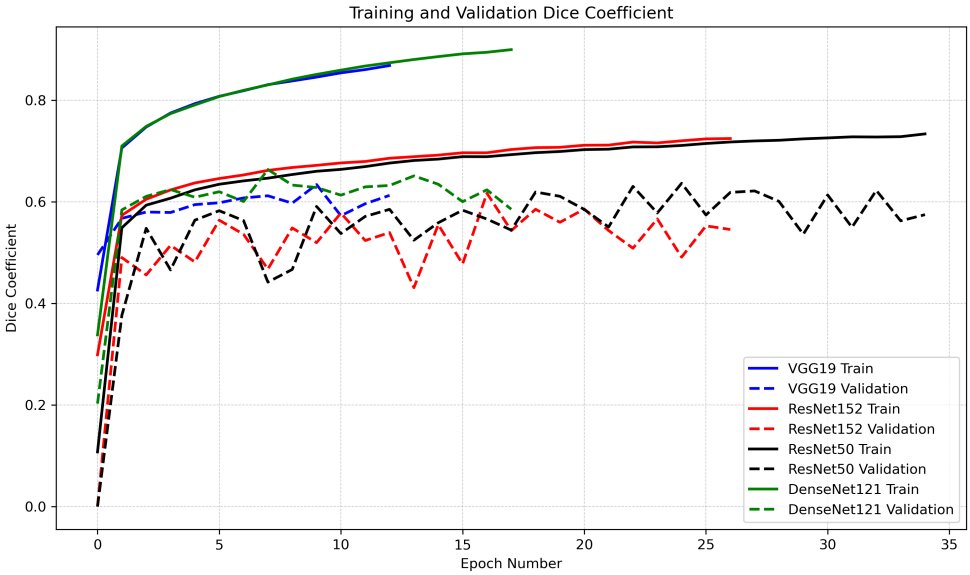


(b)

Figure 5.1: The SCCE loss (a) and dice coefficient (b) values for the training and validation sets during the **UNet-1** training process.

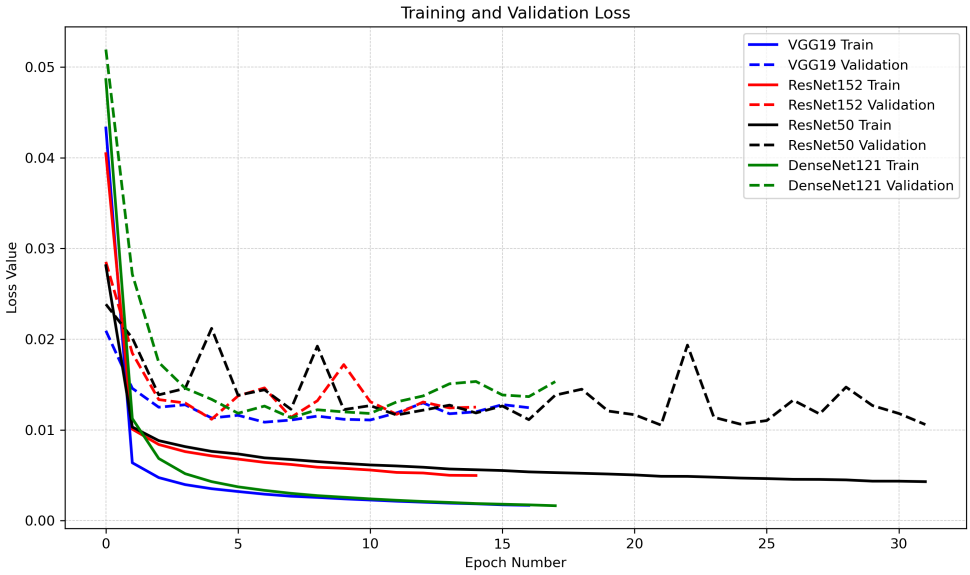


(a)

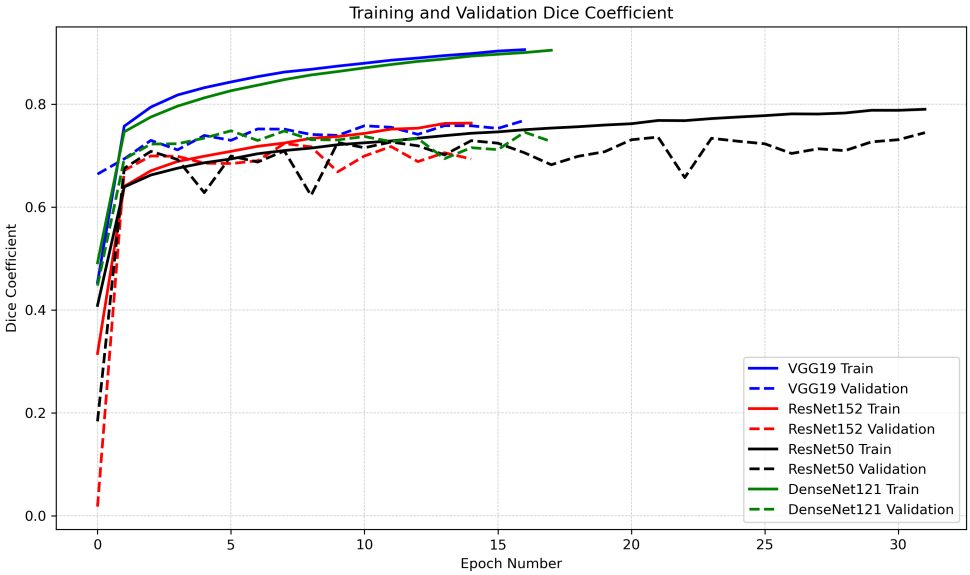


(b)

Figure 5.2: The SCCE loss (a) and dice coefficient (b) values for the training and validation sets during the **UNet-3** training process.



(a)

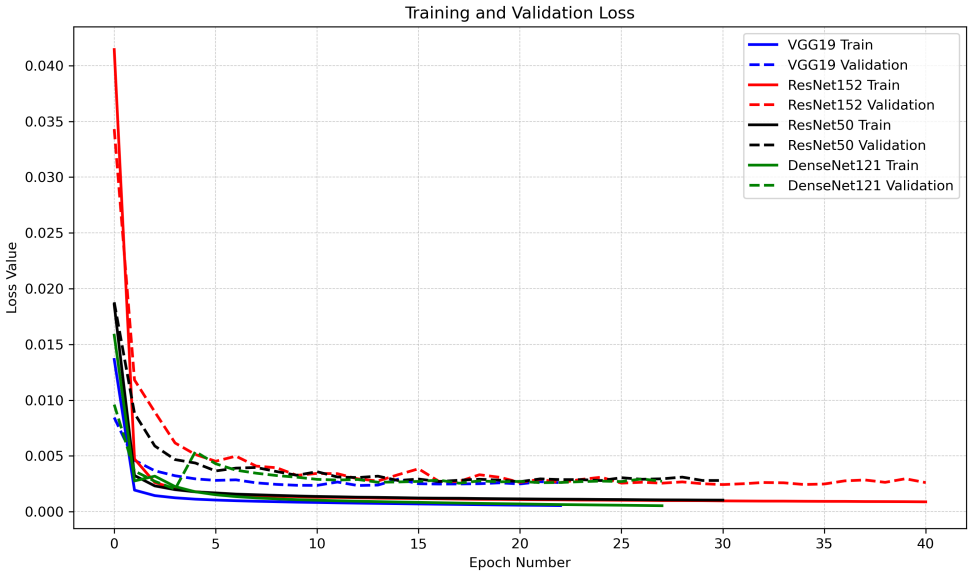


(b)

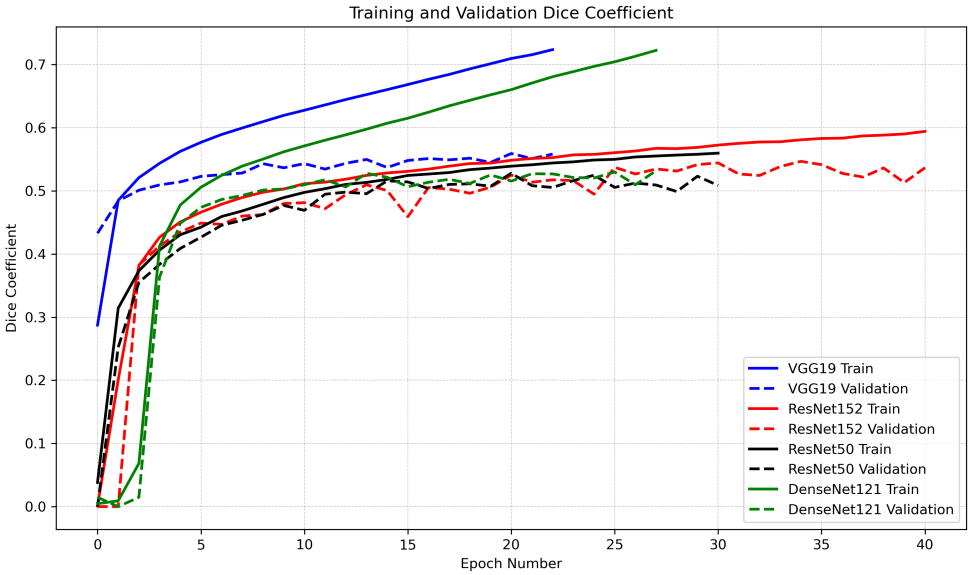
Figure 5.3: The SCCE loss (a) and dice coefficient (b) values for the training and validation sets during the **UNet-4** training process.

5.1.1.2 Edge Detection Network

We used UNet-2 for edge detection on transverse US frames. The "femur head edge detection" dataset, as described in Section 4.5.2.2, was utilized for training UNet-2. Since the number of edge pixels was very small compared to the total number of pixels in each mask, we employed a weighted SCCE loss function to prioritize the accurate identification of edge pixels, assigning a coefficient of 3 to errors for the edge pixels. The training and validation losses as well as the dice coefficient values during training are depicted in Figure 5.4. Moreover, the final dice coefficient values for all models are summarized in Table 5.1. The results indicated that VGG19 outperformed other encoders overally, obtaining a higher average dice coefficient for UNet-1 and UNet-4, and a close to the highest for UNet-3. For the sake of consistency and superior overall performance of VGG19 across all three models, we decided to employ VGG19 as the encoder for these models.



(a)



(b)

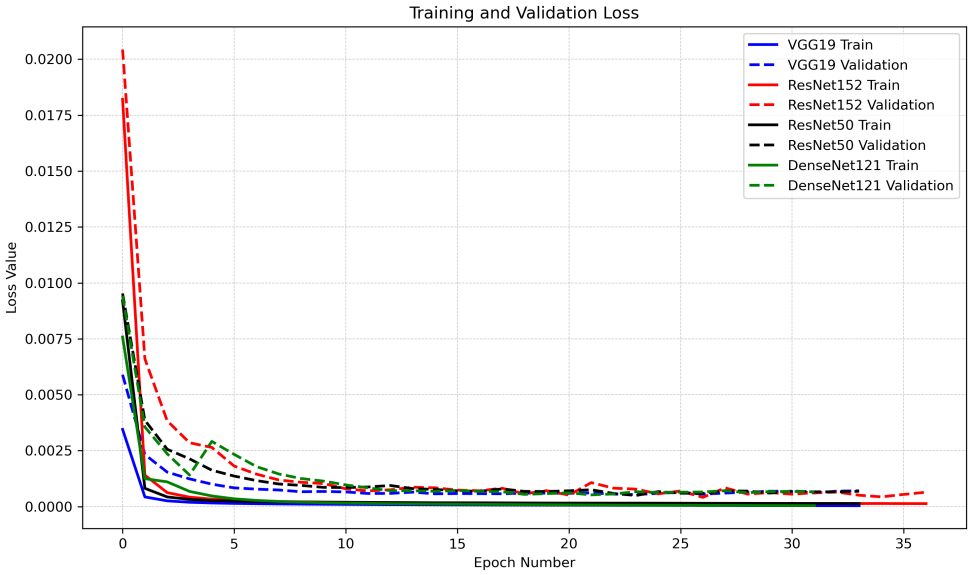
Figure 5.4: The weighted SCCE loss (a) and dice coefficient (b) values for the training and validation sets during the **UNet-2** training process.

5.1.1.3 Landmark Detection Networks

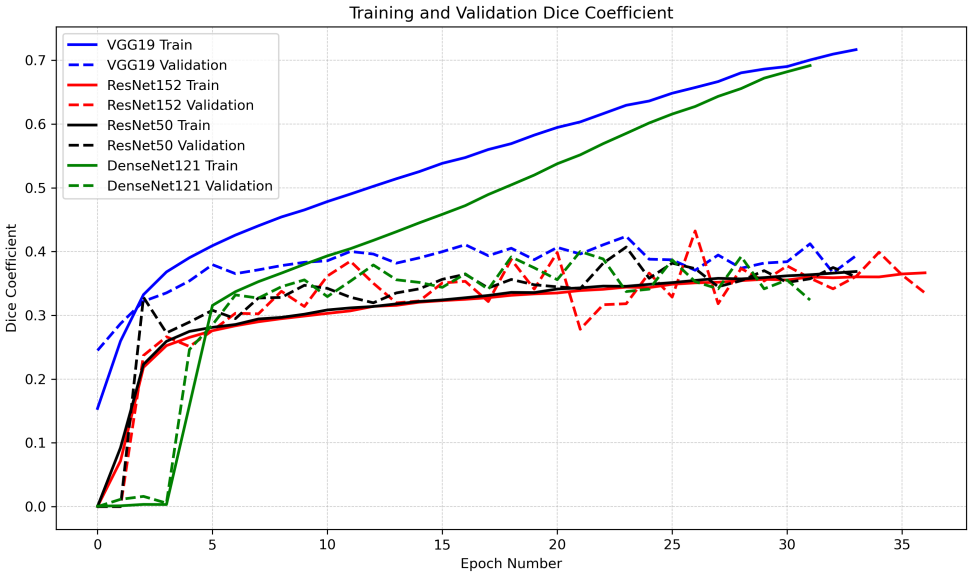
We employed UNet-5 and UNet-6 to detect acetabulum and femur head landmarks on coronal US frames. The "acetabulum landmark detection" and "femur head landmark detection" datasets (refer to Section 4.5.3.2) were utilized for training UNet-5 and UNet-6, respectively. During training, we utilized a weighted SCCE loss function, where landmark pixels were assigned a weight 10 times higher than background pixels. The training and validation losses as well as the dice coefficient values of all networks during training are depicted in Figure 5.6 for UNet-5 and Figure 5.5 for UNet-6. Table 5.1 displays the final dice coefficient values for UNet-5 and UNet-6, respectively. As observed, UNet models with VGG19 as the encoder demonstrated better performance, leading to their selection as the encoder for both UNet-5 and UNet-6.

Table 5.1: The mean Dice coefficient values of the validation set after the training of UNet models with various encoder architectures.

	ResNet50	ResNet152	VGG19	DenseNet121
UNet-1	0.81	0.81	0.84	0.83
UNet-2	0.53	0.55	0.56	0.53
UNet-3	0.64	0.62	0.63	0.66
UNet-4	0.74	0.72	0.77	0.66
UNet-5	0.32	0.31	0.43	0.38
UNet-6	0.41	0.43	0.42	0.40

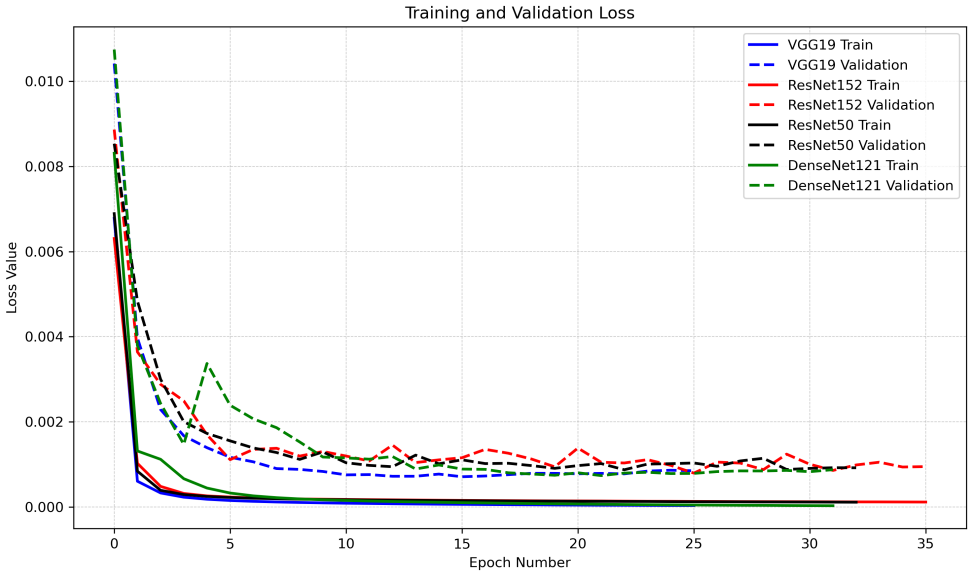


(a)

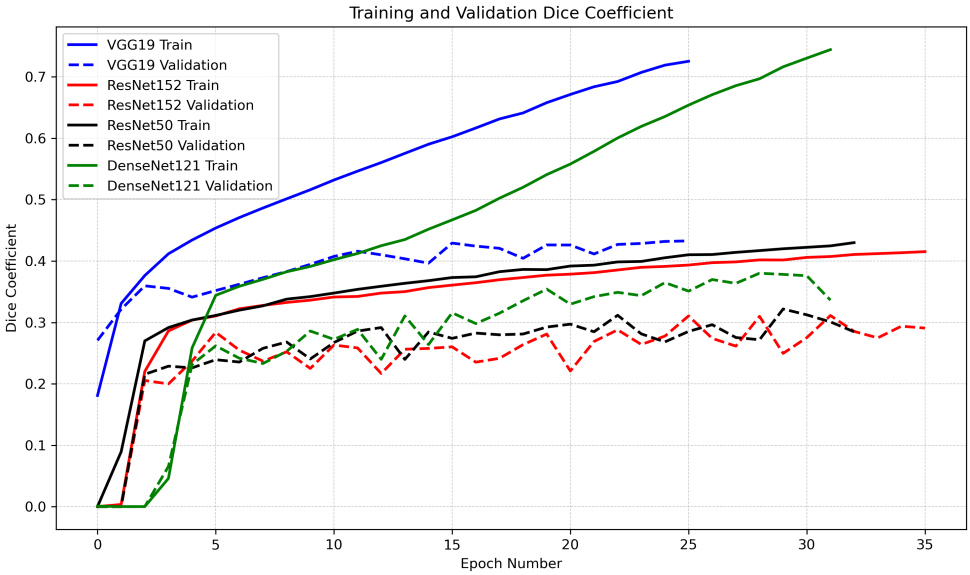


(b)

Figure 5.5: The weighted SCCE loss (a) and dice coefficient (b) values for the training and validation sets during the **UNet-5** training process.



(a)



(b)

Figure 5.6: The weighted SCCE loss (a) and dice coefficient (b) values for the training and validation sets during the **UNet-6** training process.

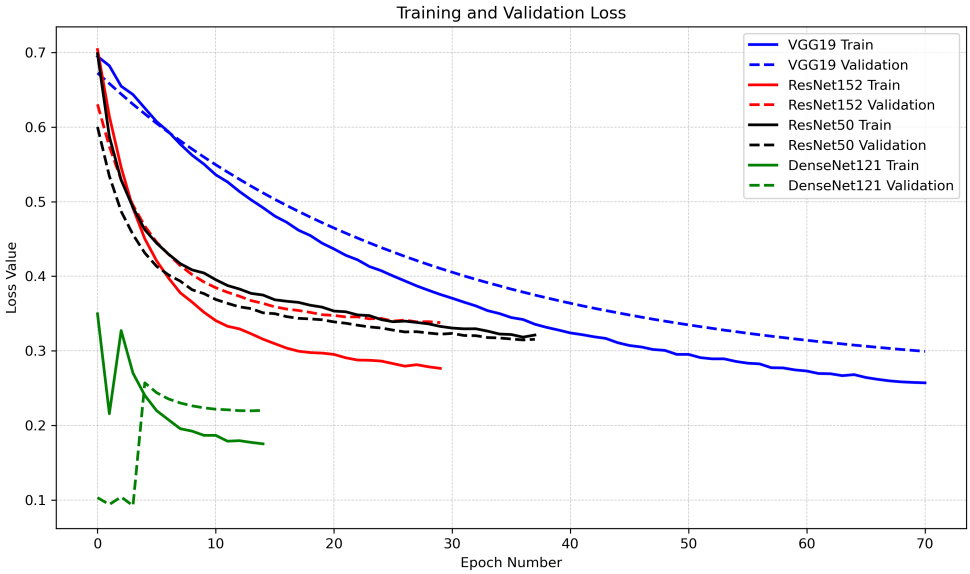
5.1.2 Classification Networks

We used CNN-1 and CNN-2 to classify US frames into "good" and "bad" frames, defined as frames with and without visible desired features, respectively. "transverse frame classification" (Section 4.5.2.3) and "coronal frame classification" (Section 4.5.3.3) datasets were used for training CNN-1 and CNN-2, respectively. We trained four CNN models, DenseNet121, ResNet152, ResNet50, and VGG19, using the binary cross-entropy (BCE) loss function and the Adam optimizer with a learning rate of 0.0001. The training process was halted after 10 epochs if no improvement in validation loss was observed. Classification accuracy was chosen as the metric to evaluate the models' performance. For both transverse and coronal frame classification, we have depicted the training and validation losses and accuracy values in Figures 5.7 and 5.8, respectively. Additionally, Table 5.2 presents the final validation accuracy values for each CNN model for both transverse and coronal frame classifications.

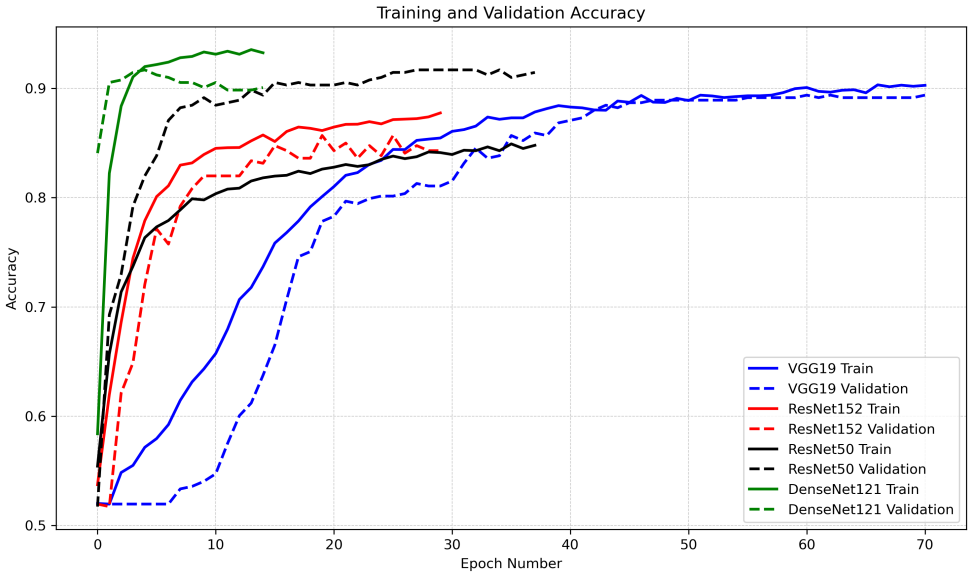
Our findings indicated that ResNet50 achieved highest accuracy for coronal classification. For transverse classification, ResNet50 and DenseNet121 achieved the highest accuracy with the same value. Based on consistency and simpler architecture, we selected ResNet50 as the preferred model for transverse frame classification.

Table 5.2: The mean accuracy values of the validation set after the training of CNN models with various architectures.

	ResNet50	ResNet152	VGG19	DenseNet121
CNN-1	0.92	0.86	0.89	0.92
CNN-2	0.86	0.74	0.83	0.85

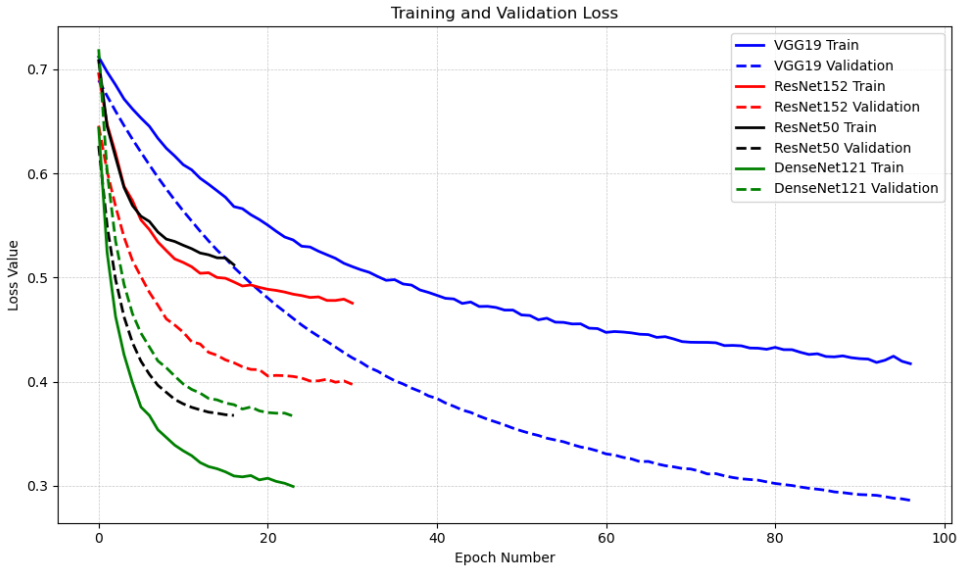


(a)

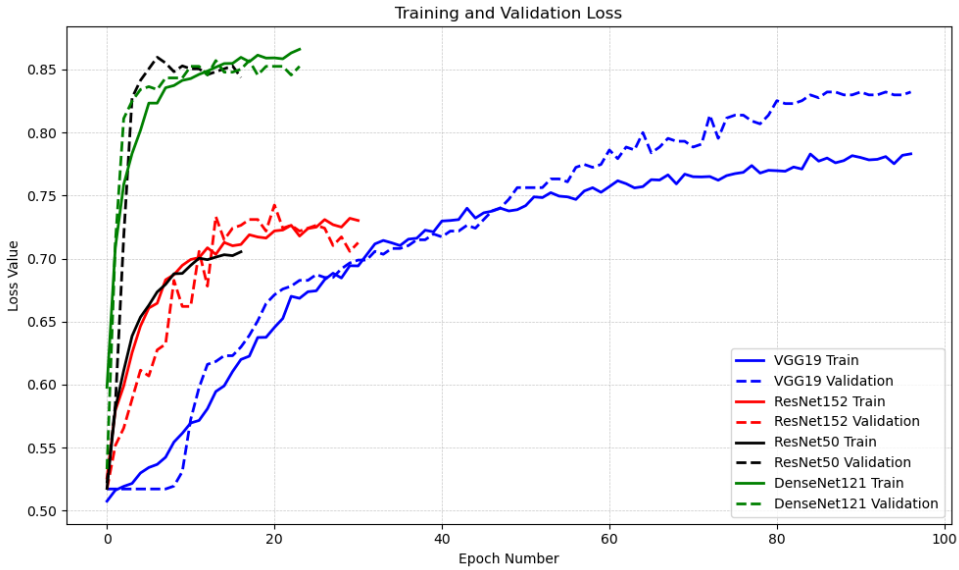


(b)

Figure 5.7: The BCE loss (a) and accuracy (b) values for the training and validation sets during the **CNN-1** training process.



(a)



(b)

Figure 5.8: The BCE loss (a) and accuracy (b) values for the training and validation sets during the **CNN-2** training process.

5.1.3 Comparison of Moving Average and Gaussian Filters for Frame Selection

We used CNN-1 and CNN-2 to score US frames, with our primary objective being to select a continuous range of frames that includes those with the highest probability. To achieve this, we applied two distinct filters, the moving average (MA) and Gaussian filters (detailed in Section 4.6.1), to the scores of individual frames. We tested values of 20, 30, and 40 to determine the optimal window length (N) for the MA filter. Heuristic analysis of frame score functions and selected frames for the scans within the training dataset led to the selection of 30. The value of standard deviation (σ) for the Gaussian filter was also heuristically selected as 1 after examining results with 0.5, 1, and 2.

The first and last frames of the chosen range, determined by each filter, were compared against the first and last frames of the manually selected range by R_1. Our evaluation metric of choice was ICC(2,1). A summary of the comparative results is presented in Table 5.3. The outcomes reveal that, on average, the moving average filter achieved a higher ICC(2,1) value in contrast to the Gaussian filter for determination of both the starting and ending frames of the selected range of coronal and transverse US scans.

Table 5.3: Comparison of ICC(2,1) values for moving average and Gaussian filters vs. the manual selections in coronal and transverse views.

Filter	Coronal		Transverse		Average
	begin	end	begin	end	
MA	0.97	0.96	0.92	0.93	0.94
Gaussian	0.95	0.80	0.94	0.89	0.89

5.2 Comparison of Measurements on Test Set by Human and AI Model in Different Modalities

We tested the final optimized model on the test dataset consisting of 18 hips, with each hip having a coronal and a transverse US scan. In our evaluation, we compared the "B", "A", and MP measurements obtained by the AI model on US (AI-B_{US}, AI-A_{US}, and AI-MP_{US}), with the measurements by R_1 on US (M-B_{US}, M-A_{US}, and M-MP_{US}) and the measurements by R_2 on X-ray as the ground truth (B_{Xray}, A_{Xray}, and MP_{Xray}). The required time for the manual US measurements by R_1 was also obtained as 300s on average.

5.2.1 "B" Measurements

Figures 5.9 to 5.11 display the comparisons of M-B_{US} with B_{Xray}, AI-B_{US} with B_{Xray}, and AI-B_{US} with M-B_{US}, respectively. Overall, the estimated "B" value was lowest with M-B_{US}, followed by AI-B_{US}, and highest with B_{Xray}. This suggested an underestimation of the "B" value when using US images. Additionally, high R-squared values indicated strong correlations within all comparisons.

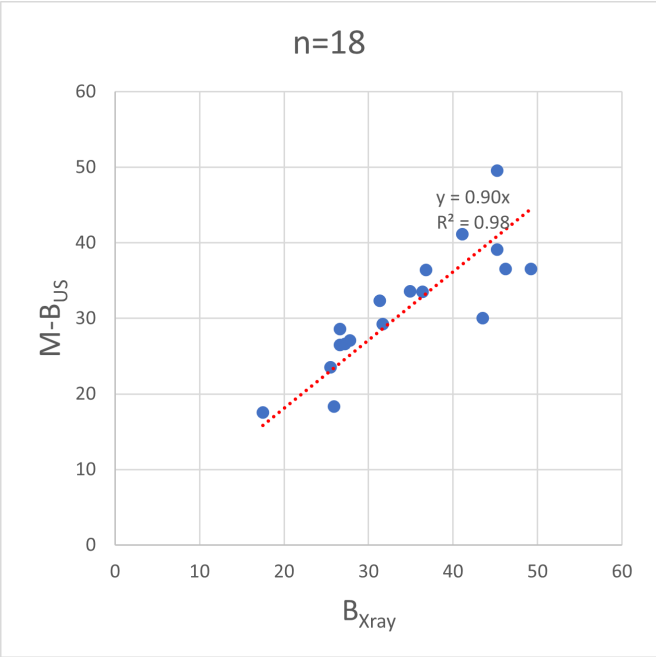


Figure 5.9: **M-B_{US} vs. B_{Xray}** measurements on the test set, with the fitted regression line.

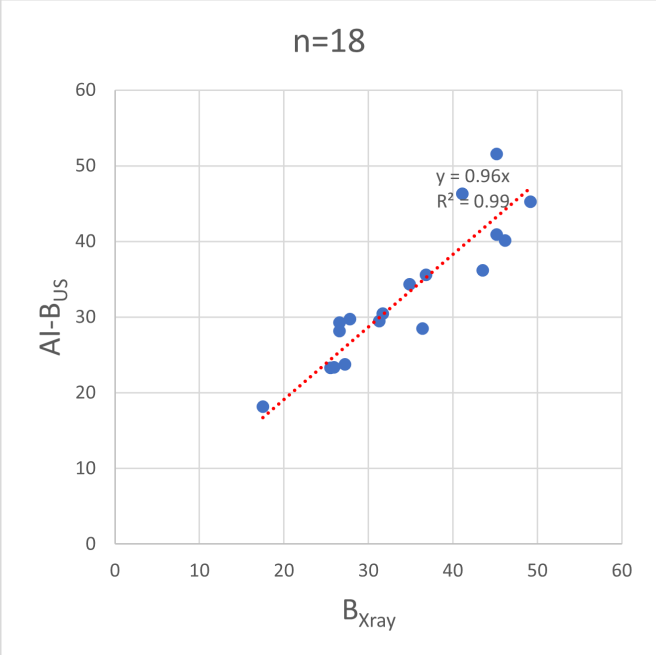


Figure 5.10: **AI-B_{US} vs. B_{Xray}** measurements on the test set, with the fitted regression line.

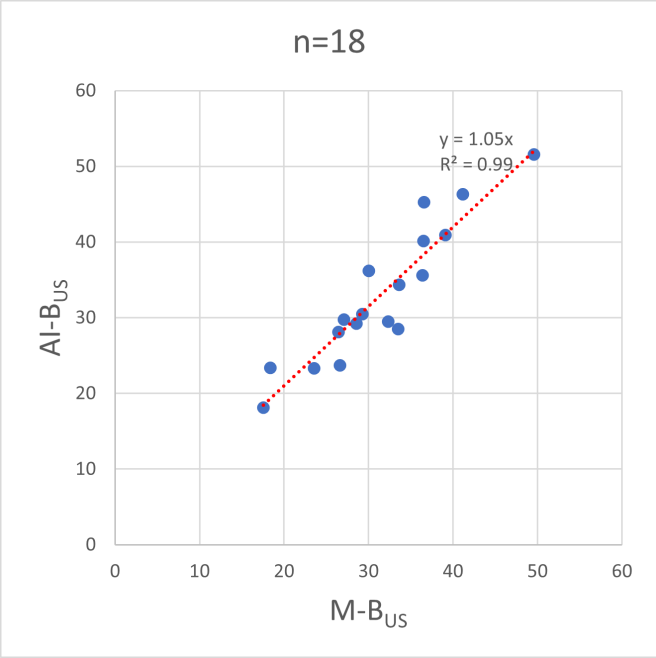


Figure 5.11: **AI-B_{US} vs. M-B_{US}** measurements on the test set, with the fitted regression line.

Table 5.4 summarizes the measured range, average and SD of B_{Xray}, M-B_{US}, and AI-B_{US}. In addition, Table 5.5 provides a full comparison between the mentioned measurements. When Compared to X-ray measurements, the AI model achieved a lower mean absolute error (MAD) than the manual US measurements. Furthermore, AI model demonstrated an excellent level of agreement with B_{Xray} while manual US measurements depicted a good level of reliability. In addition M-B_{US} and AI-B_{US} had the lowest MAD and highest ICC(2,1), demonstrating their strong agreement and alignment.

Table 5.4: The measured range, average and SD of B_{Xray}, M-B_{US}, and AI-B_{US}.

	B_{Xray}	M-B_{US}	AI-B_{US}
Measured range (mm)	17.5 - 51	17.6 - 49.6	18.1 - 51.6
Average (mm) ± SD (mm)	33.5 ± 8.2	31.5 ± 7.7	33.0 ± 8.7

Table 5.5: Pairwise statistical comparisons of AI-B_{US}, M-B_{US}, and B_{Xray} measurements for the test set.

	MAD(mm) ± SD(mm)	ICC(2,1)
AI-B_{US} vs B_{Xray}	3.4mm ± 2.3mm	0.90 (0.76-0.96)
M-B_{US} vs B_{Xray}	3.7mm ± 4.2mm	0.79 (0.47-0.92)
AI-B_{US} vs M-B_{US}	2.9mm ± 2.2mm	0.91 (0.76-0.97)

The Bland-Altman plots for comparisons between B_{Xray}, M-B_{US}, and AI-B_{US} are shown in Figures 5.12 to 5.14. M-B_{US} and AI-B_{US} had the lowest mean difference (MD) of 1.6mm and tightest LoA (-4.9, 8.1) among the three comparisons. Moreover, Comparing to X-ray, the AI model exhibited lower MD (-1.3mm) and a significantly

improved lower LoA bound (-8.7) than manual measurements (-12.3). The upper bounds of LoA for manual (6.5) and AI measurements (6.1) were close.

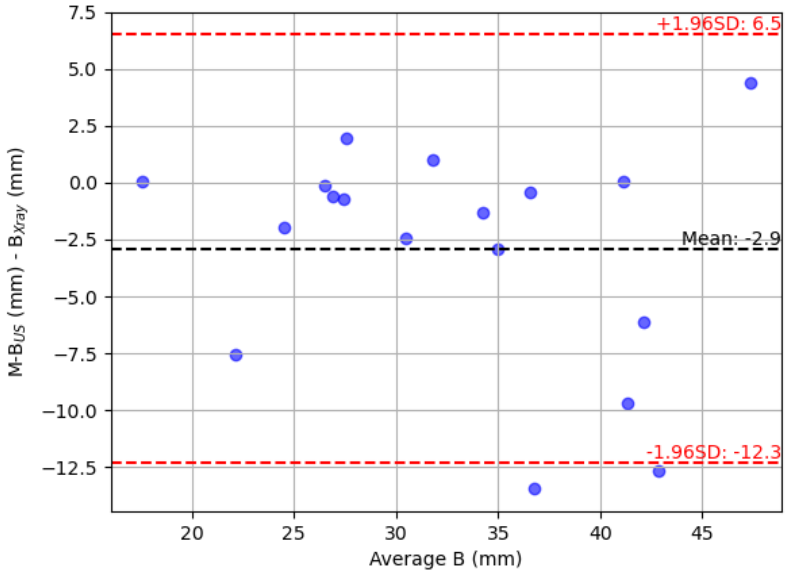


Figure 5.12: Bland-Altman plot illustrating the agreement between **M-B$_{US}$ and B$_{Xray}$** measurements on test set, with the differences plotted against their average.

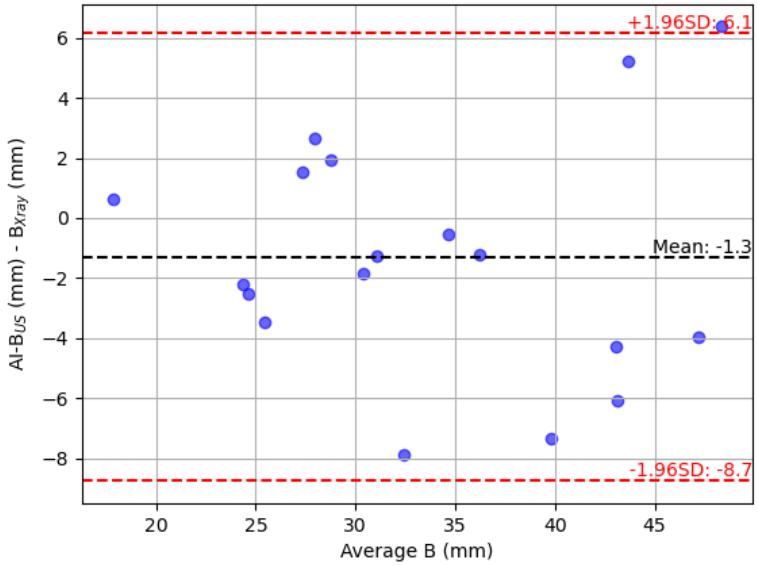


Figure 5.13: Bland-Altman plot illustrating the agreement between **AI-B$_{US}$ and B$_{Xray}$** measurements on test set, with the differences plotted against their average.

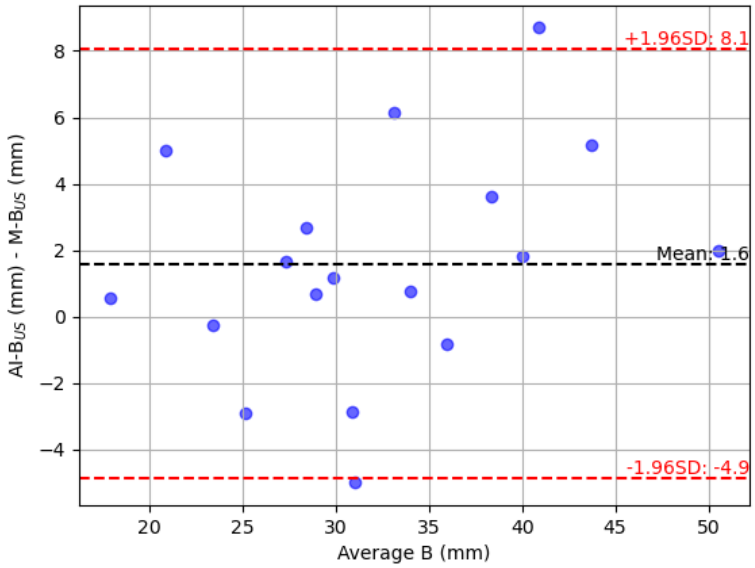


Figure 5.14: Bland-Altman plot illustrating the agreement between **AI-B$_{US}$ and M-B$_{US}$** measurements on test set, with the differences plotted against their average.

To evaluate the performance of the networks involved in "B" measurements individually, the test data was manually labeled by R_1 (as described in Section 4.5.2) and compared with the outputs of the networks. The dice coefficients for UNet-1 and UNet-2 were 0.84 and 0.55, respectively. Additionally, CNN-1 demonstrated an accuracy of 86.72%, with accuracies of 88.67% for "good" frames and 84.54% for "bad" frames. The ICC(2,1) for selection of beginning and ending frames was 0.85 and 0.97.

5.2.2 "A" Measurements

Figures 5.15 to 5.17 compare M-A_{US}, AI-A_{US}, and A_{Xray} measurements. Notably, both M-A_{US} and AI-A_{US} underestimated the "A" value compared to A_{Xray}. Additionally, AI-A_{US} and M-A_{US} measurements demonstrated close alignment, indicated by a fitted line slope of 1.01. High R-squared values suggested strong correlations between all compared measurements.

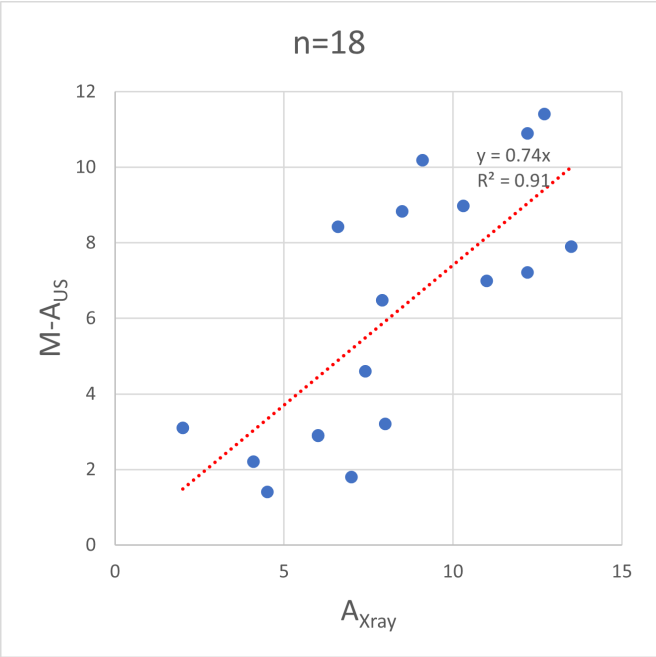


Figure 5.15: **M-A_{US} vs. A_{Xray}** measurements on the test set, with the fitted regression line.

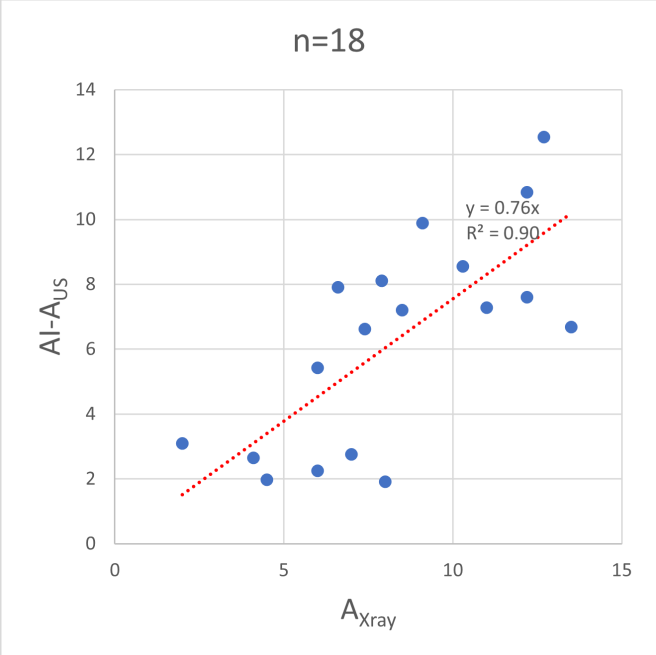


Figure 5.16: **AI-A_{US} vs. A_{Xray}** measurements on the test set, with the fitted regression line.

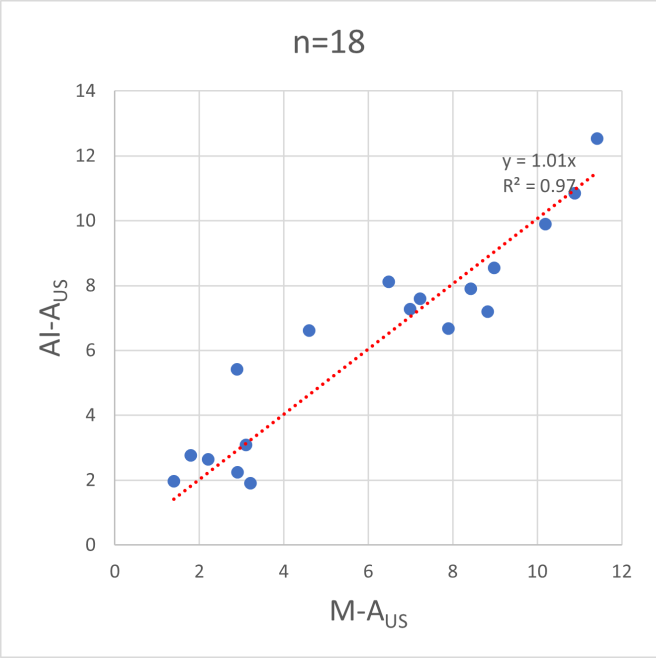


Figure 5.17: **AI-A_{US} vs. M-A_{US}** measurements on the test set, with the fitted regression line.

Table 5.6 summarizes the measured range, average and SD of A_{Xray}, M-A_{US}, and AI-A_{US}. In addition, Table 5.7 provides a full comparison between the mentioned measurements. Referencing X-ray measurements as the baseline, the M-A_{US} had a lower MAD than the AI-A_{US}. Both M-A_{US} and AI-A_{US} exhibited moderate agreement with A_{Xray}. Additionally, M-A_{US} and AI-A_{US} showed excellent mutual agreement and a significantly lower MAD than in comparisons with A_{Xray}.

Table 5.6: The measured range, average and SD of A_{Xray}, M-A_{US}, and AI-A_{US}.

	A_{Xray}	M-A_{US}	AI-A_{US}
Measured range (mm)	2 - 13.5	1.4 - 11.4	1.9 - 12.5
Average (mm) ± SD (mm)	8.3 ± 3.1	6.1 ± 3.3	6.3 ± 3.1

Table 5.7: Pairwise statistical comparisons of AI-A_{US}, M-A_{US}, and A_{Xray} measurements for the test set.

	MAD(mm) ± SD(mm)	ICC(2,1)
AI-A_{US} vs A_{Xray}	2.4 ± 2.0	0.62 (0.09-0.85)
M-A_{US} vs A_{Xray}	2.7 ± 1.6	0.63 (0.04-0.87)
AI-A_{US} vs M-A_{US}	0.9 ± 0.7	0.94 (0.85-0.98)

The Bland-Altman plots (Figures 5.18 to 5.20) illustrate comparisons between A_{Xray}, M-A_{US}, and AI-A_{US}. Compared to X-ray, both manual US and AI measurements showed similar mean differences (MD of -2.2mm and -2.0mm respectively). Manual US demonstrated a slightly tighter LoA range (-6.5mm to 2.1mm) compared to AI (-6.7mm to 2.7mm). Additionally, direct comparison of AI-A_{US} and M-A_{US} showed negligible bias (MD of 0.2mm) and a narrow LoA (-2.0mm, 2.4mm).

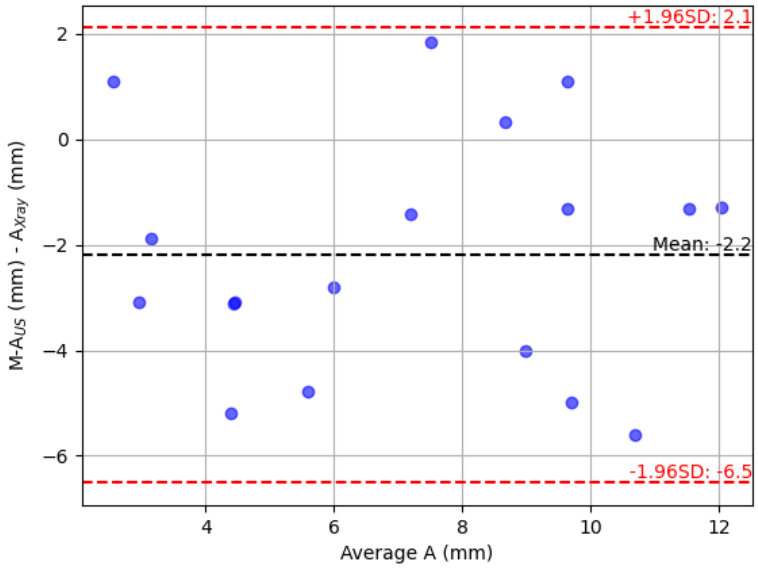


Figure 5.18: Bland-Altman plot illustrating the agreement between **M-A$_{US}$ and A$_{Xray}$** measurements on test set, with the differences plotted against their average.

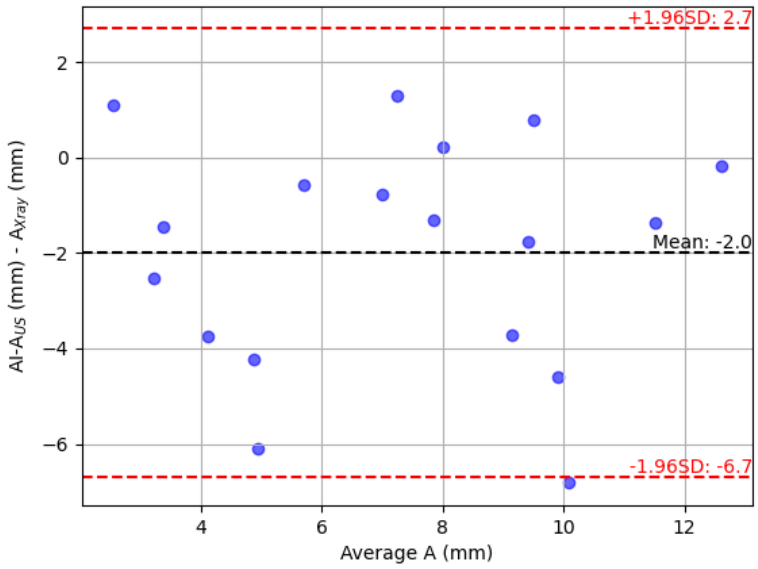


Figure 5.19: Bland-Altman plot illustrating the agreement between **AI-A$_{US}$ and A$_{Xray}$** measurements on test set, with the differences plotted against their average.

To evaluate the performance of the networks involved in "A" measurements individually, the test data was manually labeled by R_1 (as described in Section 4.5.2) and

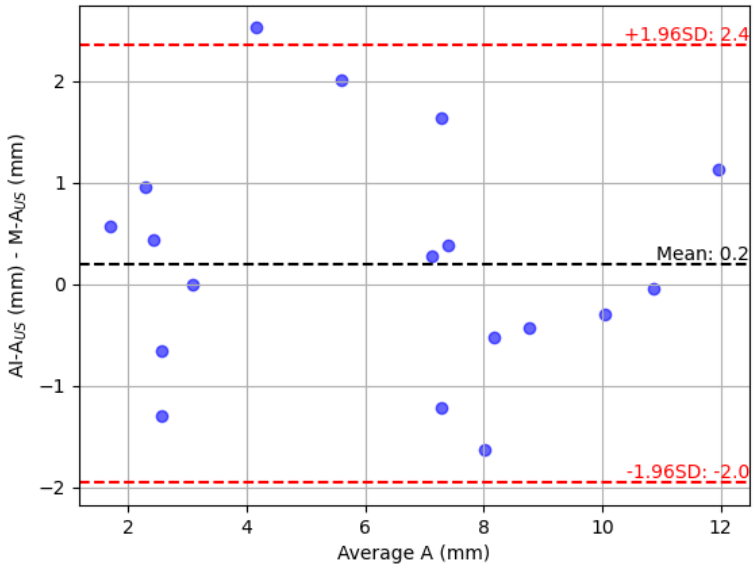


Figure 5.20: Bland-Altman plot illustrating the agreement between **AI-A_{US} and M-A_{US}** measurements on test set, with the differences plotted against their average.

compared with the outputs of the networks. The dice coefficients for UNet-3, UNet-4, UNet-5, and UNet-6 were 0.61, 0.80, 0.41, and 0.25, respectively. Additionally, CNN-2 demonstrated an accuracy of 79.95%, with accuracies of 96.23% for "good" frames and 65.56% for "bad" frames. The ICC(2,1) for selection of beginning and ending frames was 0.95 and 0.94.

5.2.3 MP Measurements

Figures 5.21 to 5.23 compare M-MP_{US}, AI-MP_{US}, and MP_{Xray} measurements. As expected due to underestimation of "A" value, both M-MP_{US} and AI-MP_{US} underestimated the "MP" value compared to MP_{Xray}. Additionally, AI-A_{US} and M-A_{US} measurements demonstrated close alignment, indicated by a fitted line slope of 1.03. High R-squared values suggested strong correlations between all compared measurements.

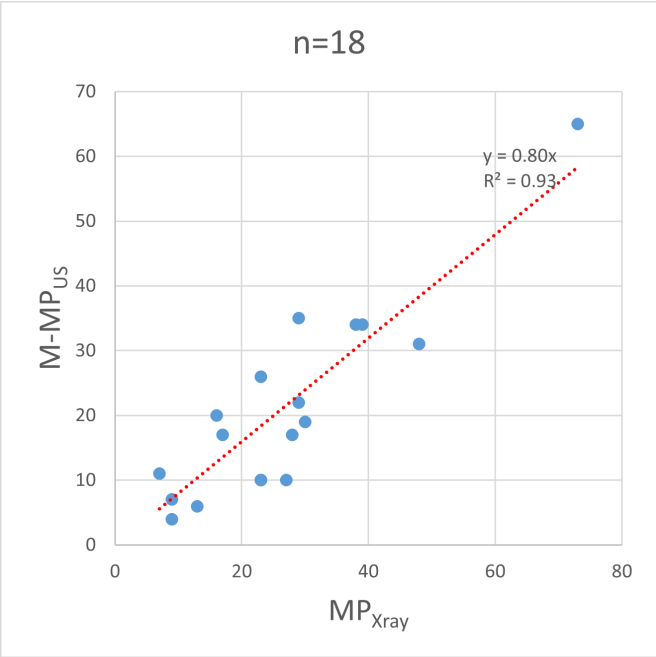


Figure 5.21: **M-MP$_{\mathbf{US}}$ vs. MP$_{\mathbf{Xray}}$** measurements on the test set, with the fitted regression line.

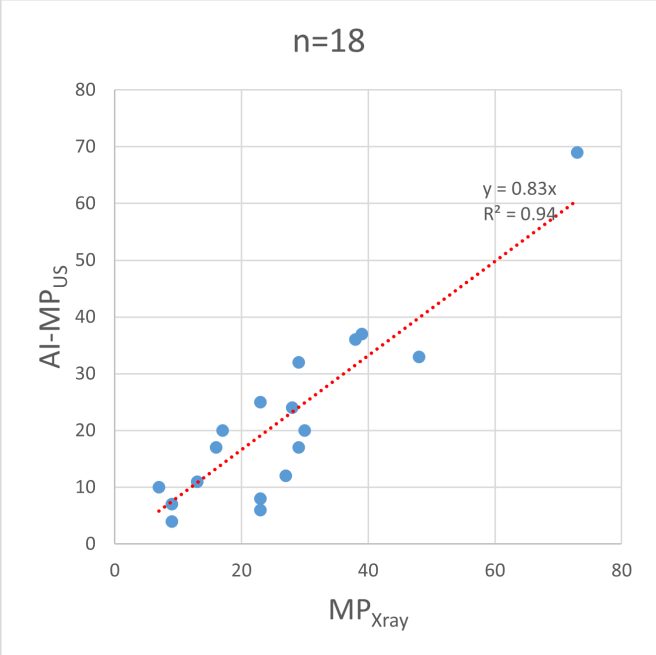


Figure 5.22: **AI-MP$_{\mathbf{US}}$ vs. MP$_{\mathbf{Xray}}$** measurements on the test set, with the fitted regression line.

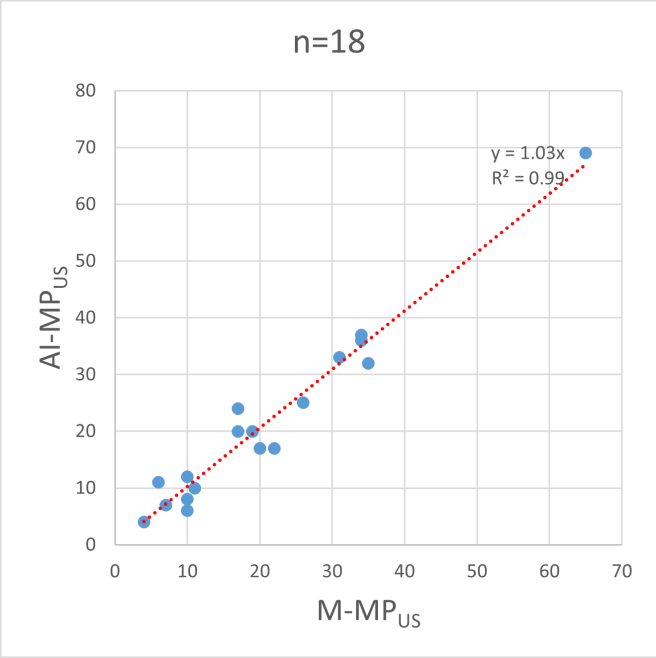


Figure 5.23: **AI-MP$_{US}$ vs. M-MP$_{US}$** measurements on the test set, with the fitted regression line.

Table 5.8 summarizes the measured range, average and SD of MP_{Xray}, M-MP_{US}, and AI-MP_{US}. In addition, Table 5.9 provides a full comparison between the mentioned measurements. Referencing X-ray measurements as the baseline, the M-MP_{US} achieved a lower MAD than the AI-MP_{US}. Both M-MP_{US} and AI-MP_{US} exhibited good agreement with MP_{Xray}. Additionally, M-MP_{US} and AI-MP_{US} showed excellent mutual agreement and a significantly lower MAD than in comparisons with MP_{Xray}. Furthermore, the AI model achieved a 72% (13 out of 18) clinical acceptance (CA) rate, exceeding the 67% CA rate of manual US measurements. CA rate is defined as the percentage of measurements with less 10% error.

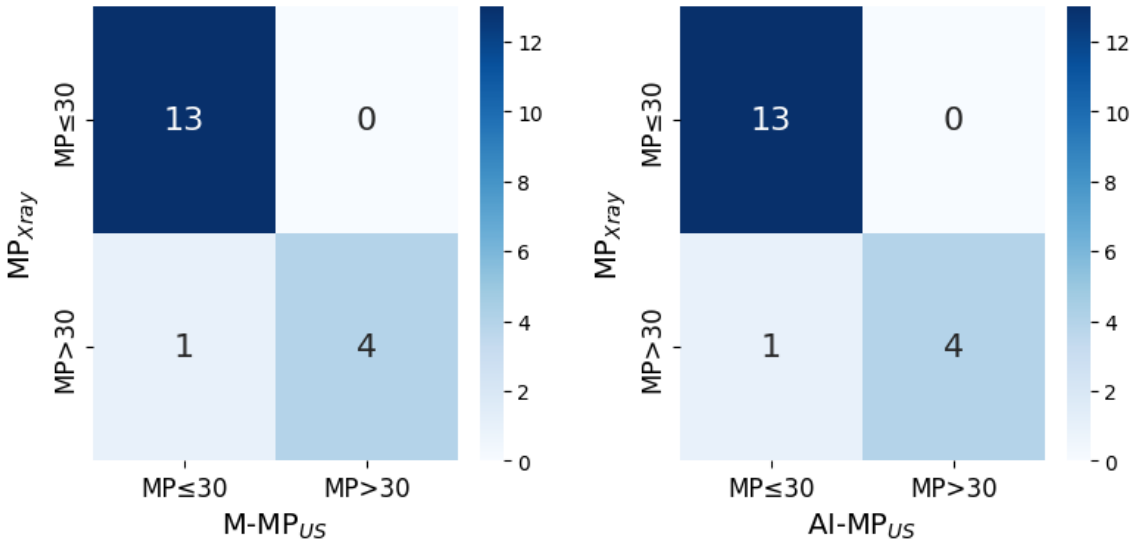
Table 5.8: The measured range, average and SD of MP_{Xray}, M-MP_{US}, and AI-MP_{US}.

	MP_{Xray}	M-MP_{US}	AI-MP_{US}
Measured range (%)	7 - 73	4 - 65	4 - 69
Average (%) ± SD (%)	26.7 ± 15.5	21 ± 14.6	21.6 ± 15.4

Table 5.9: Pairwise statistical comparisons of AI-MP_{US}, M-MP_{US}, and MP_{Xray} measurements for the test set.

	MAD(%) ± SD(%)	CA rate (%)	ICC(2,1)
AI-MP_{US} vs MP_{Xray}	6.5 ± 5.5	72	0.86 (0.52-0.95)
M-MP_{US} vs MP_{Xray}	7.6 ± 4.9	67	0.84 (0.43-0.95)
AI-MP_{US} vs M-MP_{US}	2.7 ± 1.8	100	0.98 (0.94-0.99)

The confusion matrices comparing AI-MP_{US} and M-MP_{US} against MP_{Xray} for hip displacement diagnosis are depicted in Figure 5.24. Hip displacement was defined as MP>30, a critical threshold in hip surveillance programs. The confusion matrices for



(a) (b)

Figure 5.24: (a) Confusion matrix depicting MP_{Xray} vs. M-MP_{US} classifications. (b) Confusion matrix depicting MP_{Xray} versus AI-MP_{US} classifications.

both M-MP_{US} and AI-MP_{US} demonstrate identical results, with an overall accuracy of 94.4%, specificity of 93%, and sensitivity of 100%.

The Bland-Altman plots (Figures 5.25 to 5.27) illustrate the comparisons between MP_{Xray}, M-MP_{US}, and AI-MP_{US}. When compared to X-ray, manual US measurements demonstrated a lower MD (-5.2%) and a narrower LoA (-18.5% to 8.1%) than AI measurements (-5.7% MD, -19.4% to 8.0% LoA). Furthermore, direct comparison between AI-MP_{US} and M-MP_{US} revealed a negligible bias (MD of 0.6%) and a narrow LoA (-5.7% to 6.9%).

5.2.4 Measurement Time

Using an NVIDIA Tesla V100 16GB GPU, an Intel Xeon Gold 6138 dual processor, and 64GB of RAM, the average measurement time ± SD for coronal-view US frames was 0.16 ± 0.04 seconds, and for transverse-view US frames, it was 0.12 ± 0.04 seconds. The average number of frames for coronal and transverse scans was 360 and 400, respectively. Consequently, on average, the AI model required 57.6 seconds for a coronal scan and 48 seconds for a transverse scan, and a total of 105.6 seconds to

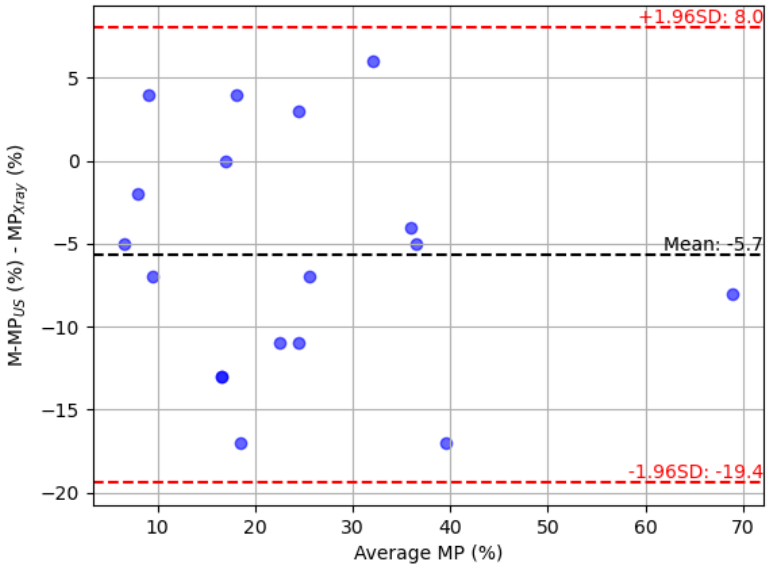


Figure 5.25: Bland-Altman plot illustrating the agreement between **M-MP$_{US}$ and MP$_{Xray}$** measurements on test set, with the differences plotted against their average.

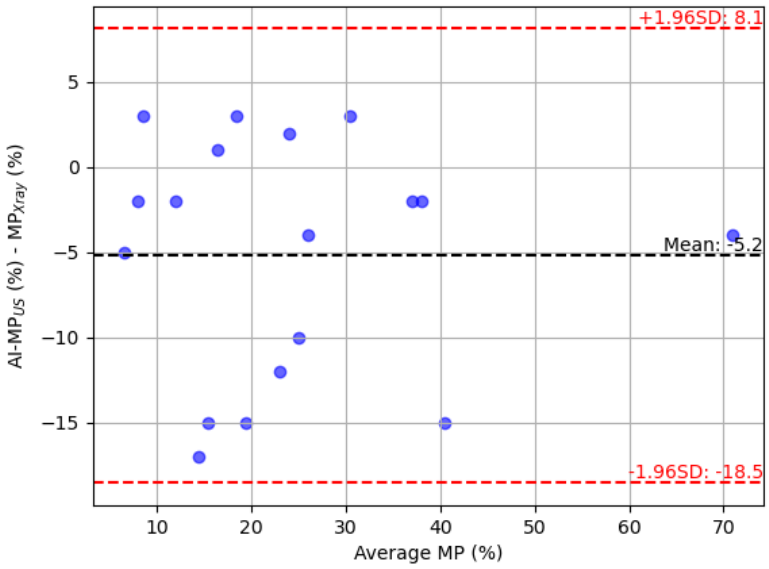


Figure 5.26: Bland-Altman plot illustrating the agreement between **AI-MP$_{US}$ and MP$_{Xray}$** measurements on test set, with the differences plotted against their average.

report the MP measurement for a single hip.

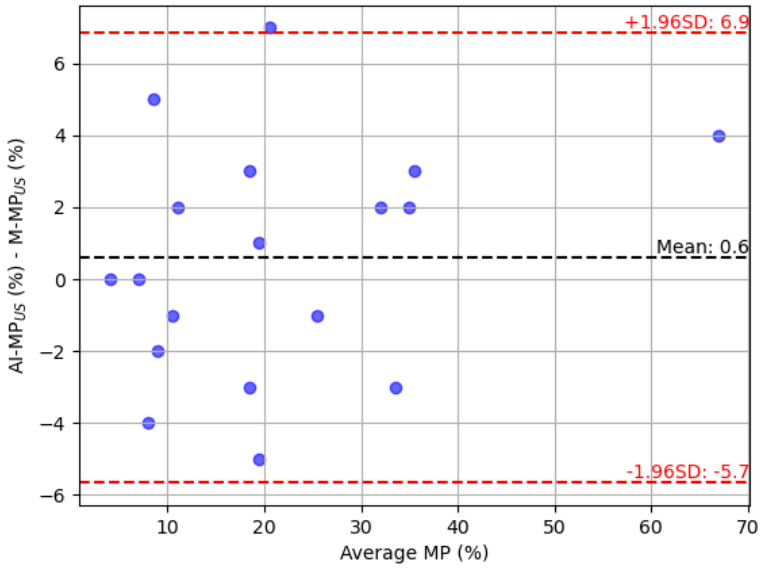


Figure 5.27: Bland-Altman plot illustrating the agreement between **AI-MP$_{US}$ and M-MP$_{US}$** measurements on test set, with the differences plotted against their average.

5.3 Analysis and Discussion on Measurements on Test Set

5.3.1 "B" Measurements

By evaluating results on "B" measurements using various metrics, it can be concluded that the model has achieved both accurate and reliable measurements. The excellent agreement between the AI-B$_{US}$ and M-B$_{US}$ signifies that the model is well-trained and capable of reproducing measurements that closely match manual ones. Furthermore, the AI measurements demonstrated excellent agreement (ICC(2,1)>0.9) with B$_{Xray}$ as well, surpassing the performance of M-B$_{US}$. This indicates that the AI model has successfully learned a consistent and reliable method of measurement, effectively reducing potential human errors.

Overall, the performance of the networks involved in "B" measurements on test set (UNet-1 and UNet-2) was close to their performance on validation set (Sections 5.1.1.1 and 5.1.1.2), indicating effective training and generalization capabilities. Further-

more, a high level of agreement was observed in the frame selection on the test set between R_1 and the developed AI model, as quantified by ICC(2,1), which indicates the effectiveness of the developed automatic frame selection method.

5.3.2 "A" measurements

The AI model demonstrated excellent agreement with manual US measurements, signifying proficient model development and training. However, both the AI measurements and manual US measurements demonstrated moderate agreement with X-ray measurements. A thorough examination of outcomes across various scans indicated that a major source of error is the segmentation of the acetabulum. This challenge primarily stems from the significant variation in acetabulum shape among patients, which complicates the model's learning process for accurate acetabulum segmentation, particularly with limited data. Moreover, in some instances, the quality of the scans was poor, making it difficult to identify the acetabulum. In summary, while measuring in the US framework is inherently challenging due to the need for precise identification of edge points—which is difficult in US images due to the absence of sharp edges and gradual changes in pixel intensity—improvements can be achieved through acquiring more data and enhancing scan quality.

Overall, the performance of the networks involved in "A" measurements (UNet-3, UNet-4, UNet-5, and UNet-6) on test set was close to their performance on validation set (Sections 5.1.1.1 and 5.1.1.2), indicating good training and generalization. Also, there was excellent level of agreement in frame selection in test set between the R_1 and the developed AI model, quantified by ICC(2,1). However, the low dice coefficient for acetabulum segmentation indicates challenges in accurately defining this structure. These difficulties likely propagate to subsequent stages of the networks, particularly the classification of "bad" coronal frames.

5.3.3 MP measurements

The MAD of the AI measurements compared to X-ray measurements was 6.5% ± 5.5%. This is slightly higher than the reported MAD of 4.5%-4.9% when comparing manual and AI-based measurements both taken on X-rays [55]. Considering the fact that we used a different modality (US), higher MAD is expected. To investigate the extent to which the differences in modalities affect the measurements, we sought the expertise of R_2, an experienced rater, to perform the measurements on US frames. The assumption was that an experienced rater could achieve the best possible results on US. R_2 measurements on US and X-ray resulted in an MAD of 6.4% ± 5.2%, showing only a difference of 0.1% with AI measurements.

The mean variation between the AI measurement on US and X-ray measurements -5.2% compared to the -5.7% mean variation of manual measurements on US vs X-ray. Reported mean differences between two raters on X-ray range from -1.2% to 3.7% [32, 33].

Moreover, the lower and upper bounds of LoA for AI measurements vs X-ray measurements were -18.5% to 8.1%, compared to -19.4% and 8.0% for manual measurements on US vs X-ray measurements. The clinical acceptance rate for AI measurements was 72% compared to 67% for R_1 measurements on US. Notably, the reported clinical acceptance rate on radiology is 90% [55].

From the literature, the reported sensitivity and specificity values are 88% and 89%, respectively, for classifying MP>25% as abnormal [42]. Another study found sensitivity and specificity values of 89% and 100% when classifying a hip with MP>33% as displaced [39]. Therefore, the developed AI model achieved a comparable sensitivity (100%) and higher specificity (93%). Nevertheless, further research involving a larger sample size is necessary to validate these findings.

Based on these findings, it can be concluded that the performance of AI can be slightly improved with expert labeling and by increasing the diversity in datasets,

especially by focusing on "A" measurements. Although some variations will still exist in hip assessment by radiology and US due to using a different modality, US can be utilized as a primary method for hip examinations, reducing the need for many unnecessary X-ray imagings.

5.3.4 Measurement Time

The developed method demonstrated a substantial efficiency gain, with a measurement time of 105.6 seconds compared to the manual US measurement's estimated time of 300 seconds. This represents a 65% reduction in measurement time. Code analysis with checkpoints revealed that approximately 33% of the runtime is dedicated to data handling, while 67% is used for GPU-based computations. This breakdown suggests potential avenues for optimization, such as streamlining data handling processes.

5.4 Chapter Summary

We optimized the architecture of deep networks used in the proposed method by trying four common architectures in the literature according to the corresponding evaluation metrics of each task. In addition, we compared Gaussian and moving average filters and selected the filter with better performance to be used in the proposed method. Furthermore, we obtained the results of the test set which showed moderate performance in "A" measurement, excellent performance in "B" measurements, and good performance in "MP" measurements. We explored each variable measurement by different metrics, and compared the results with literature. The results showed with enough training data, the proposed method has the potential to show a performance close to a human rater and facilitate replacing X-ray with US imaging in hip surveillance programs. Finally, we demonstrated the method's efficiency by showing a 65% reduction in measurement time when using the automatic process versus manual measurement.

Chapter 6

Conclusions and Recommendations

6.1 Conclusion

Migration percentage (MP), defined as the ratio of the distance between the lateral border of the acatabulum and the femur head (A) to the total width of the femur head (B), is the key parameter in hip displacement assessments and has to be measured regularly. Currently, anterposterior radiography is used for MP measurements, exposing children to ionizing radiations. Recently, a method has been proposed to measure MP from ultrasound (US) images, which is very time-consuming, and user-dependent. In this thesis, we proposed a fully automatic AI-based method to measure "A", "B", and, consequently, "MP". The AI method was developed from the 0 to 100, including the manual labeling, training the models, and evaluation of the model. AI measurements were compared with manual ultrasound measurements provided by the author and X-ray measurements provided by an experienced rater for evaluation. The model was able to deliver the final measurements in 105.6 seconds on average, significantly improving the manual measurement time.

For "B" measurements, the AI model showed excellent agreement with both ultrasound (ICC(2,1) = 0.91) and X-ray measurements (ICC(2,1) = 0.90), exceeding the agreement between manual US and X-ray measurements (ICC(2,1) = 0.79). This likely stems from the richer dataset used for "B" measurements, likely containing more diverse examples. This allowed the model to learn highly generalizable measurements,

potentially surpassing the consistency of less experienced manual assessments.

For "A" measurements, the developed method demonstrated moderate agreement with X-ray measurements (ICC(2,1) = 0.58) and excellent agreement with ultrasound measurements (ICC(2,1) = 0.94). Although the discrepancy with X-ray results is expected due to the inherent differences between imaging modalities, further analysis revealed that a key source of error lies in acetabulum segmentation. This is likely due to the high anatomical variability of the acetabulum and limitations within the training dataset.

For MP measurements, the AI model showed good agreement with X-ray measurements (ICC(2,1) = 0.85). Considering MP=30 as a threshold for hip displacement classification, the model showed high accuracy (94.4%), sensitivity (100%) and specificity (93%). Although the confirmation of the method and results require further evaluation on larger datasets, this preliminary study sets the step for replacing X-ray with US in hip assessments.

6.2 Contributions

Getting back to the objectives defined in Section 1.3, the contributions of this work are:

1. A comprehensive dataset containing approximately 1450 coronal and 1450 transverse US frames was prepared, with corresponding labels for segmentation, edge detection, and landmark detection tasks. Datasets containing approximately 3000 coronal and 3000 transverse frames were labeled as good/bad for frame classification. To the best of our knowledge, this is the first dataset of its kind, paving the way for significant advancements in the field.

2. An automated method for measuring lateral head distance (variable "A") was developed, achieving an average measurement time of approximately 58 seconds, significantly reducing the manual measurement time.

3. An automated method for measuring the total width of the femur head (variable "B") was developed, achieving an average measurement time of approximately 48 seconds, significantly reducing the manual measurement time.

4. The model's performance, using the developed "A" and "B" measurement methods, was evaluated on a test dataset. Results demonstrate good reliability, acceptable accuracy and clinical acceptance rate, and excellent sensitivity and specificity.

6.3 Future Recommendations

- This study's main limitation was the scarcity of data. Collecting a larger dataset with greater diversity in different factors such as MP value, acetabulum and femur head shape, image quality, etc., could significantly improve results – especially in "A" measurements – and lead to more conclusive findings.

- Image quality plays a vital role in measurement accuracy. Therefore, training operators to identify low-quality US scans and repeat scans when necessary can significantly improve results. Additionally, automated image quality rating systems can provide valuable support in this process.

- There is room for research and exploration in unsupervised learning, especially in feature representation learning, which is becoming more popular in medical image processing, and eliminate the need for manual labeling.

- Investing in software optimization can significantly improve calculation speed. This may involve upgrading code, streamlining how data is handled, and using tools to pinpoint areas for improvement.

Bibliography

[1] P. Baxter *et al.*, "The definition and classification of cerebral palsy," *Dev Med Child Neurol*, vol. 49, no. s109, pp. 1–44, 2007.

[2] M. Oskoui, F. Coutinho, J. Dykeman, N. Jetté, and T. Pringsheim, "An update on the prevalence of cerebral palsy: A systematic review and meta-analysis," *Developmental Medicine & Child Neurology*, vol. 55, no. 6, pp. 509–519, 2013.

[3] B. Soo *et al.*, "Hip displacement in cerebral palsy," *JBJS*, vol. 88, no. 1, pp. 121–129, 2006.

[4] J. Reimers, "The stability of the hip in children: A radiological study of the results of muscle surgery in cerebral palsy," *Acta Orthopaedica Scandinavica*, vol. 51, no. sup184, pp. 1–100, 1980.

[5] N. Herregods, F. M. Vanhoenacker, J. L. Jaremko, and L. Jans, "Update on pediatric hip imaging," in *Seminars in musculoskeletal radiology*, Thieme Medical Publishers, vol. 21, 2017, pp. 561–581.

[6] M. R. Bagg, J. Farber, and F. Miller, "Long-term follow-up of hip subluxation in cerebral palsy patients.," *Journal of pediatric orthopedics*, vol. 13, no. 1, pp. 32–36, 1993.

[7] N. H. Jung *et al.*, "Does hip displacement influence health-related quality of life in children with cerebral palsy?" *Developmental Neurorehabilitation*, vol. 17, no. 6, pp. 420–425, 2014.

[8] G. Hägglund, S. Andersson, H. Düppe, H. Lauge-Pedersen, E. Nordmark, and L. Westbom, "Prevention of dislocation of the hip in children with cerebral palsy: The first ten years of a population-based prevention programme," *The Journal of Bone and Joint Surgery. British volume*, vol. 87, no. 1, pp. 95–101, 2005.

[9] P. Sloper and J. Statham, *The national service framework for children, young people and maternity services: Developing the evidence base*, 2004.

[10] R. N. Boyd *et al.*, "Australian cerebral palsy child study: Protocol of a prospective population based study of motor and brain development of preschool aged children with cerebral palsy," *BMC neurology*, vol. 13, no. 1, pp. 1–12, 2013.

[11] T.-T. Pham, L. H. Le, T.-G. La, J. Andersen, and E. H. Lou, "Ultrasound imaging of hip displacement in children with cerebral palsy," *Ultrasound in Medicine & Biology*, 2023.

[12] P. Rosenbaum *et al.*, "A report: The definition and classification of cerebral palsy april 2006," *Dev Med Child Neurol Suppl*, vol. 109, no. suppl 109, pp. 8–14, 2007.

[13] J Robin, H. K. Graham, P Selber, F. Dobson, K Smith, and R. Baker, "Proximal femoral geometry in cerebral palsy: A population-based cross-sectional study," *The Journal of bone and joint surgery. British volume*, vol. 90, no. 10, pp. 1372–1379, 2008.

[14] M. W. Jones, E. Morgan, J. E. Shelton, and C. Thorogood, "Cerebral palsy: Introduction and diagnosis (part i)," *Journal of Pediatric Health Care*, vol. 21, no. 3, pp. 146–152, 2007.

[15] E. Wood and P. Rosenbaum, "The gross motor function classification system for cerebral palsy: A study of reliability and stability over time," *Developmental Medicine & Child Neurology*, vol. 42, no. 5, pp. 292–296, 2000.

[16] M Wynter, N Gibson, M Kentish, S. Love, P. Thomason, and H Kerr Graham, "The consensus statement on hip surveillance for children with cerebral palsy: Australian standards of care," *Journal of pediatric rehabilitation medicine*, vol. 4, no. 3, pp. 183–195, 2011.

[17] T. Terjesen, "Development of the hip joints in unoperated children with cerebral palsy: A radiographic study of 76 patients," *Acta orthopaedica*, vol. 77, no. 1, pp. 125–131, 2006.

[18] B. J. Shore *et al.*, "Reliability of radiographic assessments of the hip in cerebral palsy," *Journal of Pediatric Orthopaedics*, vol. 39, no. 7, e536–e541, 2019.

[19] P. Larnert, O. Risto, G. Hägglund, and P. Wagner, "Hip displacement in relation to age and gross motor function in children with cerebral palsy," *Journal of children's orthopaedics*, vol. 8, pp. 129–134, 2014.

[20] J. Parrott *et al.*, "Hip displacement in spastic cerebral palsy: Repeatability of radiologic measurement," *Journal of Pediatric Orthopaedics*, vol. 22, no. 5, pp. 660–667, 2002.

[21] C. K. Boese *et al.*, "The femoral neck-shaft angle on plain radiographs: A systematic review," *Skeletal radiology*, vol. 45, pp. 19–28, 2016.

[22] D. R. Cooperman, E. Bartucci, E. Dietrick, and E. A. Millar, "Hip dislocation in spastic cerebral palsy: Long-term consequences," *Journal of Pediatric Orthopaedics*, vol. 7, no. 3, pp. 268–276, 1987.

[23] D. Scrutton and G. Baird, "Surveillance measures of the hips of children with bilateral cerebral palsy," *Archives of disease in childhood*, vol. 76, no. 4, pp. 381–384, 1997.

[24] J. Robin *et al.*, "A classification system for hip disease in cerebral palsy," *Developmental Medicine & Child Neurology*, vol. 51, no. 3, pp. 183–192, 2009.

[25] J Vidal, P Deguillaume, and M Vidal, "The anatomy of the dysplastic hip in cerebral palsy related to prognosis and treatment," *International orthopaedics*, vol. 9, pp. 105–110, 1985.

[26] G. Gordon and D. Simkiss, "A systematic review of the evidence for hip surveillance in children with cerebral palsy," *The Journal of Bone and Joint Surgery. British Volume*, vol. 88, no. 11, pp. 1492–1496, 2006.

[27] F Dobson, R. Boyd, J Parrott, G. Nattrass, and H. Graham, "Hip surveillance in children with cerebral palsy: Impact on the surgical management of spastic hip disease," *The Journal of bone and joint surgery. British volume*, vol. 84, no. 5, pp. 720–726, 2002.

[28] M Onimus, G Allamel, P Manzone, and J. Laurain, "Prevention of hip dislocation in cerebral palsy by early psoas and adductors tenotomies.," *Journal of pediatric orthopedics*, vol. 11, no. 4, pp. 432–435, 1991.

[29] S Faraj, W. Atherton, and N. Stott, "Inter-and intra-measurer error in the measurement of reimers' hip migration percentage," *The Journal of Bone and Joint Surgery. British volume*, vol. 86, no. 3, pp. 434–437, 2004.

[30] E. Segev *et al.*, "Intra-and interobserver reliability analysis of digital radiographic measurements for pediatric orthopedic parameters using a novel pacs integrated computer software program," *Journal of children's orthopaedics*, vol. 4, no. 4, pp. 331–341, 2010.

[31] S. M. Kim, E. G. Sim, S. G. Lim, and E. S. Park, "Reliability of hip migration index in children with cerebral palsy: The classic and modified methods," *Annals of rehabilitation medicine*, vol. 36, no. 1, pp. 33–38, 2012.

[32] L. Cliffe, D. Sharkey, G. Charlesworth, J. Minford, S. Elliott, and R. E. Morton, "Correct positioning for hip radiographs allows reliable measurement of hip displacement in cerebral palsy," *Developmental Medicine & Child Neurology*, vol. 53, no. 6, pp. 549–552, 2011.

[33] T. Terjesen and R. B. Gunderson, "Reliability of radiographic parameters in adults with hip dysplasia," *Skeletal radiology*, vol. 41, pp. 811–816, 2012.

[34] G. Hägglund, H. Lauge-Pedersen, and M. Persson, "Radiographic threshold values for hip screening in cerebral palsy," *Journal of children's orthopaedics*, vol. 1, no. 1, pp. 43–47, 2007.

[35] J. Berger-Groch, N. M. Jandl, A. Strahl, U. Bechler, F. T. Beil, and M. H. Stuecker, "Ultrasound as a diagnostic tool for femoral head containment disorders in children between one and 12 years of age," *Journal of Children's Orthopaedics*, vol. 15, no. 5, pp. 496–502, 2021.

[36] R Graf, "Classification of hip joint dysplasia by means of sonography," *Archives of orthopaedic and traumatic surgery*, vol. 102, pp. 248–255, 1984.

[37] T. Terjesen, T. Ø. Runden, and Å. Tangerud, "Ultrasonography and radiography of the hip in infants," *Acta orthopaedica Scandinavica*, vol. 60, no. 6, pp. 651–660, 1989.

[38] N. Clarke, H. T. Harcke, P. Mchugh, M. S. Lee, P. F. Borns, and G. D. MacEwen, "Real-time ultrasound in the diagnosis of congenital dislocation and dysplasia of the hip," *The Journal of Bone and Joint Surgery. British volume*, vol. 67, no. 3, pp. 406–412, 1985.

[39] T. TERJESEN, T. Ø. RUNDEN, and H. M. JOHNSEN, "Ultrasound in the diagnosis of congenital dysplasia and dislocation of the hip joints in children older than two years," *Clinical Orthopaedics and Related Research®*, vol. 262, pp. 159–169, 1991.

[40] I. Šmigovec, T. Đapić, and V. Trkulja, "Ultrasound screening for decentered hips in children with severe cerebral palsy: A preliminary evaluation," *Pediatric radiology*, vol. 44, pp. 1101–1109, 2014.

[41] A. Tegnander and T. Terjesen, "Ultrasound measurements in hips of children above 2 years of age: Normal variations in 232 hips," *Acta Orthopaedica Scandinavica*, vol. 66, no. 3, pp. 229–233, 1995.

[42] A. Tegnander and T Terjesen, "Reliability of ultrasonography in the follow-up of hip dysplasia in children above 2 years of age," *Acta Radiologica*, vol. 40, no. 6, pp. 619–624, 1999.

[43] R. H. Kay *et al.*, "3d ultrasound to quantify lateral hip displacement in children with cerebral palsy: A validation study," *Developmental Medicine & Child Neurology*, vol. 62, no. 12, pp. 1389–1395, 2020.

[44] T.-T. Pham, T.-G. La, L. H. Le, J. Andersen, and E. Lou, "2d ultrasound validation to assess the accuracy of hip displacement measurement: A phantom study," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, 2021, pp. 3173–3176.

[45] N. Chernov, *Circular and linear regression: Fitting circles and lines by least squares*. CRC Press, 2010.

[46] D. Golan, Y. Donner, C. Mansi, J. Jaremko, M. Ramachandran, and CUDL, "Fully automating graf's method for ddh diagnosis using deep convolutional neural networks," in *Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1*, Springer, 2016, pp. 130–141.

[47] A. R. Hareendranathan *et al.*, "Toward automatic diagnosis of hip dysplasia from 2d ultrasound," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, IEEE, 2017, pp. 982–985.

[48] Z. Zhang, M. Tang, D. Cobzas, D. Zonoobi, M. Jagersand, and J. L. Jaremko, "End-to-end detection-segmentation network with roi convolution," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, IEEE, 2018, pp. 1509–1512.

[49] C. G. Chee *et al.*, "Performance of a deep learning algorithm in detecting osteonecrosis of the femoral head on digital radiography: A comparison with assessments by radiologists," *American Journal of Roentgenology*, vol. 213, no. 1, pp. 155–162, 2019.

[50] O. Paserin, K. Mulpuri, A. Cooper, A. J. Hodgson, and R. Garbi, "Real time rnn based 3d ultrasound scan adequacy for developmental dysplasia of the hip," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, Springer, 2018, pp. 365–373.

[51] Q. Li *et al.*, "Auxiliary diagnosis of developmental dysplasia of the hip by automated detection of sharp's angle on standardized anteroposterior pelvic radiographs," *Medicine*, vol. 98, no. 52, 2019.

[52] W. Yang *et al.*, "Feasibility of automatic measurements of hip joints based on pelvic radiography and a deep learning algorithm," *European Journal of Radiology*, vol. 132, p. 109 303, 2020.

[53] W. Liu, Y. Wang, T. Jiang, Y. Chi, L. Zhang, and X.-S. Hua, "Landmarks detection with anatomical constraints for total hip arthroplasty preoperative measurements," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, Springer, 2020, pp. 670–679.

[54] C. Liu, H. Xie, S. Zhang, Z. Mao, J. Sun, and Y. Zhang, "Misshapen pelvis landmark detection with local-global feature learning for diagnosing developmental dysplasia of the hip," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3944–3954, 2020.

[55] T.-T. Pham, M.-B. Le, L. H. Le, J. Andersen, and E. Lou, "Assessment of hip displacement in children with cerebral palsy using machine learning approach," *Medical & Biological Engineering & Computing*, vol. 59, no. 9, pp. 1877–1887, 2021.

[56] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[57] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[58] 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*, 2017, pp. 1492–1500.

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

[60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

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

[62] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *kdd*, vol. 96, 1996, pp. 226–231.

[63] 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.