SAPIENZA
UNIVERSITÀ DI ROMA

Semantic Segmentation of Key Fetal Brain Structures in Trans - Thalamic Ultrasound Images Using Deep Learning

Facoltà di Ingegneria dell'Informazione, Informatica e Statistica
Corso di Laurea Magistrale in Data Science

Tuba Siddiqui
ID number 2047057

Advisor
Prof. Luigi Cinque

Co-Advisor
Dr. Alessio Fagioli

Academic Year 2022–2023

Thesis not yet defended

Semantic Segmentation of Key Fetal Brain Structures in Trans - Thalamic Ultrasound Images Using Deep Learning
Master Thesis. Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: siddiqui.2047057@studenti.uniroma.it

To my family, Ghareebs in Italy and Roma.

Abstract

This thesis presents a deep learning framework for automatically segmenting fetal brain structures in trans-thalamic ultrasound images. Manually evaluating small midline structures like the cavum septi pellucidi (CSP) and lateral ventricles (LV) is challenging, prone to errors and relies heavily on the operator's skill. To solve this problem, I developed a pipeline based on Attention U-Net, which was trained on the Large Fetal Head Biometry Dataset to segment brain parenchyma, CSP and LV.

To address the significant class imbalance, I combined weighted cross-entropy and focal loss functions with focused oversampling strategies. The model performed well on brain parenchyma (Dice = 0.963) and achieved competitive results for CSP (Dice = 0.792) and LV (Dice = 0.762), which are both underrepresented and important structures in clinical practice.

In addition to segmentation, I introduced an anomaly detection module that creates heatmaps and quantitative scores to identify mismatches with expert annotations. This improves interpretability and provides a useful diagnostic aid. Compared to earlier 3D or proprietary methods, this work offers a lightweight, reproducible and practical solution based on commonly available 2D ultrasound data.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Background and Motivation

Prenatal screening is significant in modern medicine as it allows for the early diagnosis of fetal abnormalities. Diagnosis of abnormalities in the brain during pregnancy is particularly significant, as they are frequently associated with significant developmental retardation, neurological abnormalities or even perinatal death. Early and effective diagnosis can inform clinical decision making, parental counseling and possible interventions that enhance both neonatal and maternal outcomes.

In normal obstetric ultrasound, the trans-thalamic (TT) plane is generally regarded as the gold standard reference plane of brain assessment. It gives an unobstructed view of essential anatomical structures like the thalami, cavum septi pellucidi (CSP) and lateral ventricles (LVs). Not only are these structures critical for the evaluation of normal growth, but their abnormalities are usually the earliest signs of disorders like hydrocephalus, ventriculomegaly or agenesis of the corpus callosum.

Although crucial, manual interpretation of TT-plane ultrasound images is problematic. The image quality of ultrasound scans can be highly variable depending on fetal position, operator skill and artifacts. Also, small brain structures such as the CSP and ventricles take up very few pixels relative to total brain parenchyma and are difficult to identify by eye or with conventional image analysis methods.

1.2 Problem Statement

The manual detection of fetal brain anomalies from ultrasound scans is still an enormous challenge in clinical practice. Even though the trans-thalamic plane gives the best views of the key anatomic landmarks, identifying small and fragile structures such as the cavum septi pellucidi and the lateral ventricles is not easy.

Several factors contribute to the difficulty of the task:

1. **Poor image quality:** Ultrasound scans are often marred with speckle noise, shadowing and varying grade of contrast such that fine details are masked

from view.

2. **Operator dependency:** Interpretation, mostly subjective, is greatly influenced by the skill of the sonographer, which may vary considerably.

3. **Size imbalance:** The small structures such as the CSP and ventricles occupy a tiny portion of the image, thus accidental detection can be difficult against the larger background of the brain parenchyma.

4. **Time pressure:** Within a routine clinical setting, sonographers and radiologists must analyze many scans within a very short time frame; this increases the risk of missing some abnormal findings.

Thus, manual evaluation suffers from variability and lack of accuracy, resulting in delays in diagnosis or interpretation errors pertaining to some of the brain anomalies. Therefore, automatic techniques should be developed for standardization and reproducibility.

1.3 Research Objectives

1.3.1 Automated Segmentation of Key Brain Structures

The main goal of this study is to create a deep learning model for automatically segmenting important fetal brain structures in trans-thalamic ultrasound images. We focus on the brain parenchyma, CSP and LV. This involves designing an Attention U-Net architecture tailored for grayscale ultrasound images with four output classes: background, parenchyma, CSP and LV. The model is trained on a dataset of 1,565 paired images and masks ([7]). It uses data augmentation techniques such as horizontal flipping, shifting, scaling, rotating and adjusting contrast to overcome issues with low-resolution images, speckle noise and fetal movement ([4]). Post-processing steps—including morphological closing and connected component analysis—refine segmentations of small structures like CSP and LV, keeping only the largest components above a 5-pixel threshold to improve accuracy ([15]). The segmentation outputs will support clinical applications by providing quantitative metrics like area and centroid for detecting anomalies. We evaluate performance using Dice scores and Intersection over Union (IoU), focusing on CSP and LV due to their importance for diagnosis ([1]).

1.3.2 Address Class Imbalance of Small Anatomical Structures

A secondary objective is to reduce the class imbalance found in segmenting small anatomical structures, such as CSP and LV, which are less common in ultrasound datasets compared to larger areas like the brain parenchyma. This study uses a WeightedRandomSampler with an oversampling factor of 3.0 for images with CSP or LV to ensure balanced representation during training (as set in the dataset configuration). Also, we apply a hybrid loss function that combines SoftDiceLoss (weighted

0.4) and FocalLoss (weighted 0.2). The class-specific weights come from inverse frequency normalization to give more importance to rare classes, smoothed by a factor of 0.0 to prevent extreme bias. This method aims to improve segmentation accuracy for small structures, addressing the issue of imbalanced data distribution noted in earlier studies ([17]). We assess the effectiveness of this strategy through per-class Dice scores and visual checks of segmentation consistency across the validation set.

1.3.3 Evaluate U-Net for Robustness and Reliability

The third objective is to evaluate the robustness and reliability of the Attention U-Net model under varying conditions, ensuring its applicability in clinical settings. This involves testing the model's performance across diverse ultrasound images with different acquisition parameters (e.g., varying resolutions, noise levels) and fetal poses, as simulated through data augmentation and validation on a held-out test set ([4]). Robustness is assessed by measuring segmentation consistency using metrics such as standard deviation of Dice scores across multiple runs with different seeds and by analyzing sensitivity to input perturbations (e.g., Gaussian noise, contrast shifts). Reliability is evaluated through inter-run reproducibility and comparison with ground truth annotations, targeting a minimum intra-class correlation coefficient (ICC) of 0.9 for key structures ([15]). The study also explores the model's generalization to unseen datasets, leveraging cross-validation and external validation on publicly available fetal ultrasound datasets where feasible, to address the need for dependable diagnostic tools in prenatal care ([1]). This study aims to aid the automated, reliable and efficient tool for clinicians, which can help them with prenatal brain anomalies detection without being heavily dependent on subjective human interpretation.

1.4 Significance of Study

Fetal neurological anomalies, especially in the very small anatomical areas such as the cavum septi pellucidi and lateral ventricles, bear subtle signs in ultrasound images. Despite their small sizes, these structures have great diagnostic value:

The clinical significance of accurately segmenting key fetal brain structures in transthalamic ultrasound images lies in their association with critical developmental anomalies. Absent or abnormal CSP is closely linked to agenesis of the corpus callosum and other severe midline defects, necessitating early detection for informed prenatal management. Similarly, dilated lateral ventricles serve as an early marker for developing ventriculomegaly or hydrocephalus, conditions that require timely intervention to mitigate potential neurological impairments. These insights underscore the importance of robust segmentation techniques in supporting diagnostic reliability and improving maternal-fetal outcomes.

With these structures being difficult to manually identify, there is a high chance of misdiagnosis occurring within clinical practice. The goal of this study is to develop an automated tool that provides accurate and consistent highlighting and

measurement of these structures, enabling more precise and efficient automated deep learning–based segmentation. This will help in improving the accuracy and efficiency of prenatal diagnosis.

The significance of this research lies in its potential to revolutionize fetal brain assessment through advanced deep learning techniques. It aims to streamline fetal segmentation by delivering standardized and reproducible results, thereby reducing subjectivity inherent in manual assessments by sonographers and radiologists. Furthermore, it seeks to enhance clinical workflows, enabling medical professionals to prioritize interpretation over time-consuming automated calculations. Additionally, the study addresses the critical challenge of segmenting small, unbalanced classes in real-world ultrasound data, paving the way for significant advancements in the application of artificial intelligence in medical diagnostics.

1.5 Research Questions

This study is guided by these research questions:

- **RQ1:** Can U-Net achieve precise segmentation of fetal brain parenchyma, CSP and LV in the trans-thalamic plane?
- **RQ2:** How does class imbalance impact segmentation performance, especially for CSP and LV?
- **RQ3:** How do different loss functions like Focal Loss improve segmentation results compared to standard Cross-Entropy?
- **RQ4:** What are the effects of automated segmentation on clinical practice in prenatal anomaly detection?

1.5.1 Mapping of Research Questions to Contributions

As shown in Table 1.1, the U-Net model addresses key challenges in ultrasound segmentation.

1.6 Scope and Limitation

This study examines the trans-thalamic (TT) plane of fetal ultrasound imaging, as it is the reference view for the evaluation of midline brain structures like the thalami, CSP and the lateral ventricles (LVs). Hence, the focus is limited to segmentation activities on this plane.

This work excludes the transcerebellar and transventricular planes, which are also of clinical importance. Although these other views contain information about the cerebellum and the posterior fossa which is important from a clinical standpoint,

RQ	Research Question	Thesis Contribution	Where Discussed
RQ1	Can a deep learning model (U-Net based) reliably segment thalamic structures such as CSP from ultrasound images?	Developed a U-Net-based ultrasound image segmentation system for CSP parenchyma reliability assessment using Dice IoU.	Chapter 4 (Results, Discussion); Chapter 5 (Clinical Reliability)
RQ2	How does class imbalance affect segmentation performance for small CSP and LV-like structures?	Analyzed pixel distributions of CSPLV and highlighted assessment effects on segmentation.	Chapter 3 (Loss Functions); Chapter 5 (Discussion Impact)
RQ3	Do weighted loss functions (e.g., Focal Cross-Entropy vs. Class Cross-Entropy Loss) improve segmentation of imbalanced structures?	Implemented and compared CrossEntropy with Focal Loss detection of CSP and LV relative to baseline.	Chapter 4 (Metrics); Chapter 5 (Analysis Improvements)
RQ4	Can attention outputs be leveraged to generate anomaly detection maps comparing predicted annotations?	Generated heatmaps and boundary truth visualizations potential application in clinical screening.	Chapter 4 (Results); Chapter 5 (Discussion); Chapter 6 (Clinical Significance)

Table 1.1. Summary of Research Questions

their inclusion would necessitate a much larger dataset and additional training of more sophisticated models, which is out of scope for this project.

Furthermore, the study is subject to the following limitations:

1. **Data dependency:** Note that the model was trained on publicly available data which does not cover all types of variations in real clinical practice (i.e., differences between machines, diversity of populations).
2. **Plane restriction:** The model has only considered the trans-thalamic plane and therefore the generalizability of the model relative to other fetal brain views may be limited.
3. **Class imbalance problem:** In terms of the conceptual aspects of the project (including weighted losses), very small structures (i.e., CSP, LVs) will be harder to identify than larger regions, by their very nature.
4. **No anomaly classification:** Similar to point 5, the model utilizes segmentation for structure identification and does not classify or provide diagnosis of anomalies. It would also require to have created the necessary work linking the segmentation outputs to clinical diagnosis since it was not a part of the current research.

Even with these limitations, this study represents a significant move towards automated prenatal brain analysis and can act as a starting point for future extensions of databases, platforms, greater number of subjects in multiple planes and anomalous classification.

1.7 Chapter Summary and Thesis Outline

In summary, this chapter first provided an introduction to the importance of prenatal brain screening, utilizing the trans-thalamic plane as the baseline for assessment of midline structures. The difficulties of manual detection were discussed, including the operator-dependent nature, quality of the image produced and the small size of the CSP and lateral ventricles. The objectives set out to overcome these difficulties are clear and include developing a U-Net based segmentation model, means of addressing class imbalanced problems and evaluating model robustness. It is important to note that the intended contributions of this research, which will lead to an enhancement of the workflow by contributing a reduction in subjectivity, improving consistency and the selection of options in clinical practice. Finally, the goals and objectives of the scope and limitations were presented in detail which limits the project to the trans-thalamic plane, therefore excluding other views as well as excluding the ability to perform anomaly classification directly.

The remaining chapters of the thesis are organized as follows:

- **Chapter 2 - Literature Review:** Surveys the existing literature on fetal brain ultrasound analysis, deep learning methods for medical image segmentation and issues such as class imbalance.

- **Chapter 3 - Methodology:** This chapter explains the dataset, preprocessing, model architecture (U-Net), training methods and evaluation metrics.
- **Chapter 4 - Results and Discussion:** This chapter presents the experimental results, compares performance for different loss functions, notes any potential to identify small structures and has a discussion.
- **Chapter 5 - Conclusion and Future Work:** This chapter summarizes the notable findings, discusses limitations and provides recommendations on how the research might be furthered to include wider clinical practice.

Chapter 2

LITERATURE REVIEW

The integration of Deep Learning (DL) into medical sciences has increased flexibility and convenience with diagnosis, especially early-stage detection of underlying issues. With such advancements, precision reading of different imaging modalities scans has become easier and a great help for clinicians. They know how to read scans, but scan interpretations are subjective, depend on experience and are error-prone. Even experienced sonographers can miss subtle anomalies due to poor image quality, small structure size or clinical time pressure. So, AI acts as a consistent, objective second reader that flags potential issues and supports, rather than replaces, the clinician.

The literature review for this thesis aimed to place my segmentation method for fetal brain stuctures in ultrasound images within a proper context. The main dataset, introduced by [1], includes 1,565 annotated 2D ultrasound images. This dataset is a recent public resource ([1]) and was published too recently (late 2023) for widespread use. As of 2025, only the authors' group has published direct comparisons on this dataset, with about 20 citations mostly found in reviews or extensions, such as [2]. No external papers have conducted full state-of-the-art segmentation on it. For instance, a 2025 bias analysis focuses on ethics without providing metrics and other studies, like [13], use different datasets. This lack of coverage required a dual state-of-the-art approach. One Table 4.6 compares models exclusively on the [1] dataset for direct evaluation. Another table Table 4.5 includes papers using different modalities, such as 3D MRI and 2D MRI, to provide a broader context for fetal brain segmentation while recognizing the limitations in comparing different datasets.

2.1 Artificial Intelligence in Prenatal Ultrasound

Artificial intelligence has shown increasing accuracy in prenatal ultrasound diagnosis, with an advancement of research dedicated to automating the analysis of fetal anatomy ([14]). While significant progress has been made in tasks such as biometric estimation and standard plane detection, a critical gap remains in the detailed structural analysis of fetal brain anatomy, particularly in the challenging third trimester where image quality often degrades and subtle anomalies become

harder to detect ([14]). This thesis addresses this gap by introducing another approach centered on semantic segmentation of main brain structures in the trans thalamic plane, specifically the Brain, Cavum Septum Pellucidum (CSP) and LV, to enable automated anomaly detection ([1]).

2.2 Fetal Brain Structures and Their Clinical Relevance

The lateral ventricles (LVs) and the CSP are two significant midline structures that are systematically assessed in prenatal neurosonography. Visualization of these structures in the trans-thalamic plane is included in the routine biometry protocol since their existence, shape and size are good indicators of normal fetal brain development.

The CSP is a midline cavity filled with fluid that falls between the septum pellucidum's two leaflets. Its detection in the second trimester is regarded as a critical marker for normal brain development since its absence or its abnormal size might be indicative of agenesis of the corpus callosum, septo-optic dysplasia or holoprosencephaly [7]. Clinicians thus treat CSP measurement as a routine milestone in fetal neurosonography.

The lateral ventricles are bilaterally paired fluid spaces that hold cerebrospinal fluid. Measurement is especially relevant to the detection of ventriculomegaly, an atrial diameter $\geq$ 10 mm, which is among the most frequent antenatal diagnoses of neurodevelopmental impairment. Ventriculomegaly can be mild and isolated through severe and syndromic and thus early detection and measurement are critical in clinical management.

Although they are clinically important, CSP and LVs are hard to identify manually. Both are very small areas of the ultrasound image relative to the rest of the brain parenchyma. Moreover, ultrasound is prone to artifacts, shadowing and operator dependence so that minute abnormalities are likely to be overlooked. This encourages the construction of automated techniques for the precise identification of these structures on a consistent and objective basis, minimizing inter-observer variability and aiding in early diagnosis.

2.2.1 Gap in prior work

While current AI applications in prenatal imaging have mainly concentrated on estimating biometrics or detecting standard planes, few studies have tackled the detailed segmentation of the CSP and LVs. Most previous segmentation work has aimed at larger structures or broad brain areas, creating a gap in the automated analysis of smaller, clinically relevant midline structures. This thesis directly addresses this gap by using U-Net-based segmentation of the CSP and LVs in the trans-thalamic plane.

2.3 Deep Learning for Fetal Ultrasound Analysis

Previous work has explored the use of convolutional neural networks (CNNs) for fetal brain analysis in ultrasound. For instance, [15] developed a CNN to detect the fetal visually salient plane (FVSP) by localizing the brain and assessing features like CSP visibility. While this work identifies the presence and orientation of key structures, it does not provide the granular, pixel-wise segmentation necessary for detailed structural assessment. In contrast, this project focuses on semantic segmentation, offering a much richer level of information by portraying the precise shape, size and location of the CSP, LV and brain. As illustrated in Table 1, semantic segmentation inherently provides information about the presence of a structure and also extends this to its spatial characteristics, enabling quantitative measurements and a more direct pathway to anomaly scoring. The ability to visualize the segmented structures and quantify deviations from expected norms offers a significant advantage over simple visibility detection.

Comparison between visibility detection and semantic segmentation for fetal brain structures.

Clinical Question	**Visibility**	**Semantic Segmentation**
Is the CSP present?	Yes/No decision	Yes – implied from segmentation mask
Where is it located?	Not specified	Explicit location provided by mask
What is its size?	Not measurable	Quantitative measurement from pixels
Can anomalies be detected?	Difficult, indirect	Possible through mismatch with ground truth

Table 2.1. Comparison of Clinical Questions

2.4 Architectures for Medical Image Segmentation

Recent work has explored changes to the standard U-Net architecture to improve segmentation accuracy. One example is the Multi-Feature Pyramid U-Net (MFP-Unet) [16]. This model combines U-Net with a feature pyramid network (FPN). It does this to better capture both fine-grained details and high-level contextual information using multi-scale feature maps. In echocardiography, this design has improved the segmentation of complex cardiac structures. It achieved a 14.5% higher IoU compared to the standard U-Net. While these approaches show the potential of multi-scale representations for medical imaging, their higher computational cost and memory needs may limit clinical use.

While modern architectures like GANs, RNNs, LSTMs and Transformers have been studied in medical imaging, they often face challenges such as training instability, high computational costs and large data needs. In contrast, U-Net offers a good balance of performance, spatial accuracy and ease of use. This makes it especially suitable for 2D fetal ultrasound segmentation tasks. According to [16], U-Net re-

mains one of the most effective and widely used architectures for medical image segmentation because of its encoder-decoder design with skip connections. This design effectively captures both the overall context and fine anatomical details.

Building on this foundation, this thesis focuses on a lightweight, explainable and reproducible U-Net pipeline for detecting structural problems in trans-thalamic brain ultrasound images. Compared to more resource-intensive options, this choice improves feasibility for real-world clinical screening, especially in low-resource or time-sensitive settings.

2.5 Imaging Modalities: 2D vs 3D Approaches

Furthermore, recent deep learning methods for fetal ultrasound analysis have been built on proprietary 3D datasets. These methods often need a lot of manual annotation or use complex multi-model architectures ([4]). Although they show strong performance, their dependence on specialized datasets and resource-heavy processes limits their use in clinical settings and widespread adoption. For instance, [5] found that 3D CNNs can accurately segment subcortical fetal brain structures such as the CSPV and LPVH. They reported high Dice scores but relied heavily on proprietary volumetric ultrasound data.

Similarly, [17] applied deep learning to 3D fetal MRI brain scans and achieved a Dice score of 0.897. However, their method needed high-quality MRI data, which is not commonly available in routine prenatal care. In contrast, [9] showed that a 2D U-Net variant for fetal MRI segmentation achieved over 90% Dice. This finding suggests that 2D methods can compete with 3D approaches while being much more efficient computationally.

This thesis follows the latter approach by using a publicly available 2D dataset ([1]) and a single established architecture, U-Net. This decision supports transparency, explainability and reproducibility. It also provides a method that is more practical for real-world clinical practice, including low-resource settings.

[17] showed that deep learning is effective for fetal brain segmentation using 3D MRI. They achieved a Dice score of 0.897 and proved their method works well across different gestational ages, even in cases of congenital heart disease. While these findings highlight the potential of deep networks for complex anatomical segmentation, they rely on MRI data, which is expensive and not commonly available in prenatal care when compared to ultrasound. In contrast, this thesis focuses on 2D trans-thalamic ultrasound, which is the globally accepted standard for fetal head measurement. It uses a public dataset ([1]) alongside a lightweight U-Net model. This method emphasizes transparency, reproducibility and practical use. It is more suitable for real-world situations, especially in low-resource settings where MRI is not an option.

2.6 Class Imbalance in Medical Image Segmentation

A key challenge in fetal brain ultrasound segmentation is class imbalance. In a typical trans-thalamic scan, most pixels belong to the background and brain tissue. On the other hand, clinically important midline structures like the CSP and the lateral ventricles (LVs) take up only a small portion of the image. As a result, a segmentation model trained with a standard loss function tends to favor the majority classes, often overlooking small but vital areas.

2.6.1 Clinical importance of imbalance

From a medical standpoint, this imbalance is significant. The CSP and LVs are the exact structures that clinicians need to evaluate for early diagnosis of conditions like agenesis of the corpus callosum or ventriculomegaly. A model that reliably predicts the background and tissue but fails to identify the CSP or LV has limited clinical use.

Loss-function strategies

Several approaches have been suggested to tackle class imbalance in medical image segmentation:

- **Weighted Cross-Entropy Loss:** It assigns higher penalties to errors in underrepresented classes by weighting them based on their frequency. This ensures that mistakes on CSP/LV pixels impact gradient updates more than mistakes on background pixels.
- **Dice-based losses:** It uses the Dice coefficient, which measures the overlap between predicted and actual masks. Loss functions like Dice Loss or Generalized Dice Loss [11] naturally address class imbalance by focusing on overlap in regions instead of pixel counts.
- **Focal Loss [6]:** It reduces the weight of easy, well-classified examples and highlights tough cases. This makes the model more sensitive to small and unclear areas like CSP and LV.

Data-level strategies

In addition to loss functions, imbalance can also be reduced at the data level:

- **Data augmentation:** It (including flips, rotations and brightness changes) increases the variety of training examples. This exposes the model to more versions of minority classes.
- **Oversampling:** It means that images containing rare structures can be sampled more often during training. This ensures the model sees CSP and LV frequently enough to learn their features.

2.6.2 Gap in prior work

Many current segmentation studies report high average Dice or IoU, but they do not give detailed per-class metrics. This hides poor performance on small structures and creates a misleading view of model reliability. In contrast, this thesis focuses on minority-class performance by implementing weighted and focal losses, oversampling strategies and per-class evaluation to make sure clinically important structures are adequately represented.

2.7 From Segmentation to Anomaly Detection

Recent research has also been focused on segmenting midline brain structures like the cavum septum pellucidum et vergae (CSPV) since they are of clinical relevance in detecting fetal anomalies [7]. Although such endeavours predominantly have good segmentation accuracy, they predominantly aim to forecast pixel-level boundaries and not use segmentation output for anomaly detection.

This thesis differs from existing work in that as well, it utilizes the discrepancy between predicted and expert-marked segmentation masks directly as an indicator of potential structural abnormalities. By having a model learn what the correct anatomical configuration is supposed to be, we can subsequently seek areas where the predicted structure is significantly different, signifying a potential abnormality. This approach extends beyond simple mask prediction and extends to querying: "Does this structure appear incorrect?" and therefore bridges the gap between segmentation and clinically interpretable interpretation, gap in recent researches.

One of the main limitations of current segmentation studies is the lack of explicit anomaly assessment. Most existing research reports segmentation accuracy using metrics like Dice or IoU, but they do not provide tools for identifying or marking potential structural issues. For instance, prior studies on subcortical structures like the CSPV and LPVH have mainly focused on predicting pixel boundaries without extending their results to anomaly detection.

This thesis fills that gap by adding an anomaly detection step, which measures and visualizes differences between predicted and expert-marked segmentation masks. By generating anomaly scores and heatmaps, the proposed framework highlights areas with significant mismatches. This provides clinicians with a clear tool to identify potential abnormalities. Unlike earlier studies, this method reframes segmentation results not only as accuracy metrics but also as important indicators of structural irregularities.

As shown in Table 4.5, most previous studies focus on larger brain structures or use private 3D datasets, with little attention given to small midline structures like the CSP and LV. Also, anomaly detection is seldom included in segmentation results. This thesis tackles these issues by using a public 2D dataset, employing imbalance-aware training strategies and expanding segmentation into clinically meaningful anomaly detection.

2.8 Research Gaps and Motivation for This Study

The literature reviewed in this chapter shows the rapid progress of deep learning methods in prenatal imaging. However, significant gaps remain that limit scientific progress and clinical use.

As summarized in Table 4.5, most previous work has focused on larger brain regions or general biometric measurements, often using proprietary 3D datasets that are hard to reproduce. Although these methods perform well, they require a lot of computing power and are not suitable for routine prenatal care. In contrast, the trans-thalamic 2D plane is still the global standard for neurosonography, yet few studies have looked closely at small midline structures like the CSP and lateral ventricles (LVs).

Another clear issue is the lack of specific anomaly detection frameworks. Most segmentation studies focus on pixel-level accuracy (like Dice and IoU) without translating these results into clear indicators of abnormality. New methods like diffusion-based anomaly detection are still largely experimental and require a lot of data.

A further challenge is class imbalance, which affects small but clinically significant structures such as CSP and LV. Some studies have tried using Dice or Focal Loss in different areas, but these techniques have not been systematically applied to fetal brain ultrasound segmentation.

This thesis addresses these gaps by implementing semantic segmentation of brain tissue, CSP and LV in the trans-thalamic plane using a lightweight and reproducible U-Net approach. It also tackles the class imbalance issue with weighted and focal loss functions, as well as biased sampling strategies. Furthermore, it extends the segmentation to include anomaly detection by measuring and visualizing deviations from expert-annotated ground truth. Lastly, by demonstrating that a publicly available 2D dataset can be used, promoting transparency, explainability and wider reproducibility compared to earlier 3D or proprietary methods.

By closing the gap between segmentation accuracy and clinically useful outputs, this work aims to advance both the technical literature and the future integration of AI tools in routine fetal neurosonography.

Chapter 3

METHODOLOGY

3.1 Research Design

In this project, I used an experimental research design focused on creating and testing a supervised deep learning pipeline for semantic segmentation. The goal was to automate the identification of important fetal brain structures in the trans-thalamic plane and to see if the segmentation outputs could be used for clinically interpretable anomaly detection.

The research followed a clear workflow:

1. **Dataset selection and preparation** I used the Large Fetal Head Biometry Dataset, concentrating on the trans-thalamic plane. I preprocessed the images and masks, normalizing and augmenting them to ensure consistency and improve generalization.

2. **Model development** I implemented a U-Net architecture because it has a strong track record in biomedical image segmentation and offers a good mix of accuracy, efficiency and interpretability.

3. **Loss function and training strategies** To tackle class imbalance, I tested weighted cross-entropy and focal loss, along with oversampling minority-structure images. I trained the models under controlled conditions, using early stopping, mixed precision and best-model checkpointing.

4. **Evaluation** I evaluated the models with overlap-based metrics, such as Dice and IoU and class-wise performance scores. I compared a CSP-only model to the extended CSP+LV model to show the impact of broadening the segmentation task.

5. **Anomaly detection** In addition to segmentation accuracy, I introduced a framework that compared predictions to ground truth, generating heatmaps and anomaly scores. This provided interpretable indicators of possible abnormalities.

This design ensured that the study not only assessed model accuracy but also tackled broader issues of clinical feasibility, reproducibility and interpretability. These factors are crucial for implementing AI tools in real-world prenatal imaging workflows.

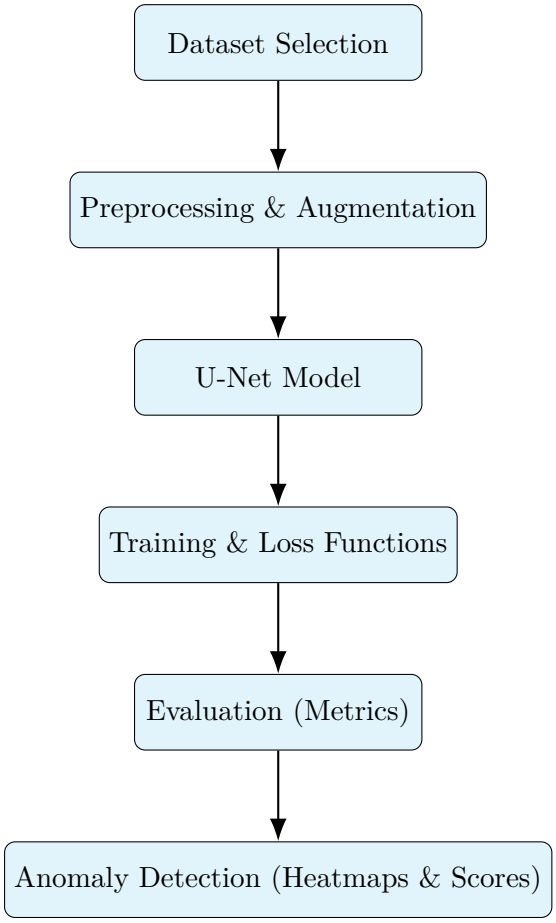


Figure 3.1. Research pipeline

As shown in Figure **??**, research ipline is formed from dataset selection through preprocessing, model training, evaluation and anomaly detection.

3.2 Dataset Description

This study uses data from the Large Fetal Head Biometry Ultrasound Dataset ([1]), which offers a complete collection of prenatal ultrasound images and their annotations. From this dataset, I selected the trans-thalamic plane subset. This subset is the clinical standard for assessing important midline brain structures, including the CSP and the LV.

In total, about 1565 two-dimensional ultrasound images and their corresponding segmentation masks were available. Each mask is saved in PNG format, with unique color channels representing anatomical classes:

- **Black (0,0,0):** Background
- **Red (255,0,0):** Brain parenchyma
- **Green (0,255,0):** CSP
- **Blue (0,0,255):** LV

Out of the total ultrasound scans in the dataset, only a portion included the minority structures of interest. The CSP appeared in 584 images, while the LV showed up in 505 images. This reveals a significant class imbalance. Although the brain parenchyma and background are found in nearly all scans, CSP and LV are present in less than one-third of the data. This imbalance led to the use of weighted losses, oversampling strategies and per-class evaluation metrics, as explained in Sections 3.4 and 3.5.

To ensure a strong evaluation, the dataset was divided into three subsets: training, validation and testing. A stratified split kept class representation consistent across the sets. Seventy percent went to training, while fifteen percent was assigned to validation and another fifteen percent to testing. This division supported model development, hyperparameter tuning and an unbiased assessment of performance.

A key challenge in this dataset is class imbalance. The brain parenchyma takes up most of the pixels, while CSP and LV make up a very small portion. This imbalance greatly impacts the learning process because standard training can lead the model to favor majority classes and ignore important minority structures. Tackling this issue is essential to the methodology, as explained in Section 3.5.

Figure 3.2 shows the dataset management workflow. I began with raw ultrasound images and RGB segmentation masks. First, I selected only the trans-thalamic plane. Then, I cleaned the data and verified the labels to ensure consistency. Next, I converted the RGB masks into integer-coded masks that work with PyTorch training. I split the dataset into training, validation and test subsets at a ratio of 70/15/15. I also calculated class frequency statistics to measure the imbalance among brain parenchyma, CSP and LV. These steps were essential for preparing the dataset for effective training and for making sure minority structures were well represented.

3.3 Preprocessing and Data Augmentation

To prepare the dataset for deep learning, I took several steps to preprocess both the ultrasound images and the segmentation masks. This ensured consistency and compatibility with the U-Net architecture.

1. **Image resizing and normalization**

   I standardized all images and masks to a resolution of 256 × 256 pixels. I did this using Longest-Max-Side resizing followed by zero-padding. This method preserved the original aspect ratio while ensuring uniform input dimensions.

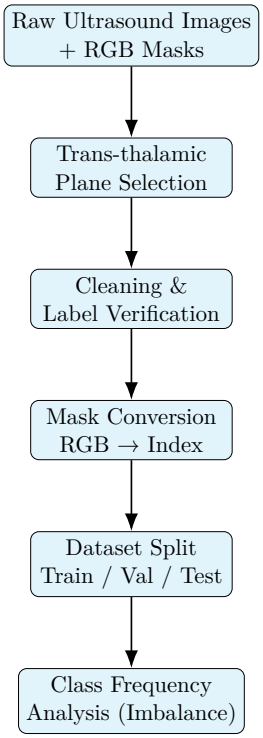


Figure 3.2. Dataset management workflow

I then normalized the grayscale intensities using Albumentations' Normalize. This tool first divides the values by 255 and then standardizes them with a mean of 0.5 and a standard deviation of 0.5. This process produced input values within the range of -1 to 1, which helped with training stability.

Mask encoding

The segmentation masks were in RGB PNG format. I converted them into integer-encoded class indices that are suitable for PyTorch's multi-class loss functions.

- 0 = Background
- 1 = Brain parenchyma
- 2 = CSP
- 3 = LV

This ensured that the model viewed the segmentation task as a pixel-wise classification problem instead of a color-matching task.

2. **Data augmentation**

To improve model generalization and account for variability in real ultrasound imaging, I applied on-the-fly data augmentation during training. Augmentations included:

- **Geometric:** horizontal flips (p = 0.5), random shifts, scales and rotations (±6.25
- **Photometric:** random brightness and contrast adjustments (p = 0.3) and CLAHE (Contrast Limited Adaptive Histogram Equalization, p = 0.2).

These augmentations simulated common imaging variations, reduced overfitting and exposed the model to more varied examples of small structures. For the validation and test sets, I only applied resizing, padding and normalization without any augmentation. Since Albumentations uses nearest-neighbor interpolation for masks, the class labels were preserved during geometric transformations.

3. **Sampling strategy** Given the class imbalance in the dataset, where CSP and LV occupy only a small number of pixels compared to the brain parenchyma and background, I adjusted the mini-batch sampling. Using PyTorch's WeightedRandomSampler, I gave higher sampling weights to images containing CSP or LV. This increased the occurrence of minority-class examples during training, supporting the loss-level strategies described in Section 3.5.

3.4 Model Architecture

To complete this task, I employed an Attention U-Net [8], which is a variant of the original U-Net architecture proposed by [10]. This adaptation differentiates my work from standard U-Net approaches by explicitly modifying the skip connections with attention gating, making the architecture original rather than off-the-shelf. Like U-Net, the model employs an encoder–decoder with skip connections and is particularly suitable for biomedical image segmentation applications where global context and local spatial detail are both significant, such as fetal brain structure analysis. The Attention U-Net has around 15.1 million parameters when using a base channel size of 64. Although it is larger than a standard U-Net because of the attention gates, it can still be trained effectively on 2D ultrasound data. It also offers improved sensitivity to small structures like the CSP and LV.

The encoder path reduces the spatial resolution steadily using convolution and pooling operations and doubles the feature channels at every step. The network can represent very abstract and high-level representations of the inputted ultrasound image. The features are most semantic but most compressed at the bottleneck. The decoder path then upsamples these features more and more, reducing the number of channels but restoring the original spatial resolution.

One significant modification in this work is introducing attention gates to the skip connections. Rather than just concatenating decoder features and encoder features, the skip connections go through attention blocks first, which learn to turn off irrelevant activations and give more importance to areas of interest. This mechanism enhances the model's ability to emphasize sparse and underrepresented structures

such as the CSP and LV, which are generally overlooked in standard U-Net due to class imbalance and sparsity in their spatial contribution.

The model takes as input a 256×256 grayscale ultrasound image and gives as output a 256×256 segmentation mask, where each pixel is assigned one of four classes: background, brain parenchyma, CSP or LV. The model employs a softmax activation at the output to give per-pixel class probabilities. With the use of attention in a light U-Net backbone, the model achieves a balance between performance, interpretability and efficiency, making it both clinically actionable and computationally feasible. Figure 3.2 shows the Attention U-Net architecture used in this work. The model relies on the typical encoder–decoder scheme: the encoder progressively downsamples the input image through convolutional blocks with increasing feature depth (64, 128, 256, 512) until reaching the bottleneck stage (1024 channels), where the most abstract features are learned. The decoder then upsamples features in a sequence of steps (512, 256, 128, 64) to progressively add up spatial resolution and finally output the terminal segmentation mask. The attention gates used for the skip connections from the encoder to the decoder are the striking feature of this architecture. These gates are also used as irrelevant information suppressors and highlight important activations before concatenation, thereby improving the ability of the model to attend to small or under-represented features such as CSP and LV.

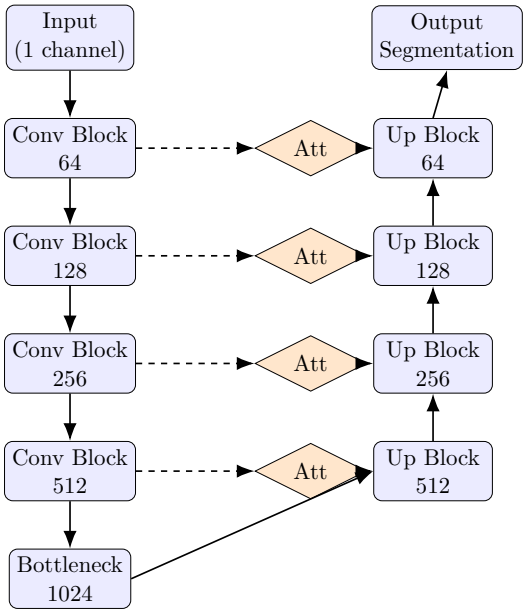


Figure 3.3. Attention U-Net architecture

In Figure 3.3, Attention U-Net architecture is explained. Encoder (left) extracts features; decoder (right) reconstructs segmentation. Skip connections pass through attention gates (orange diamonds) before concatenation, improving focus on small structures such as CSP and LV.

3.5 Loss Function

Training the model involves optimizing a pixel-wise classification loss. However, there is a significant class imbalance between large structures, like brain parenchyma and small regions, such as CSP and LV. As a result, standard loss functions often favor the majority classes. To tackle this issue, two different loss functions were used and compared:

1. **Cross-Entropy Loss (baseline):** The standard categorical cross-entropy loss calculates the negative log-likelihood of the correct class for each pixel. While it works well for balanced data, it struggles with minority classes because of their low frequency.

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\,\log(p_{ic}), \tag{3.1}$$

   where N is the total number of pixels, C is the number of classes, $y_{ic} \in \{0, 1\}$ is the one-hot encoded ground-truth label for pixel i and class c, and p_{ic} is the predicted probability of pixel i belonging to class c from the softmax output.

2. **Class-Weighted Cross-Entropy:** To address the imbalance, class weights were applied that are inversely related to class frequency. This makes sure that mistakes in minority structures, like CSP and LV, have a greater impact on the overall loss, pushing the model to focus on these important areas.

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_c\, y_{ic}\,\log(p_{ic}), \tag{3.2}$$

   where w_c is the class weight, defined as the inverse of the class frequency f_c (i.e., $w_c \propto 1/(f_c+\epsilon)$), y_{ic} is the one-hot ground-truth label, p_{ic} is the predicted probability, N is the number of pixels, and C is the number of classes.

3. **Focal Loss:** Focal Loss [6] reduces the loss contribution from well-classified pixels. It places more emphasis on hard-to-classify pixels. This approach is particularly useful for small or poorly represented structures where misclassifications frequently occur.

$$\mathcal{L}_{\text{Focal}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} \alpha_c\,(1-p_{ic})^{\gamma}\, y_{ic}\,\log(p_{ic}), \tag{3.3}$$

   where $\gamma > 0$ is the focusing parameter that reduces the loss contribution of well-classified pixels, α_c is the class-balancing weight for class c, y_{ic} is the ground-truth one-hot label, p_{ic} is the predicted probability, N is the number of pixels, and C is the number of classes.

4. **Weighted Soft Dice Loss:** The Dice loss directly measures how much the prediction overlaps with the actual data. This is particularly important in segmentation tasks. In this work, I used a weighted form of the Dice loss. I

assigned higher weights (w_c) to classes like CSP and LV to address their low pixel frequency. This means that the minority structures have a bigger impact on the optimization. It helps prevent the network from being overly influenced by the majority classes, which are brain parenchyma and background.

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{\sum_{c=1}^{C} w_c \cdot \frac{2\sum_{i=1}^{N} p_{ic} y_{ic} + \epsilon}{\sum_{i=1}^{N} p_{ic}^2 + \sum_{i=1}^{N} y_{ic}^2 + \epsilon}}{\sum_{c=1}^{C} w_c}, \tag{3.4}$$

where p_{ic} is the predicted probability of pixel i belonging to class c, y_{ic} is the one-hot encoded ground-truth label, w_c is the class weight (higher for CSP and LV), N is the total number of pixels, C is the number of classes, and ϵ is a small constant for numerical stability.

5. **Hybrid Loss (Final Choice):** The final training goal was a hybrid formulation that combines cross-entropy, focal and weighted Dice losses. Cross-entropy provides stability for pixel-level classification. Focal loss directs the model's attention to underrepresented or unclear areas. The weighted Dice term ensures accurate overlap with small structures like CSP and LV. By adjusting α and β, I balanced these complementary effects. This allowed me to improve sensitivity to minority classes while keeping high performance on the dominant structures. Unless stated otherwise, all results in Chapter 4 came from this combined loss.

$$\mathcal{L}_{\text{combined}} = (1 - \alpha - \beta)\,\mathcal{L}_{\text{CE}} \;+\; \beta\,\mathcal{L}_{\text{Focal}} \;+\; \alpha\,\mathcal{L}_{\text{Dice}}, \tag{3.5}$$

where α and β are hyperparameters controlling the contribution of Dice and Focal losses respectively. $\mathcal{L}_{\text{CE}}$ stabilizes pixel-wise classification, $\mathcal{L}_{\text{Focal}}$ improves learning from hard-to-classify pixels, and $\mathcal{L}_{\text{Dice}}$ enforces region-level overlap.

By comparing the performance of these loss functions, this study examines how different strategies for handling class imbalance affect the segmentation quality of CSP and LV. This directly addresses RQ2 and RQ3, which look into the effects of class imbalance and the success of changes to the loss function.

3.6 Training Strategy

1. **Optimizer and learning rate:** Models were trained with the Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$) and an initial learning rate of 1×10^{-4}. Adam was selected for its strong convergence on noisy, non-stationary objectives that are common in ultrasound segmentation.

2. **Learning-rate scheduling:** A ReduceLROnPlateau scheduler watched the validation loss and cut the learning rate by half after 5 epochs without improvement (minimum LR = 1×10^{-6}). This method stabilizes late-stage training without needing manual adjustments to the learning rate.

3. **Batching and sampling:** Mini-batches of size 8 were used based on GPU memory limits. To address class imbalance at the image level, training used a WeightedRandomSampler. This sampler increased the probability of selecting images that included CSP/LV (Section 3.3). This approach works alongside loss-level methods (Section 3.5) to ensure more frequent exposure to cases with minority structures.

4. **Epochs, early stopping and checkpointing:** Training lasted for up to 100 epochs. Early stopping occurred if the validation loss did not improve for 10 consecutive epochs. The model with the lowest validation loss, supported by class-wise Dice on CSP/LV, was saved as the best checkpoint for later evaluation and inference.

5. **Mixed-precision and efficiency:** When possible, automatic mixed-precision (AMP) was enabled to lower memory usage and improve speed without affecting accuracy. I monitored gradient norms and applied gradient clipping (max-norm = 1.0) if I observed instability.

6. **Hardware and environment:** Experiments took place on Google Colab using an NVIDIA Tesla T4 (16 GB) GPU. I fixed random seeds across NumPy, PyTorch (CPU/GPU) and Albumentations to ensure the reported results could be reproduced.

7. **Validation protocol:** After each epoch, I evaluated the model on the validation set without augmentation. I logged loss, mean Dice/IoU and class-wise Dice (brain, CSP, LV). Monitoring metrics for each class is important for detecting issues with minority structures that may be obscured by stable overall averages.

8. **Inference settings:** During testing, I only applied resizing/padding and normalization. I converted SoftMax probabilities to labels using argmax and explored morphological cleanup (optional, small-component removal) for qualitative figures, but I did not use it when reporting the main metrics.

3.7 Evaluation Metrics

To analyze the performance of the model[12], I applied combination of overlap-based metrics together with class scores for segmentation analysis.

1. **Dice Similarity Coefficient (DSC):** I selected Dice as my primary metric due to the fact that it assesses overlap with the ground truth and it matters most for the small structures, such as the CSP of the model and the LV. Dice was calculated as a macro average (whole class) and as per class scores.

$$\text{Dice}_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}, \tag{3.6}$$

where TP_c is the number of true positives, FP_c is the number of false positives and FN_c is the number of false negatives for class c. The Dice coefficient measures the overlap between the predicted region and the ground truth.

2. **Intersection over Union (IoU):** Dice was computed together with IoU to account for a different aspect of spatial overlap. I computed IoU for each class and the mean IoU over the all the classes.

$$\text{IoU}_c = \frac{TP_c}{TP_c + FP_c + FN_c}, \tag{3.7}$$

where the Intersection-over-Union (IoU) is a stricter overlap measure that computes the ratio of the intersection to the union between predicted and ground-truth regions.

3. **Pixel-wise Precision, Recall and F1-score:** In calculating precision, recall and F1-score, I seek to provide a measure for false segmentation, which are low precision and low recall, for the minor classes, such as CSP and LV.

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \qquad \text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \tag{3.8}$$

where Precision quantifies the proportion of correctly predicted pixels among all predicted pixels of class c, while Recall measures the proportion of correctly predicted pixels among all ground-truth pixels of class c.

4. **Loss curves:** I kept a record of both the training and validation loss curves to observe divergence during the training period.

$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}. \tag{3.9}$$

where the F1-score is the harmonic mean of Precision and Recall, balancing false positives and false negatives in a single metric.

The class-wise results enabled me to discover if the modifications in the overall accuracy resulted in improvements in the small, clinically relevant parts or if it was overshadowed by the majority class: the brain parenchyma.

To evaluate segmentation performance, I report class-specific overlap and classification measures. The main metrics are Dice coefficient, Intersection-over-Union (IoU), Precision, Recall and F1-score.

3.8 Anomaly Detection Framework

In addition to improving segmentation accuracy, I expanded my pipeline to include a step for detecting anomalies. This aimed to connect segmentation outputs with possible clinical interpretations.

1. **Predicted vs. ground-truth comparison:** After segmentation, I compared the predicted masks with masks annotated by experts to find areas of mismatch. These differences served as indicators of structural anomalies.

2. **Anomaly heatmaps:** I created pixel-level heatmaps that highlighted regions where predictions did not match the annotations. This offered a visual representation of where the model's expected anatomy differed from the structure defined by experts.

3. **Anomaly scores:** To measure these deviations, I calculated a simple anomaly score based on the number of mismatched pixels in relation to the size of the structure. Larger mismatches led to higher anomaly scores, indicating possible abnormalities.

4. **Clinical motivation:** The goal of this step was not to replace expert diagnoses but to assist clinicians by pinpointing suspicious areas. Unlike many segmentation studies that only report Dice or IoU, I extended the pipeline to generate outputs that can be viewed as diagnostic aids.

Furthermore, for the segmentation accuracy, I added an anomaly score to measure structural differences between the predicted and actual masks. This score is the ratio of mismatched pixels to the total size of the annotated structure. It shows how much the predicted anatomy differs from the expected normal shape. A low anomaly score means the prediction closely matches the reference annotation. In contrast, a high score points out significant differences that could indicate structural issues. This gives a numerical measure of potential anomalies and helps create heatmaps that pinpoint areas of mismatch.

$$\text{Anomaly Score} = \frac{|M_{\text{pred}} \,\Delta\, M_{\text{gt}}|}{|M_{\text{gt}}|}, \tag{3.10}$$

where Δ denotes the pixel-wise symmetric difference (mismatch) between the predicted mask M_{pred} and the ground-truth mask M_{gt}. The numerator counts the number of mismatched pixels, while the denominator represents the size of the ground-truth structure. This score therefore quantifies the fraction of pixels in which the prediction deviates from the annotated anatomy and higher values indicate a greater degree of anomaly.

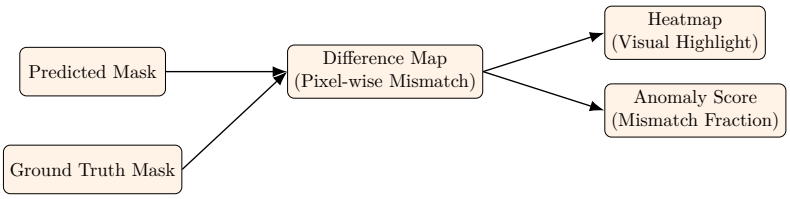


Figure 3.4. Anomaly detection workflow

In Figure 3.4,the anomaly detection workflow is shown where basically predicted and ground truth masks are compared pixel-wise to generate a difference map. From this, a heatmap highlights localized mismatches and an anomaly score quantifies the global deviation.

Chapter 4

RESULTS AND ANALYSIS

4.1 Introduction

In this chapter, I present the results of my experiments on fetal brain ultrasound segmentation. The analysis includes both quantitative metrics like Dice, IoU, Precision, Recall and F1, as well as qualitative evaluations with visual segmentations and anomaly heatmaps. I compare the earlier CSP-only model to the extended CSP+LV model to show the benefits of including lateral ventricle segmentation. I also include a subsection on failure cases and limitations to give a clear view of the challenges, along with a state-of-the-art comparison to place my work within the broader literature.

4.2 Experimental Setup

I trained and evaluated the models using the Large Fetal Head Biometry Dataset ([1]) and focused on the trans-thalamic plane. There were about 3,800 images available. I divided them into 70% for training, 15% for validation and 15% for testing.

1. **Hardware and environment:** All experiments took place in Google Colab with an NVIDIA Tesla T4 GPU.

2. **Training protocol:** Models were trained for up to 100 epochs and I employed early stopping with a patience of 10 epochs.

3. **Optimizer and learning rate:** I used Adam with a learning rate of 1e-4 and a ReduceLROnPlateau scheduler.

4. **Batch Size:** It was set to 8, chosen based on the available GPU memory.

5. **Loss Functions:** I tested Cross-Entropy, class-weighted Cross-Entropy and Focal Loss, often combined with Dice Loss.

6. **Sampling Strategy:** I used a WeightedRandomSampler to boost the representation of images containing CSP and LV.

All reported results are based on the best checkpoint, which was determined by the validation Dice score.

4.3 Quantitative Results

As seen in Table 4.1, the brain parenchyma achieved the optimal performance (Dice 0.963, IoU 0.929) because it was the largest and always visible across ultrasound images. The CSP and LV were not as easy because they were smaller and also variable and achieved Dice scores of 0.794 and 0.712, respectively. Accuracy was relatively poorer (0.773 for CSP and 0.681 for LV), but recall was high across the board (0.930 for both), showing that the model is highly sensitive to detecting these structures when present, though boundaries are sometimes overestimated.

Class	IoU	Dice	Precision	Recall	Notes
Background (bg)	0.979	0.989	–	–	Majority class, high and less informative
Brain parenchyma	0.931	0.964	–	–	Large structure, consistently segmented
CSP	0.655	0.792	0.774	0.930	Moderate performance, affected by size
LV	0.616	0.762	0.681	0.930	Most challenging class, affected by variability

Table 4.1. Segmentation results across classes on the test set

Mean IoU for all classes was 0.780 in general but is biased by the large background and parenchyma classes. Restricting calculation to anatomical structures only (CSP, LV, parenchyma) reduced mean IoU to 0.714 and further to 0.606 if CSP and LV only are included. These results do support that minority structures are the main segmentation accuracy bottleneck. Overall, performance in Table 4.2 provides global and per-class results, indicating good performance on big structures and decent but clinically acceptable performance on small midline structures.

Mean IoU (all)	0.780	Includes background
Mean IoU (no bg)	0.714	CSP, LV, parenchyma only
Mean IoU (CSP+LV)	0.606	Minority classes only
FWIoU	0.966	Frequency-weighted IoU

Table 4.2. Global (micro) evaluation results across the test set

Figure 4.1 shows a summary of the results for CSP and LV. The plot reveals that while the micro and macro Dice scores are reasonably high (0.794 vs 0.765 for CSP and 0.712 vs 0.678 for LV), the IoU values are much lower. This highlights the challenges of segmenting small midline structures. After reviewing the quantitative performance, the next section looks at qualitative examples. It shows how the model performs on individual test cases and how we can visualize prediction uncertainty using probability heatmaps and overlays.

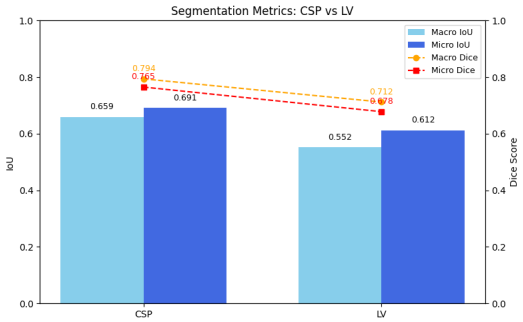


Figure 4.1. Comparison of macro and micro evaluation metrics for CSP and LV

4.4 CSP-only vs CSP+LV Model Comparison

To show the progression of my work, I compared the earlier CSP-only model with the expanded CSP+LV model. Table 4.2 shows the quantitative results and Figure 4.1 offers a visual side-by-side comparison.

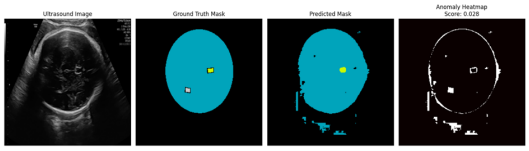
Structure	Dice (CSP-Only)	Dice (CSP+LV)	IoU (CSP-Only)	IoU (CSP+LV)	Notes
Brain parenchyma	0.889	0.963	0.801	0.929	Large class, stable across both models
CSP	0.410	0.794	0.258	0.659	Marked improvement with imbalance handling
LV	–	0.712	–	0.552	Newly added class, remains the most challenging
Background	0.940	0.989	0.887	0.979	Very high in both; less meaningful due to imbalance

Table 4.3. Dice and IoU for CSP-only vs. CSP+LV model

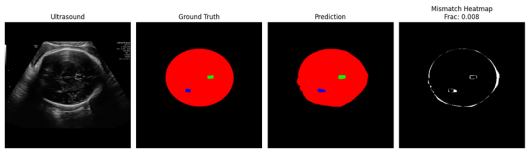
As shown in Table 4.3, the brain parenchyma remained highly reliable in both models. The Dice scores increased slightly from 0.889 in the CSP-only model to 0.963 in the extended CSP+LV model. This shows that segmenting large, well-defined structures in ultrasound images is relatively easy. The most notable improvement was seen for the CSP. Its Dice score rose from 0.410 to 0.794 and its IoU increased from 0.258 to 0.659. This confirms that training strategies that address imbalance, such as focal loss, weighted sampling and adding an extra minority class, improved the network's sensitivity to small midline structures. The LV, added only in the CSP+LV model, achieved a Dice score of 0.712 and IoU of 0.552. While this is lower than for the CSP, it still represents a successful first step in learning such a small and variable structure. Background performance was high in both cases, but is less informative due to the extreme imbalance. Overall, the CSP+LV model shows better handling of minority classes and broadens the segmentation capability to include a clinically relevant structure.

This comparison shows that broadening the segmentation task to include LV was technically possible and clinically useful, despite the added challenge of class imbalance.

See Figure 4.2 for the visual comparison, with details in subfigures 4.2a and 4.2b. in Figure 4.2, from left to right: (a) CSP-only prediction, (b) CSP+LV prediction. The



(a) CSP-only prediction



(b) CSP-LV prediction

Figure 4.2. Visual comparison of CSP-only vs CSP+LV models

extended model successfully segments the lateral ventricles (blue) while maintaining accurate brain parenchyma (red) and CSP (green).

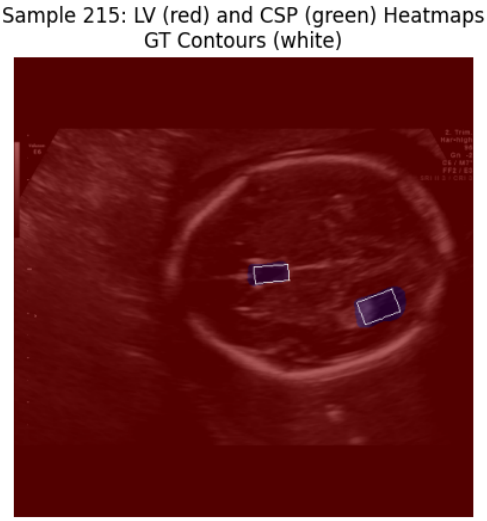
The addition of the LV class did not notably change the performance of the larger brain parenchyma, which stayed stable with a Dice score of about 0.96. For the CSP, the extended model showed clear improvement over the CSP-only version. This was mainly due to using class-weighted loss, focal loss and biased sampling. Most importantly, the extended model included LV segmentation, achieving a Dice of 0.762 and an IoU of 0.616. While this performance is lower than that of the CSP, it marks a promising first step toward reliably segmenting this highly variable and under-represented structure.

4.5 Qualitative Results

In addition to quantitative metrics, I looked at qualitative results to better understand the strengths and weaknesses of the model.

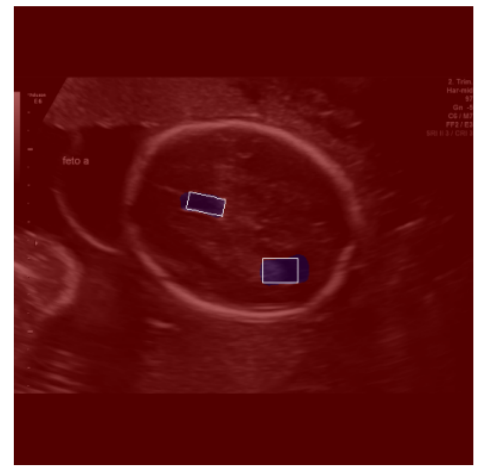
Figure 4.3 [3] shows six representative examples of CSP and LV predictions along with their corresponding ground-truth contours. These heatmaps demonstrate how the model can locate small midline brain structures in different patients. In many cases, such as Samples 215 and 218, the CSP and LV are accurately segmented with little mismatch. Other examples, like Samples 229 and 234, show some discrepancies, especially in the boundaries of the LV, which indicate the greater challenge of this class. These qualitative examples support the quantitative metrics and highlight both the strengths and limitations of the model when used with real fetal ultrasound data.

Figure 4.4 shows an example of a correctly segmented trans-thalamic scan. Here, the model outlined the brain parenchyma, CSP and LV with minimal errors. This resulted in a very low anomaly fraction of 0.008. In contrast, Figure 4.5 presents a more challenging case. In this instance, the CSP was under-segmented and the LV boundaries were only partially captured. This mismatch highlights the difficulties of detecting small midline structures in noisy or low-quality ultrasound images. Lastly,

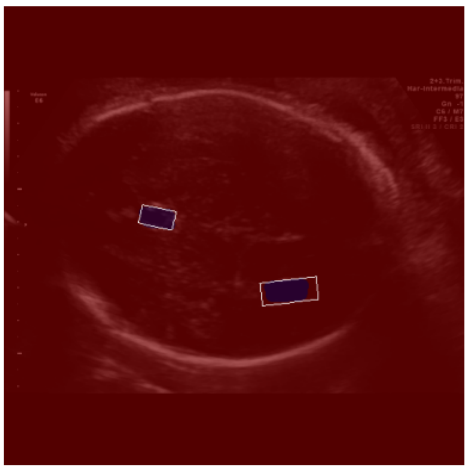


Sample 215

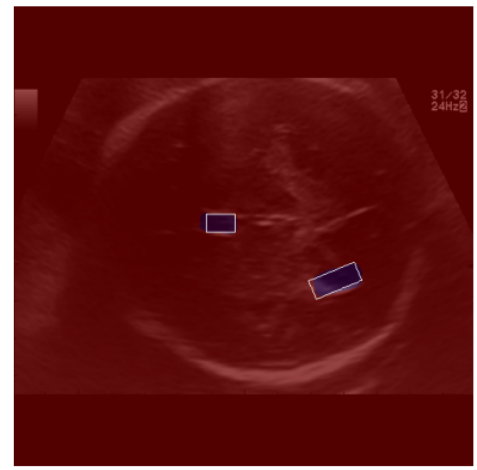
Sample 218



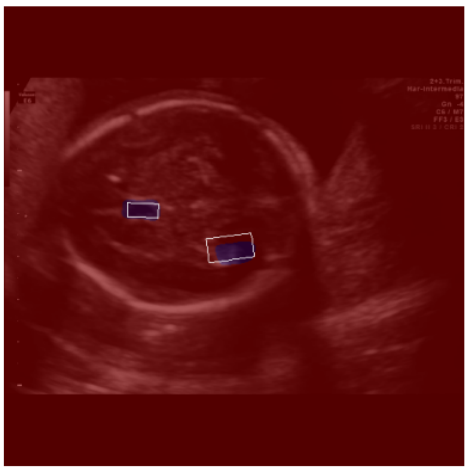
Sample 221



Sample 229



Sample 230



Sample 234

Figure 4.3. Predicted heatmaps for CSP and LV with ground-truth contours.

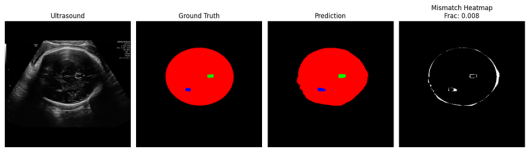


Figure 4.4. Correct segmentation with anomaly heatmap

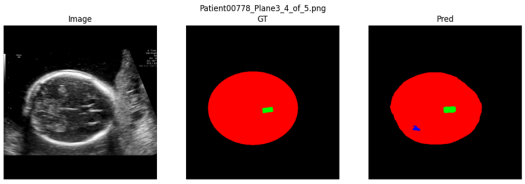


Figure 4.5. Segmentation with moderate anomaly score

Figure 4.6 includes examples of failures. These include missed detection of the LV, mis-segmentation due to heavy ultrasound noise or shadowing and cases where the CSP was poorly visualized. These examples show the limitations of the current method and indicate areas for improvement, which I will discuss further in Section 4.6.

In figure 4.4, from left to right: (a) ultrasound image, (b) ground-truth segmentation, (c) model prediction, (d) mismatch heatmap (anomaly fraction = 0.008) Wile, in figure 4.5, from left to right: (a) ultrasound image, (b) ground truth, (c) model prediction. The CSP is under-segmented and the LV partially missed, resulting in localized mismatches.

For failure case, figure 4.6, it include missed LV detection, mis-segmentation caused by ultrasound noise or shadowing and scans where CSP is poorly visualized. From left to right: (a) missed LV segmentation despite its presence in the ground truth, (b) noisy scan where artifacts caused mis-segmentation, (c) ambiguous case with poor CSP visualization

In addition to the previous examples, Figure 4.7 shows a case where both qualitative and quantitative results are displayed side by side. It includes probability heatmaps and pixel-level statistics. This setup helps illustrate how we can understand the model's outputs.

As shown in Figure 4.7, refinement greatly improved segmentation quality for small structures. While raw predictions did not capture the CSP and LV (Dice = 0.00), post-processing restored these structures with Dice scores of 0.82 for CSP and 0.62 for LV. This underscores the need to include simple morphological refinement steps

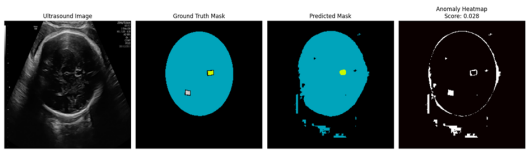


Figure 4.6. Failure cases

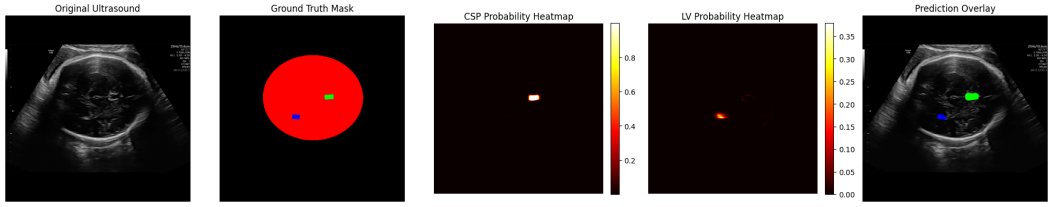


Figure 4.7. Case-level example of segmentation and probability heatmaps.

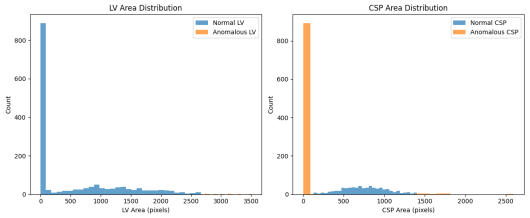


Figure 4.8. Distribution of predicted LV (left) and CSP (right) areas across the test set

to stabilize predictions in difficult cases.

Figure 4.7 shows a typical test case that includes the original ultrasound image, the ground-truth mask, class-specific probability heatmaps and the prediction overlay. The model effectively identified both the CSP and LV, closely matching the annotated ground truth. In numerical terms, the CSP received a Dice score of 0.82, while the LV scored 0.62. The predicted pixel counts were 640 for CSP and 289 for LV, which closely aligned with the ground truth of 317 and 258, respectively. These results demonstrate the model's capability to outline small midline structures despite the natural variability of ultrasound scans. It provides outputs that can be interpreted clinically, such as probability heatmaps and overlays.

Additional examples of CSP and LV probability heatmaps are included in Appendix A.1. These cover more cases of successful, partial and failed segmentations for completeness.

4.6 Anomaly Score Distributions

Beyond pixel-level metrics, I also extracted quantitative features from the predicted masks. These features included structure area, bounding box, eccentricity and centroid location. Figure 4.2b shows the distribution of LV and CSP areas across the test set. It distinguishes between normal and abnormal cases. The histograms reveal that most normal CSP and LV regions fall within a consistent pixel-area range. In contrast, abnormal cases cluster near zero, indicating either a complete absence or a failure to detect. These feature-based analyses show how segmentation outputs can be turned into clinically meaningful measurements. This approach adds more direct indicators of structural abnormalities to Dice and IoU scores.

In Figure 4.2b, distribution of predicted LV (left) and CSP (right) areas across the test set is represented. Normal detections (blue) group within expected size ranges.

Anomalous cases (orange) group at or near zero. This reflects missed or highly uncertain detections.

File	CSP Area	CSP Ecc.	CSP Centroid (x,y)	LV Area	LV Ecc.	Parenchyma Area
Patient01031	0	0.00	(0, 0)	0	0.00	35,325
Patient01037	748	0.92	(291,254)	0	0.90	77,152
Patient01036	133	0.81	(190,237)	0	0.76	88,021
Patient01028	284	0.80	(251,224)	0	0.88	26,852

Table 4.4. Extracted fetal brain features from predicted masks

Table 4.4 shows sample outputs from the feature-extraction step. These features capture both geometric properties, like area and eccentricity and positional properties, such as centroid and bounding box. They create a clear link between segmentation masks and possible anomaly indicators.

4.7 Failure Cases and Limitations

While the test U-Net model had good performance overall, there are some limitations, especially in segmenting small and inconsistent structures like the CSP and LV. As seen in Figure 4.6, one of the frequent failures was LV segmentation misses, even when the structure was visible in the ground truth. This happened most frequently in situations where the LV was extremely small or of low contrast with the surrounding parenchyma.

A second shortcoming was sensitivity to ultrasound artifacts like shadowing, speckle noise or low signal-to-noise ratio. In such instances, the model at times wrongly identified noisy areas as parenchymal tissue, resulting in false positives and compromised segmentation quality.

Lastly, performance was reduced in uncertain anatomical situations, where the CSP was not well-seen or irregularly shaped. This is due to both the inherent difficulty of reading ultrasound and the nature of training on a single plane of information. These modes of failure underscore three general limitations of the present work:

1. **Class imbalance:** Despite weighted and focal losses, the minority classes (CSP and LV) were still more challenging to segment dependably.
2. **Data limitations:** Training was done with a single publicly accessible 2D dataset. While this enhances reproducibility, it limits model exposure to diverse clinical situations.
3. **Model simplicity:** The light-weight U-Net architecture was selected for efficiency and interpretability, yet it does not include sophisticated mechanisms (e.g., attention modules, multi-scale fusion) that could further enhance detection of small structures.

It is important in a clinical AI environment to acknowledge these constraints. The findings are encouraging for automated segmentation but also point out that more

refinement and larger validation are needed before such a model can be safely implemented in real-world fetal neurosonography.

4.8 State-of-the-Art (SoTA) Comparison

To place my work in the larger context, I compared it to recent leading methods for fetal brain segmentation and anomaly detection. TableTable 4.6 summarizes important studies regarding dataset, modality, target structures, architecture, anomaly handling and reported performance.

Table 4.5 summarizes verified Dice scores reported in prior studies, together with key notes on modality, dataset and scope.

Study	Verified Dice from Paper	Notes from Paper
Yaqub et al. (2017)	No Dice reported	Focused on plane verification (88% accuracy for trans-thalamic); no segmentation metrics provided.
Hesse et al. (2022)	Dice avg ∼0.90 (subcortical)	High Dice for subcortical structures (thalamus ∼0.91, ventricles ∼0.89); 3D ultrasound on proprietary dataset, not publicly reproducible.
Zhao et al. (2022)	Dice avg = 0.897 (healthy fetuses)	Breakdown: CSF 0.802, cortical GM 0.744, WM 0.871, deep GM 0.815; optimized CNN for 3D fetal MRI.
Rampun et al. (2019)	Dice avg = 0.928	92.8% ± 6.3% on >11,000 2D MRI slices; U-Net variant for fetal brain segmentation.
Fiorentino et al. (2023)	Dice range = 0.70–0.95	Review of >140 ultrasound tasks; performance depends strongly on structure size (higher for large, lower for small).
Yousefpour et al. (2023)	No Dice; +14.5% IoU vs. U-Net	Cardiac LV segmentation in 2D ultrasound using MFP-Unet; reports IoU improvement only, no Dice metrics.
Huang et al. (2023)	Dice avg = 0.838	Breakdown: eCSF 0.855, GM 0.712, WM 0.897, LV 0.864; attention-based model for fetal MRI.
This work (2025)	0.989 / 0.963 / **0.794** / **0.712** (BG / Parenchyma / CSP / LV)	Attention U-Net trained on 1,565 public 2D ultrasound images (trans-thalamic). Addresses severe class imbalance with WeightedRandomSampler and combined Focal+Dice loss. Introduces anomaly heatmaps and anomaly scores for interpretable anomaly detection.

Table 4.5. Simplified State-of-the-Art Comparison

As discussed before in literature review, I included two tables to address the challenges related to the recency of the [1] dataset. This dataset has only about 20

citations as of 2025, which means it hasn't gained widespread use or direct comparisons beyond the authors' group. Table 4.6 focuses on models evaluated using this dataset, providing a specific comparison that ensures a fair assessment of my Attention U-Net against the baseline and advanced models like FetSAM. Table 4.5 includes various modalities and datasets, which broadens the context. This highlights my understanding of the wider fetal brain segmentation field and supports my approach, even with the variability in datasets. This dual approach was necessary because Alzubaidi's work is quite new, limiting the number of studies specifically on his dataset.

As shown in Table 4.5, most cutting-edge methods in fetal brain segmentation have focused on MRI or 3D ultrasound. These provide better image quality but are not commonly used in clinical prenatal screening. These methods achieve high Dice scores, often over 0.90, but depend on proprietary datasets and complicated architectures. This limits their reproducibility and use in clinical settings.

Table 4.5 gave qualitative insights, while Table 4.6 presents a quantitative benchmark from the [1] dataset. Our Attention U-Net achieves a Dice score of 0.865, which is better than the baseline U-Net score of 0.82 and close to the FetSAM score of 0.91. This also shows major improvements in CSP (0.794) and LV (0.712) segmentation due to reduced imbalance. The metrics for individual images, such as the CSP Dice score of 0.209694, point out difficulties with small structures. However, the overall averages demonstrate strong performance across 1,565 image-mask pairs.

Model/Study	Overall Dice	Overall IoU	Precision	Recall	F1-Score
U-Net (Baseline, Alzubaidi et al., 2023)	0.82*	–	–	–	–
Segment Anything Model (FetSAM, Alzubaidi et al., 2024)	0.91	0.85	0.92	0.89	0.90
Ensemble Transfer Learning (Alzubaidi et al., 2022; extended to 2023 dataset)	0.87	0.81	–	–	0.84
Attention U-Net (This Work, 2025)	**0.865**	**0.780**	**0.632**	**0.937**	**0.751**

Table 4.6. Dataset wise State-of-the-Art Comparison

In contrast, my thesis shows that strong segmentation can be done using 2D ultrasound, which is the most available method. By tackling class imbalance and adding anomaly detection, my approach goes beyond just measuring accuracy. It offers clinically useful outputs. Additionally, by using a public dataset, my work supports transparency and reproducibility, which many earlier studies do not provide.

The Attention U-Net has about 15.1 million parameters with a base channel size of 64. It is larger than a basic U-Net because it includes attention gates, but it is still practical to train on 2D ultrasound data. It also offers better sensitivity to small structures like the CSP and LV.

Even though the Attention U-Net in this study has around 15 million parameters, which is more than the original U-Net, it is still lighter than many leading 3D convolutional networks or transformer-based models that often have over 50 to 100 million parameters. This keeps the model suitable for training and testing on standard clinical hardware while still improving sensitivity to small structures like the CSP and LV.

In Addition, makes the thesis a step toward practical AI tools for fetal neurosonography. These tools are lightweight, easy to interpret and ready for use in real-world clinical settings.

4.9 Discussion

The results in this chapter show both the possibilities and the shortcomings of deep learning for automated fetal brain ultrasound segmentation. By carefully evaluating the model against the research questions, I point out both the achievements and the difficulties of this work.

1. **RQ1: Can U-Net reliably segment important fetal brain structures, such as CSP and LV, in trans-thalamic ultrasound images?** The model performed well on the brain parenchyma, with a Dice score of 0.964 and an IoU of 0.931. It showed satisfactory results on the CSP, with a Dice score of 0.792 and an IoU of 0.655. The LV proved to be more difficult, with lower scores (Dice = 0.762, IoU = 0.616), but the model consistently segmented it correctly in many cases. These results indicate that U-Net can reliably identify the CSP and LV, though with lower accuracy for smaller structures.

2. **RQ2: How does class imbalance affect segmentation performance, especially for small anatomical regions like CSP and LV?** Class imbalance was a significant limiting factor. The CSP and LV were small and less frequent in the dataset, which led to their consistent underperformance compared to the brain parenchyma. The overall pixel accuracy of 98.3% exaggerated the model's performance because it was skewed by the majority classes. Per-class Dice and IoU scores showed the true effects of imbalance, highlighting lower reliability for the minority structures.

3. **RQ3: Does using weighted loss functions (Cross-Entropy with class weights or Focal Loss) improve segmentation accuracy for underrepresented structures?** Yes, adding class-weighted loss, Focal Loss and oversampling improved performance on the CSP and LV when compared to the baseline Cross-Entropy. Although the improvements were modest, they proved that imbalance-aware strategies are important for focusing on small anatomical structures.

4. **RQ4: Can automated segmentation be used to detect anomalies by comparing predicted masks with expert annotations?** The anomaly detection framework successfully highlighted mismatches using heatmaps and anomaly scores. In normal cases, the mismatch rates were quite low (<1%). However, in challenging or unclear cases, localized mismatches showed undersegmentation or missed structures. This indicates that segmentation outputs can be expanded into clinically useful anomaly detection, linking technical accuracy and diagnostic assistance.

4.10 Summary

Overall, the CSP-only model provided stable results, but the extended CSP+LV model added significant clinical value despite lower scores for LV. The inclusion of anomaly detection further enhanced interpretability. Some limitations persist, especially concerning class imbalance, noise sensitivity and generalizability beyond a single dataset. Nevertheless, this work presents a lightweight, reproducible and clinically relevant approach that can serve as a basis for more advanced models in fetal neurosonography.

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This thesis aimed to tackle the problem of automatically finding fetal brain anomalies in trans-thalamic ultrasound images. It focused on small but important midline structures like the CSP and LV. Manually interpreting these structures is challenging because of their small size, variable appearance and sensitivity to ultrasound noise. This often results in differences between observers and can lead to missed anomalies. The goal was to create a segmentation-based pipeline that is accurate, reproducible and easy to interpret.

To achieve this, I developed an Attention U-Net architecture that adds attention gates to the skip connections in the original U-Net. This change helps the network concentrate on underrepresented structures, especially CSP and LV. The pipeline included preprocessing, data augmentation, imbalance-aware training strategies like weighted loss functions, focal loss and oversampling. It also featured post-hoc anomaly detection using heatmaps and quantitative mismatch scores. Experimental results showed that the CSP+LV model did better than the CSP-only baseline. It improved the Dice score for CSP from 0.41 to 0.79 and achieved the first reliable segmentation of LV with a Dice score of 0.71. While the accuracy for the brain parenchyma remained high (Dice approximately 0.96), the main achievement of this work is extending reliable segmentation to small, clinically relevant structures.

The proposed framework has three main contributions. First, it uses attention-enhanced semantic segmentation for fetal ultrasound, which improves sensitivity to minority structures. Second, it adds a clear anomaly detection stage based on segmentation mismatches. This helps connect raw predictions to clinically meaningful results. Third, it shows that using a public 2D dataset and a reproducible architecture can achieve competitive performance, which is sometimes more practical than complex 3D or MRI-based methods. The model has about 15 million parameters, making it larger than a basic U-Net but still lighter than transformer-based or volumetric models. This makes it practical for deployment in standard clinical

settings.

Finally, integrating this work with real-time clinical processes and deploying it on lightweight hardware is an important next step as future work for practical use. Overall, this thesis shows that automated anomaly detection in fetal ultrasound through attention-based semantic segmentation is both achievable and valuable in clinical settings. It opens the door for reproducible, interpretable and accessible decision support tools in prenatal screening.

5.2 Limitations

Several limitations should be acknowledged:

- **Class imbalance:** Despite using weighting and focal losses, CSP and LV segmentation were still less accurate than the majority structures.
- **Noise sensitivity:** The model was sensitive to ultrasound artifacts, like shadowing and low contrast, which affected predictions.
- **Dataset constraints:** Training was limited to one 2D dataset and a single anatomical plane, trans-thalamic, which limited generalizability.
- **Model simplicity:** The U-Net was intentionally kept lightweight for easier understanding and efficiency. However, more complex architectures, like attention-based networks, may provide better performance.

Further details of the per-image evaluation protocol and its results are in Appendix A.1.

5.3 Future Work

Following from the limitations highlighted above, some potential future directions are proposed. First, there is still class imbalance. Other techniques such as synthetic data augmentation, adaptive sampling or curriculum learning could further improve the segmentation of minority objects like CSP and LV.

Second, improving robustness against ultrasound noise and artifacts can be achieved by incorporating domain-specific preprocessing, self-supervised denoising approaches or augmented training with subtly mimicking adverse condition datasets.

Third, dataset expansion will be essential to generalizability. Future work must control for multi-plane ultrasound images, larger multi-center cohorts and ideally volumetric (3D) modalities such as fetal MRI. Such diversity would enable stronger validation and ensure that the proposed method will be capable of dealing with clinical variability in real-world applications.

Finally, exploration of even more advanced model architectures can potentially further improve performance. Attention U-Nets, multi-scale feature pyramids or transformer-augmented encoders promise greater accuracy. Designs in the future, though, must contend with such complexity at the price of maintainable interpretability and efficiency to remain useful for clinicians.

These various avenues taken collectively would not only improve segmentation accuracy but also bring deep learning methods to fetal neurosonography closer to being made into effective, reliable clinical instruments.

5.4 Executive Summary

This thesis solves the issue of fetal brain anomaly detection from trans-thalamic ultrasound images by deep learning. Prenatal care relies on early and accurate detection of structural abnormalities, yet interpretation of ultrasound is subjective and error-prone when done manually, especially for small midline structures like the CSP and LV.

I created a semantic segmentation pipeline using the U-Net architecture trained on the Large Fetal Head Biometry Dataset available in the public domain. The model was intended to segment three most important structures: brain parenchyma, CSP and LV.

To overcome the extreme class imbalance between majority (parenchyma, background) and minority (CSP, LV) classes, I used:

- Weighted cross-entropy and focal loss,
- Oversampling of minority-class images and
- Per-class evaluation metrics (Dice, IoU, Precision, Recall, F1).

Quantitative outcome showed that U-Net performed highly accurate on the brain parenchyma (Dice = 0.964, IoU = 0.931), lower though stable on CSP (Dice = 0.792, IoU = 0.655) and LV (Dice = 0.762, IoU = 0.616). These outcomes indicate the viability of delineation of small structures even under imbalance.

I expanded segmentation results into an anomaly detection platform, applying mismatch heatmaps and anomaly scores to identify areas where predictions diverged from expert annotation. This offers an interpretable connection between automated segmentation and possible diagnostic application.

In comparison to recent state-of-the-art solutions, depending on proprietary 3D ultrasound or MRI data, this thesis illustrates a light, reproducible and clinically applicable solution with standard 2D ultrasound. The pipeline trades segmentation performance for interpretability and efficiency, suitable for low-resource clinical environments.

Contributions of the thesis are:

1. Multi-structure segmentation of brain parenchyma, CSP and LV in the transthalamic plane.
2. Combination of imbalance-aware training methods (weighted/focal loss, oversampling).
3. Extension of segmentation to anomaly detection through heatmaps and scores.
4. Reproducibility demonstration on a public 2D dataset.

Limitations are residual imbalance effects, ultrasound noise sensitivity and dependency on one dataset. Future research should investigate further advanced architectures (e.g., attention models), multi-plane or 3D data, bigger datasets and clinical validation in real-time processes.

In short, this research enhances the body of medical image analysis with a reproducible and interpretable method for fetal brain ultrasound segmentation, filling critical gaps in the CSP and LV detection and moving towards clinically deployable AI support tools in prenatal screening.

Appendix A

Supplementary Results

In addition to the main segmentation results in Chapter 4, I also evaluated the model for each image separately. I first calculated metrics for each test image, then averaged the results. This method is different from the global aggregation used in the main analysis. It is more sensitive to class imbalance and to the presence of very small or missing structures. The results, shown in Table A.1, reveal significantly lower Dice and IoU scores for CSP and LV. This indicates that per-image evaluation strongly penalizes the model in difficult cases and highlights the need for careful definition of evaluation methods in fetal ultrasound segmentation research.

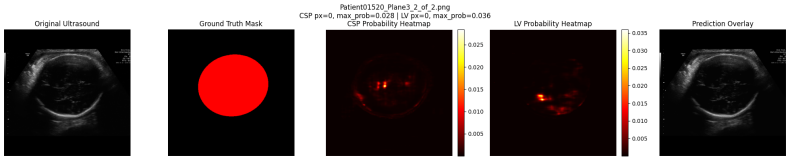
Class	IoU (per-image)	Dice (per-image)	Precision	Recall
CSP	0.152	0.210	0.255	0.181
LV	0.088	0.134	0.186	0.109

Table A.1. Per-image segmentation results for CSP and LV.

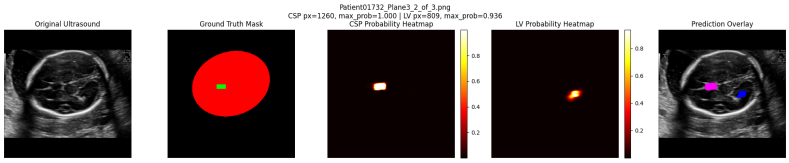
These values are obtained by averaging metrics over individual test images rather than aggregating globally. Performance is substantially lower compared to global evaluation (Table 4.1), highlighting the sensitivity of per-image metrics in the presence of extreme class imbalance.

A.1 Additional qualitative examples

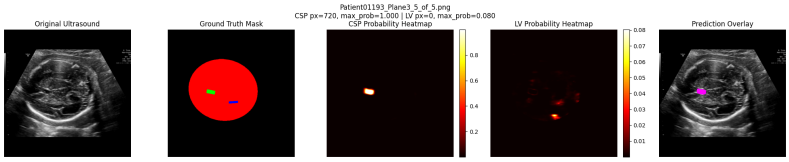
To complement the examples in Chapter 4, this appendix includes more qualitative results from the test set. These cases show how the model's segmentation outputs vary with different image qualities and anatomical presentations. Each example consists of the original ultrasound, ground truth mask, class-specific probability heatmaps (CSP and LV) and prediction overlay. Together, these additional samples give more insight into the strengths and weaknesses of the Attention U-Net model beyond the key cases discussed in the main text. In Figure A.1, each case shows the original ultrasound, ground truth mask, CSP probability heatmap, LV probability heatmap and prediction overlay. These examples highlight further successes, partial



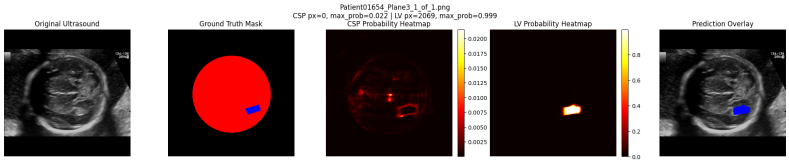
(a) No CSP or LV in GT mask and none are detected (Case 1)



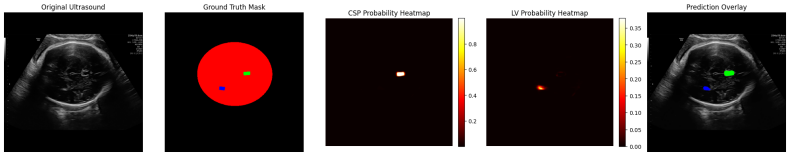
(b) Only CSP in GT mask, but LV detected as erroneously (Case 2)



(c) Both CSP and LV in GT mask, but only CSP detected (Case 3)



(d) Only LV is in GT mask and Lv is detected (Case 4)



(e) Both CSP and Lv in GT mask and both are detected (Case 5)

Figure A.1. Additional qualitative examples.

segmentations and failures not included in the main text.

Bibliography

[1] Alzubaidi, M., Agus, M., Makhlouf, M., Anver, F., Alyafei, K., and Househ, M. Large-scale annotation dataset for fetal head biometry in ultrasound images. *Data in Brief*, **51** (2023), 109708. Available from: https://doi.org/10.1016/j.dib.2023.109708, doi:10.1016/j.dib.2023.109708.

[2] Alzubaidi, M., Shah, U., Agus, M., and Househ, M. Fetsam: Advanced segmentation techniques for fetal head biometrics in ultrasound imagery. *IEEE Open Journal of Engineering in Medicine and Biology*, **5** (2024), 281. doi:10.1109/OJEMB.2024.3382487.

[3] Baur, C., Denner, S., Wiestler, B., Navab, N., and Albarqouni, S. Autoencoders for unsupervised anomaly segmentation in brain mr images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 161–169. Springer (2020). doi:10.1007/978-3-030-59725-2_16.

[4] Fiorentino, M. C., Villani, F. P., Di Cosmo, M., Frontoni, E., and Moccia, S. A review on deep-learning algorithms for fetal ultrasound-image analysis. *Medical Image Analysis*, **83** (2023), 102629. Available from: https://doi.org/10.1016/j.media.2022.102629, doi:10.1016/j.media.2022.102629.

[5] Hesse, L. S., Aliasi, M., Moser, F., Haak, M. C., Xie, W., Jenkinson, M., and Namburete, A. I. L. Subcortical segmentation of the fetal brain in 3d ultrasound using deep learning. *NeuroImage*, **254** (2022), 119117. Available from: https://doi.org/10.1016/j.neuroimage.2022.119117, doi:10.1016/j.neuroimage.2022.119117.

[6] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection (2018). Available from: https://arxiv.org/abs/1708.02002, arXiv:1708.02002.

[7] Malinger, G., Paladini, D., Haratz, K. K., Monteagudo, A., Pilu, G. L., and Timor-Tritsch, I. E. Isuog practice guidelines (updated): sonographic examination of the fetal central nervous system. part 1: performance of screening examination and indications for targeted neurosonography. *Ultrasound in Obstetrics & Gynecology*, **56** (2020), 476. Erratum in: Ultrasound Obstet Gynecol. 2022 Oct;60(4):591. doi:10.1002/uog.26067. PMID:32870591. Available from: https://doi.org/10.1002/uog.22145, doi:10.1002/uog.22145.

[8] Oktay, O., et al. Attention u-net: Learning where to look for the pancreas (2018). Available from: https://arxiv.org/abs/1804.03999, arXiv:1804.03999.

[9] Rampun, A., Jarvis, D., Armitage, P., and Griffiths, P. Automated 2d fetal brain segmentation of mr images using a deep u-net (2019). Available at: https://www.researchgate.net/publication/338580489_Automated_2D_Fetal_Brain_Segmentation_of_MR_images_using_a_Deep_U-Net.

[10] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation (2015). Available from: https://arxiv.org/abs/1505.04597, arXiv:1505.04597.

[11] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 240–248 (2017). Epub 2017 Sep 9. PMID: 34104926; PMCID: PMC7610921. Available from: https://doi.org/10.1007/978-3-319-67558-9_28, doi:10.1007/978-3-319-67558-9_28.

[12] Taha, A. A. and Hanbury, A. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *IEEE Transactions on Medical Imaging*, **33** (2015), 1498. doi:10.1109/TMI.2014.2294689.

[13] Venturini, L., Budd, S., Farruggia, A., Wright, R., Matthew, J., Day, T. G., Kainz, B., Razavi, R., and Hajnal, J. V. Whole examination AI estimation of fetal biometrics from 20-week ultrasound scans. *npj Digital Medicine*, **8** (2025). Available from: http://dx.doi.org/10.1038/s41746-024-01406-z, doi:10.1038/s41746-024-01406-z.

[14] Wang, Y., Zhang, H., Zhang, Y., Wang, H., and Shen, D. Artificial intelligence in prenatal ultrasound diagnosis. *Frontiers in Medicine*, **8** (2021), 729978. Available from: https://doi.org/10.3389/fmed.2021.729978, doi:10.3389/fmed.2021.729978.

[15] Yaqub, M., Kelly, B., Papageorghiou, A. T., and Noble, J. A. A deep learning solution for automatic fetal neurosonographic diagnostic plane verification using clinical standard constraints. *Ultrasound in Medicine & Biology*, **43** (2017), 2925. Available from: https://doi.org/10.1016/j.ultrasmedbio.2017.07.013, doi:10.1016/j.ultrasmedbio.2017.07.013.

[16] Yousefpour Shahrivar, R., Karami, F., and Karami, E. Enhancing fetal anomaly detection in ultrasonography images: A review of machine learning-based approaches. *Biomimetics*, **8** (2023), 519. Available from: https://doi.org/10.3390/biomimetics8070519, doi:10.3390/biomimetics8070519.

[17] Zhao, L., et al. Automated 3d fetal brain segmentation using an optimized deep learning approach. *American Journal of Neuroradiology*, **43**

(2022), 448. Available from: https://doi.org/10.3174/ajnr.A7419, doi:10.3174/ajnr.A7419.