Deep Learning for Ultrasound Tongue Imaging: Towards Robust, Interpretable, and Deployable Assessment of Speech Sound Disorders

Saja Al Ani

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

James Watt School of Engineering
College of Science and Engineering
University of Glasgow

March 2026

Abstract

Speech sound disorders (SSDs) are among the most common developmental communication difficulties in childhood, with long-term consequences for intelligibility, literacy, and psychosocial well-being. Clinical assessment relies primarily on auditory–perceptual judgement, which, although effective, is subjective and provides limited insight into underlying articulatory mechanisms. Ultrasound tongue imaging (UTI) offers a safe, non-invasive method for visualising tongue movement during speech; however, its interpretation remains challenging due to image noise, speaker and acquisition variability, and the high cost of expert annotation. This thesis examines the systematic adaptation of deep learning (DL) to address three core challenges in automated UTI analysis: **(C1)** data variability and generalisability limitations, **(C2)** data scarcity and annotation inefficiency, and **(C3)** lack of interpretability and clinical usability.

Reproducible baseline deep neural network (DNN) models are first established for phonetic classification from raw UTI, quantifying generalisation limitations under speaker-independent evaluation. A novel multi-input FusionNet architecture is then introduced, combining raw ultrasound frames with texture-based representations to improve cross-speaker robustness. A two-stage conditional generative adversarial framework is proposed for field-of-view (FoV) standardisation and tongue region enhancement, improving image consistency and classification performance across domains. To address data scarcity, a cost-focused framework integrates statistical power-curve modelling with active learning to optimise annotation effort, achieving substantial reductions in required labelled data while maintaining clinically meaningful accuracy. Model interpretability is examined using an explainable AI (XAI) technique to assess how image representation and standardisation influence network attention and anatomical relevance. Finally, the feasibility of clinical translation is demonstrated through a prototype web-based deployment system for real-time inference and visualisation.

Collectively, this work presents an integrated DL framework that advances robustness, data efficiency, interpretability, and deployment for ultrasound-based speech assessment, contributing toward objective and scalable clinical decision-support tools for SSDs.

Keywords

Ultrasound Tongue Imaging, Speech Sound Disorders, Deep Learning, Explainable Artificial Intelligence, Generative Adversarial Networks, Speech Assessment.

Acknowledgements

I would like to begin by expressing my deepest gratitude to my supervisor, Dr Ahmed Zoha. His exceptional mentorship and unwavering support have been fundamental to both my academic development and personal growth throughout this doctoral journey. His thoughtful guidance, insightful feedback, and constant encouragement challenged me to think more critically and to strive for excellence in my work. I am equally grateful to Dr Joanne Cleland for grounding this research in clinical relevance, for encouraging the exploration of new ideas, and for pushing me to take on challenges I might not have otherwise pursued.

I am sincerely grateful to the Future Ultrasonic Engineering Centre for Doctoral Training (FUCE) for providing the environment, training, and interdisciplinary perspective that made this work possible. Being part of FUSE broadened my thinking beyond a single discipline and allowed me to explore the intersection of engineering, artificial intelligence, and clinical speech science.

My thanks also extend to colleagues who contributed through discussion, feedback, and encouragement along the way. Research is never a solitary endeavour, and I have been fortunate to be surrounded by people willing to share their ideas, time, and expertise.

I would like to express my heartfelt thanks to my family for their constant support and belief in me. Their encouragement carried me through every challenge and milestone.

Finally, I owe special thanks to my husband, Mohammed. Your unwavering support, patience, and belief in me have been the foundation of this entire journey. Through long writing days, experimental setbacks, deadlines, and the inevitable self-doubt that comes with a PhD, you stood beside me with strength and calm. I could not have done this without you.

To my sons, Saif and Muneer, you are my greatest motivation. Your love, patience, and bright energy gave me strength during the most demanding stages of this journey. This achievement is for you, and because of you.

Author’s declaration

University of Glasgow

College of Science & Engineering

Statement of Originality

Name: Saja Al Ani

Registration number: XXXXXXXX

I certify that the thesis presented for examination for the degree of PhD at the University of Glasgow is my own work, except where due acknowledgement is made to the work of others. In such cases, the extent of any jointly undertaken work is clearly identified. I further confirm that the thesis has not been edited by any third party beyond what is permitted under the University’s PGR Code of Practice.

I declare that this thesis does not contain work that has been submitted, in whole or in part, for the award of any other degree.

I declare that this thesis has been produced in accordance with the University of Glasgow’s Code of Good Practice in Research.

I understand that if any concerns related to good research practice are identified during the thesis review, the examination may be postponed until the outcome of an investigation is determined.

Signature: Saja Al Ani

Date: 09/03/2026

Dedication

To my beloved father, Adil Al Ani. Your love, wisdom, and quiet sacrifices continue to guide me, even in your absence. You shaped who I am and who I strive to be. This work is dedicated to your memory, with enduring love and gratitude.

Statement of Copyright

Contents

Abstract i

Acknowledgements ii

Author's declaration iii

Dedication iv

Statement of Copyright v

Contents vi

List of Tables xi

List of Figures xiii

List of Publications xv

List of Abbreviations xvi

1. Introduction 1

1.1 Background 1

1.2 Scope and Motivation 4

1.3 Problem Statement and Objectives 7

1.3.1 Aims and Objectives 7

1.4 Research Contributions 8

1.5 Thesis Organisation 12

2. Literature Review 14

2.1 Background 15

2.1.1 Speech Sound Disorders in Childhood 15

2.1.2 Speech Sound Disorders and Cleft Lip and Palate 16

2.1.3 Current Assessment Practices 16

2.1.4 Imaging Modalities for Tongue Visualisation 17
2.1.5 Fundamentals of Ultrasound Tongue Imaging 19
2.1.6 Deep Learning Foundations for Medical Imaging 23
2.2 Deep Learning in Ultrasound Tongue Imaging 30
2.2.1 Scope and Corpus of the Review 37
2.2.2 Application Domains and Representative Studies 40
2.3 Challenges, Research Gaps and Link with Challenges 48
2.4 Summary 52
3. Establishing Baseline Models for Phonetic Classification from Raw Ultrasound Imaging . 53
3.1 Introduction 54
3.1.1 Contributions 57
3.2 Methods 58
3.2.1 Dataset 58
3.2.2 Local Binary Patterns 60
3.2.3 Models and Experimental Setup 62
3.2.4 Training Configuration 67
3.2.5 Evaluation Strategy 68
3.3 Results 69
3.3.1 Baseline Performance 69
3.3.2 Transfer Learning Models 70
3.3.3 FusionNet Performance 71
3.3.4 Contribution of Individual Input Streams to Fusion Performance 74
3.4 Discussion 76
3.5 Summary 80

4. Improving Interpretability in Ultrasound Tongue Imaging through Representation and Harmonisation 81
4.1 Introduction 82
4.1.1 Contributions 84
4.2. Method 86
4.2.1. Datasets 86
4.2.2. Image Representation Techniques 89
4.2.3. Field of View Normalisation 93
4.2.4. Model Architecture and Training Strategy 95
4.2.5 Evaluation Metrics and Explainability 98
4.3 Results 100
4.3.1 Model Performance Across Image Representations 100
4.3.2 Impact of FoV Alignment 102
4.3.3 Grad-CAM Visualisation of Model Attention 103
4.3.4 Statistical Comparison of Grad-CAM Activations 106
4.3.5 Clustering of Grad-CAM Features 108
4.4 Discussion 111
4.5 Summary 119
5. Generative Harmonisation of Ultrasound Tongue Imaging via a Two-Stage Conditional GAN 120
5.1 Introduction 121
5.1.1 Contributions 124
5.2 Method 125
5.2.1 Datasets 127
5.2.2 Data Preprocessing 127

5.2.3 Pix2Pix cGAN Architecture 128
5.2.4 Generative Image Standardisation Pipeline 131
5.2.5 Evaluation of Generated Images 133
5.3 Results 136
5.3.1 Image quality 136
5.3.2 Classification Performance Analysis 139
5.3.3 Generalisation Testing 140
5.4 Discussion 142
5.5 Summary 148
6 Data Sufficiency and Annotation Optimisation in Ultrasound-Based Speech Classification 150
6.1 Introduction 151
6.1.1 Contributions 154
6.2 Method 155
6.2.1 Proposed Framework for Cost-Efficient Sampling and Annotation 155
6.2.2 Datasets and Preprocessing 160
6.2.3 Deep learning setup 161
6.3 Results 163
6.3.1 Phase 1- Optimised Dataset Capture 163
6.3.2 Phase 2- Optimised Annotation via Active Learning 169
6.3.3 Combined Cost Analysis 171
6.3.4 Architecture Validation 175
6.4 Discussion 177
6.5 Summary 181
7. Translating the Generative Pipeline into a Deployable Ultrasound Processing System 182

7.1 Introduction 183
7.1.1 Contributions 185
7.2 System Design 186
7.2.1 Backend Implementation 187
7.2.2 Frontend Implementation 189
7.2.3 Technical Challenges and Solutions 191
7.2.4 Considerations for Future Clinical Translation 193
7.2.5 Performance Evaluation 194
7.3 Discussion 196
7.4 Summary 198
8. Conclusion and Future Work 199
8.1 Summary of Contributions 199
8.2 Limitations and Future Research Directions 202
Acknowledgement for the use of AI 205
Bibliography 206
Appendix A: UltraSuite Speaker IDs 222
Appendix B: Supplementary Training Dynamics and Error Analysis 223
Appendix C: EfficientNet-B0 Architecture Validation 228

List of Tables

Table 2. 1: Comparative Analysis of Major Medical Imaging Modalities. 19

Table 2. 2: Inclusion and Exclusion Criteria. 32

Table 2. 3: Summary of Direct SSD Detection Studies. 35

Table 2. 4: Summary of Technical Foundations Studies. 36

Table 2. 5: Datasets Used by the Included Studies. 38

Table 2. 6: Mapping from Included Studies to the Dataset. 39

Table 2. 7: Comparative Analysis of DL-Based UTI Studies for Classifying SSDs. 43

Table 2. 8: Comparative Analysis of Foundational UTI Studies. 46

Table 2. 9: Summary of Research Gaps and Corresponding Thesis Focus. 51

Table 3.1: Demographic Summary of the Selected Subset. 59

Table 3. 2: Summary of Hyperparameters for Baseline and FusionNet Models. 63

Table 3.3: Baseline Models Accuracy Performance. 70

Table 3.4: Transfer Learning Models Accuracy Performance. 71

Table 3. 5: FusionNet Model's Accuracy Performance. 73

Table 4.1: Overview of Datasets Used in this Study. 88

Table 4.2: Tongue Contour Localisation Validation 92

Table 4.3: Overview of Dataset Variations Used for FoV Experiments. 95

Table 4. 4: Summary of Hyperparameters for EfficientNet-B0 97

Table 4.5: Model Performance Across Image Representations and Datasets. 101

Table 4.6: Summary of Significant Grad-CAM Activation Differences between TD and CP±L Speech. 107

Table 5. 1: Summary of Hyperparameters for the Two-Stage Pix2Pix cGAN Pipeline..... 131

Table 5.2: Summary of Evaluation Metrics used to Assess Generated Image Quality, Classification Performance, and Generalisation. 134

Table 5.3: Image Quality Assessment of Generated Outputs Across both Pix2Pix Stages.... 138

Table 5.4: Classification Performance of the Proposed Two-Stage Generative Pipeline..... 140

Table 5.5: Stage 1 Cross-Domain Generalisation Results. 142

Table 5.6: Comparison of FoV Standardisation Approaches..... 143

Table 6. 1: Summary of Hyperparameters for Power-Curve and AL Experiments..... 161

Table 6.2: Comparison of Candidate Learning Curve Models. 164

Table 6.3: Estimated UTI Dataset Requirements for AlexNet Baseline Performance using Exponential Decay Model..... 168

Table 6.4: Summary of Performance and Cost Across Collection and Annotation Strategies. 172

Table 6.5: Cost Analysis Sensitivity to Collection: Annotation Cost Ratios..... 174

Table 6.6: Framework Performance Across Architectures and Imaging Modalities..... 175

List of Figures

Figure 2. 1: Midsagittal UTI Showing Key Anatomical Landmarks and Characteristic Artefacts. 20

Figure 2. 2: Components of a UTI Scanning Setup. 21

Figure 2. 3: Illustration of a CNN Processing an UTI. 24

Figure 2. 4: Overview of DL for UTI in SSD Assessment. 31

Figure 2. 5: PRISMA Flow Diagram of the Study Selection Process. 33

Figure 3. 1: Raw UTI Frame and its LBP Texture Representation (a) Example Mid-Sagittal UTI from the UXTD Corpus. (b) LBP Code Image Computed from the Same Frame. 61

Figure 3. 2: The Proposed FusionNet Architecture. 66

Figure 3. 3: All Models' Accuracy Performance. 72

Figure 3. 4: Confusion Matrix for Speaker-Independent on the Testing Dataset. 74

Figure 3. 5: Contribution of Individual Input Streams. 76

Figure 4. 1: Original UTI frame (left) and automatically tracked tongue surface AAA's built-in DLC module (right, red overlay), with the tongue tip on the right. 90

Figure 4. 2: ROI image (left) showing the isolated tongue region, and binary segmentation mask (right) created via thresholding, with the tongue tip on the right. 91

Figure 4. 3: Grad-CAM Heatmaps from a Model Trained on Raw UTI. 104

Figure 4. 4: Grad-CAM Heatmaps for Models Trained on ROI and Masked UTI. 105

Figure 4. 5: PCA Projection of Grad-CAM++ Activation Features Across all Speakers. 109

Figure 4. 6: Phoneme-Specific PCA projections of Grad-CAM++ Activation Features for Alveolar, Palatal, and Velar Consonants Across all Speakers. 110

Figure 5. 1: Overview of the proposed system, comprising the two-stage Pix2Pix generative preprocessing pipeline (Stage 1: FoV standardisation; Stage 2: ROI refinement) followed by the EfficientNet-B0 downstream classifier. The generative stages operate at the image level to standardise inputs; the classifier operates on the standardised outputs to perform binary SSD classification. 126

Figure 5. 2: Structure of the Proposed cGAN (Pix2Pix). 129

Figure 6. 1: Iterative Workflow for Estimating Dataset Sufficiency via Power-Curve Modelling. 156

Figure 6. 2: Workflow of Phase 2 AL for Cost-Effective Annotation. 158

Figure 6. 3: Relationship Between Training Set Size and Classification Accuracy. 163

Figure 6. 4: Residual Diagnostics for the Exponential Decay Model. 166

Figure 6. 5: Exponential Learning Curve for UTI Dataset Scaling. 167

Figure 6. 6: Comparison of AL vs Random Sampling Strategies. 169

Figure 6. 7: Cost Breakdown Across Data-Collection and Annotation Strategies. 172

Figure 6. 8: Sensitivity Analysis. 174

Figure 7. 1: Deployment architecture showing the Streamlit frontend for user interaction and the FastAPI backend hosting the FoV and ROI Pix2Pix models, alongside the operational user interface. 184

Figure 7. 2: High-level Lifecycle of the Deployed UTI Pix2Pix Prototype System. 186

Figure 7. 3: Example of the deployed Streamlit interface showing ROI refinement with input–output comparison and SSIM/PSNR metrics. 190

List of Publications

1. Al Ani, Saja, Cleland, Joanne and Zoha, Ahmed (2024) Automated Classification of Phonetic Segments in Child Speech Using Raw Ultrasound Imaging. In: 17th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOIMAGING 2024), Rome, Italy, 21-23 February 2024, pp. 326-331. ISBN 9789897586880 (doi: 10.5220/0000184700003657).

2. Al Ani, Saja, Cleland, Joanne and Zoha, Ahmed (2025) Deep learning in ultrasound tongue imaging: a systematic review toward automated detection of speech sound disorders. Frontiers in Artificial Intelligence, 8, 1631134. (doi: 10.3389/frai.2025.1631134) (PMID:41069929) (PMCID:PMC12504283).

3. Al Ani, Saja, Cleland, Joanne and Zoha, Ahmed (2025) Two-Stage GAN for Field-of-View Standardisation and Tongue Region Enhancement in Ultrasound for Cleft Palate Speech Pattern Analysis. In: 6th International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD 2025), London, UK, 19-21 Nov 2025, (Accepted for Publication).

4. Al Ani, Saja, Cleland, Joanne, and Zoha, Ahmed (2025) A Framework for Assessing and Optimising Data Sufficiency in Ultrasound Tongue Imaging. In: 13th International Conference on Bioimaging (BIOIMAGING 2026), Marbella, Spain, 02-04 Mar 2026, (Accepted for Publication).

5. Al Ani, Saja, Cleland, Joanne, and Zoha, Ahmed (2026) Explainable Deep Learning for Ultrasound Tongue Imaging with Standardised Image Representations and Field of View. Scientific Reports (under review).

List of Abbreviations

SSD	Speech Sound Disorder
SLT	Speech and Language Therapist
UTI	Ultrasound Tongue Imaging
ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
FoV	Field of View
AI	Artificial Intelligence
MRI	Magnetic Resonance Imaging
CT	Computed Tomography
XAI	Explainable Artificial Intelligence
DNN	Deep Neural Network
ROI	Region of Interest
cGAN	Conditional Generative Adversarial Network
AL	Active Learning
LBP	Local Binary Pattern
Grad- CAM	Gradient- Weighted Class Activation
SSIM	Structural Similarity Index Measure
PSNR	Peak Signal-to-Noise Ratio
CP±L	Cleft Lip and Palate
EPG	Electropalatographic
GPU	General Processing Unit
RNN	Recurrent Neural Network
GAN	Generative Adversarial Network
SSL	Self-Supervised Learning
MSD	Mean Surface Distance

LIME	Local Interpretable Model Agnostic Explanation
SHAP	SHapley Additive exPlanations
TD	Typically Developing
IoU	Intersection over Union
MSE	Mean Squared Error
UXTD	Ultrax Typically Developing Children
TaL	Tongue and Lips
UXSSD	Ultrax Speech Sound Disorders
UPX	UltraPhonix
PCA	Principal Component Analysis
DCT	Discrete Cosine Transform
ViT	Vision Transformer
VGG	Visual Geometry Group
ConvLSTM	Convolutional Long Short-Term Memory
AAA	Articulate Assistant Advanced
ReLU	Rectified Linear Unit
MLP	Multi-Layer Perceptron
SGD	Stochastic Gradient Descent
HPC	High Performance Computing
FPS	Frame Per Second
DLC	DeepLabCut
QC	Quality Control
BGR	Blue, Green, Red
HSV	Hue, Saturation, Value
BCE	Binary Cross Entropy
SD	Standard Deviation
SCAN	Structure Correcting Adversarial Network
PET	Positron Emission Tomography

CLAHE	Contrast-Limited Adaptive Histogram Equalization
CI	Confidence Interval
AIC	Akaike's Information Criterion
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
GPU	Graphical Processing Unit
CPU	Central Processing Unit
API	Application Programming Interface
HTTP	Hypertext Transfer Protocol
UK	United Kingdom
MHRA	Medicines and Healthcare products Regulatory Agency
FDA	Food and Drug Administration

Chapter 1

1. Introduction

1.1 Background

SSDs represent one of the most common developmental communication difficulties in childhood, affecting approximately 3.8% to 6.4% of 5 to 8-year-old children [1]. These disorders manifest as consistent or inconsistent misarticulations of speech sounds, often leading to reduced intelligibility and long-term consequences for literacy, academic attainment, and psychosocial well-being [2], [3]. Early and accurate identification is therefore essential, allowing targeted intervention by speech and language therapists (SLTs) to prevent persistent difficulties. Conventional diagnostic practice relies primarily on auditory–perceptual assessment, in which SLTs judge a child's speech accuracy through listening and transcription. Although this remains the clinical gold standard, it is inherently subjective: inter-rater agreement can vary substantially, and subtle articulatory errors may escape detection [4]. As a result, there is a growing demand for objective, instrument-based methods that provide reproducible evidence of articulatory function.

UTI has emerged as a powerful technique for visualising tongue shape and motion during speech. Using a small transducer placed under the chin, UTI captures high-frequency sound reflections from the tongue surface to generate a dynamic, mid-sagittal view of articulatory movement [5]. It is safe, non-invasive, and radiation-free, making it well-suited to paediatric and clinical populations [6]. In speech therapy, UTI enables clinicians to show children how their tongue moves in real time, providing a visual feedback channel that supports articulation correction [7]. Research applications have likewise expanded, with UTI being used to study coarticulation and articulatory variability across languages [8], [9].

Despite its value, however, the interpretation of UTI remains challenging. UTIs are inherently noisy due to acoustic speckle, exhibit low contrast, and vary with factors such as probe angle, pressure, and coupling medium. Manual analysis, such as tracing tongue contours or visually judging shape differences, is time-consuming, operator-dependent, and unsuitable for large datasets [10]. A comprehensive discussion of UTI principles, image characteristics, and clinical applications is provided in Chapter 2, Section 2.1.5. These limitations have motivated the development of computational methods to automate the interpretation of UTIs.

Parallel progress in machine learning (ML) and DL has transformed medical imaging across many domains. Convolutional neural networks (CNNs) and related architectures have achieved human-level performance in radiology and pathology tasks, including tumour classification, organ segmentation, and image reconstruction [11], [12]. DL models excel at learning hierarchical feature representations directly from raw image data, eliminating the need for hand-crafted features and enabling robust generalisation when sufficient data are available. Within clinical linguistics, ML has been applied to acoustic analysis, articulatory modelling, and automatic phoneme recognition [13]. Extending these advances to UTI offers the prospect of automated detection of misarticulations or atypical tongue gestures, providing clinicians with objective indicators of speech production accuracy.

Nevertheless, the application of DL to UTI introduces unique challenges. Ultrasound differs markedly from photographic or radiological imagery in its physics and appearance: speckle noise [14], variable FoV [15], and ambiguous tissue boundaries complicate feature learning. Moreover, datasets are typically small and heterogeneous, collected under differing experimental setups and probe calibrations. These factors limit model generalisability, with systems often performing well on individual speakers but failing on unseen subjects [16]. In addition, model interpretability is a critical issue. Clinical adoption of artificial intelligence (AI) requires transparency; SLTs need to understand why a system predicts a particular articulatory category or diagnosis [17]. Without explainable mechanisms, automated decisions risk being viewed as unreliable or untrustworthy in clinical contexts. These challenges form the foundation of the methodological framework developed across Chapters 3-7 of this thesis.

Despite these obstacles, recent studies demonstrate the feasibility of DL for UTI. Ribeiro et al. [16] showed that CNNs can classify phonetic segments directly from raw ultrasound frames, while Fabre et al. [18] and Xu et al. [19] reported success in tongue-contour extraction and self-supervised feature learning, respectively. These efforts highlight both the promise and the persistent gaps in current approaches: accuracy remains limited by small datasets, and cross-speaker generalisation is poor. Furthermore, most research has focused on algorithmic performance rather than clinical translation, leaving issues such as data standardisation, interpretability, and deployment unresolved.

Consequently, the field now faces a dual imperative. From a clinical perspective, there is a growing need for objective visual tools that complement auditory–perceptual assessment, particularly when clinicians need to understand how a child is producing a sound rather than simply whether the sound is correct. Auditory methods remain highly effective for the diagnosis of SSDs, but they cannot reveal the hidden articulatory patterns, such as covert contrasts or compensatory gestures, that often guide therapy planning and progress monitoring. From a technological perspective, there is a need for DL frameworks that can learn from limited, heterogeneous ultrasound data while remaining explainable and practical for real-world use. The convergence of these aims defines the motivation of this thesis: to explore how DL can be systematically adapted to UTI to produce models that are accurate, interpretable, and clinically deployable as supportive tools for assessment and therapy, rather than replacements for auditory diagnosis.

In summary, while UTI provides clinicians with a unique window into the articulatory processes underlying speech, its interpretation remains constrained by the same factors that limit many imaging-based diagnostics: variability, noise, and reliance on expert judgement. As DL continues to transform medical imaging through automation and pattern discovery, it offers a compelling opportunity to translate these visual observations into quantifiable, objective measures of speech production. Bridging this clinical–technological divide forms the foundation of this thesis, exploring how DL can be adapted to the challenges of ultrasound data to support accurate, reliable differential diagnosis of SSDs.

1.2 Scope and Motivation

The integration of UTI with ML holds great promise for the objective differential diagnosis of childhood SSDs. Yet, translating this promise into a clinically viable system requires confronting several interrelated challenges at the intersection of data acquisition, model design, and practical deployment. These challenges define the scope of this research and motivate the methodological choices made throughout the thesis.

The central motivation is twofold. Clinically, there is an urgent need for diagnostic tools that provide objective, reproducible evidence of articulatory behaviour, complementing traditional perceptual assessments. Technologically, there is a need for computational frameworks that can handle the intrinsic variability and data limitations of UTI while remaining interpretable to non-technical users. The work presented in this thesis is therefore motivated by the goal of bridging the gap between speech science and AI: to develop DL methods that are robust to heterogeneous ultrasound data, efficient in their data requirements, and transparent in their decision-making.

Despite steady progress in both ultrasound-based speech analysis and ML, the translation of these advances into robust, clinically usable systems has remained limited. Many proposed methods demonstrate promising performance under controlled conditions yet fail to generalise across speakers, recording sessions, or acquisition setups. These limitations arise not from a single deficiency but from three interrelated constraints that collectively define the scope of this thesis:

> **C1 – Data Variability and Generalisability Limitations:** A primary obstacle to reliable automated analysis is the variability inherent in UTI data. Each ultrasound session depends on the equipment used, the probe's angle and pressure, and the individual anatomy of the speaker. These parameters produce significant differences in image geometry, intensity distribution, and visible anatomical coverage. Even when speakers produce the same phoneme, the corresponding ultrasound frames may vary significantly in spatial scale or contrast. This heterogeneity leads DL models to overfit to dataset-specific characteristics rather than to the underlying articulatory structure, causing models trained on one dataset to perform poorly on new speakers or acquisition conditions [20].

This challenge mirrors the well-documented phenomenon of domain shift in medical imaging, where differences in scanners, acquisition protocols, and patient populations degrade generalisability when algorithms are deployed in new settings [21]. Previous UTI research has attempted to mitigate variability through manual preprocessing, such as contour tracing, fixed cropping, or headset stabilisation, but these approaches are labour-intensive and fail to capture the full articulatory context [22]. Hence, the first challenge (**C1**) addressed in this thesis is the development of strategies that enhance model robustness to acquisition variability and improve generalisability across speakers, sessions, and recording conditions.

C2 – Data Scarcity and Annotation Efficiency: A second, equally critical challenge concerns the availability of annotated data. Unlike large public repositories for magnetic resonance imaging (MRI) or computed tomography (CT), UTI datasets remain small and fragmented. Collecting such data requires the cooperation of clinical participants, ethical approval, and the use of specialist equipment. Annotation adds a further bottleneck: each frame must be examined by an SLT or trained researcher to label the phonetic category or articulatory event, a process that is both time-consuming and costly [5]. This limitation is particularly pronounced in paediatric research, where recording time is constrained by the child's attention span and ethical considerations. From an ML perspective, the lack of labelled data restricts model complexity and undermines statistical confidence. Models trained on limited examples risk overfitting, learning idiosyncratic features that do not generalise [23]. Attempts to compensate through data augmentation [24], or transfer learning from natural images has yielded mixed results because the statistical properties of ultrasound differ markedly from photographic imagery [25]. Consequently, the second challenge (**C2**) targeted in this research is the efficient use of limited data, achieved through approaches that minimise the quantity of labelled samples required while preserving diagnostic accuracy.

C3 – Lack of Interpretability and Clinical Usability: A third limitation concerns the interpretability of DL models and their practical deployment. Neural networks can achieve high accuracy but often operate as opaque systems, providing little insight into how decisions are made.

In a clinical setting, this opacity undermines trust: clinicians must be able to verify that a system bases its classification on anatomically meaningful evidence rather than on spurious correlations or background artefacts [26]. XAI techniques [27], such as gradient-based visualisation, have been proposed to expose model attention; however, their application to UTI remains limited. Beyond interpretability, usability itself poses a challenge. Most studies stop at experimental proof-of-concepts, leaving clinicians without accessible tools for real-time analysis or feedback. Addressing this third challenge (**C3**), therefore, requires two complementary advances: first, incorporating XAI to make model reasoning transparent and clinically interpretable; and second, translating research models into deployable systems with intuitive interfaces that integrate seamlessly into therapy and research environments.

The scope of this thesis extends across all three challenges. To address **C1**, the thesis explores discriminative and generative approaches to enhance generalisability, from baseline CNN/DNN architectures and multi-input fusion models to a generative adversarial framework for FoV and ROI standardisation. To mitigate **C2**, a cost-aware methodology combining statistical power analysis with active learning is introduced to determine minimum dataset requirements and reduce dependency on exhaustive manual labelling. To overcome **C3**, XAI tools are integrated for visual interpretation of model behaviour, and a prototype system is deployed within a web-based environment using FastAPI and Streamlit.

The overarching motivation, therefore, is to enable accurate, interpretable, and deployable AI for ultrasound-based speech assessment. By systematically targeting **C1**–**C3**, the thesis moves beyond algorithmic innovation toward clinically meaningful translation, demonstrating how DL can improve diagnostic performance while enhancing understanding, efficiency, and accessibility in speech pathology.

1.3 Problem Statement and Objectives

Despite major advances in ML and DL, the automated analysis of UTI for SSD diagnosis remains limited by three persistent barriers: data variability and generalisability issues, data scarcity and annotation efficiency, and lack of interpretability and clinical usability. Variability in the FoV, probe alignment, and participant anatomy introduces large inconsistencies across datasets, reducing the generalisability of trained models and undermining reproducibility across speakers and recording sessions. At the same time, the scarcity of annotated UTI data, driven by the time-intensive nature of expert labelling, restricts the ability to train data-hungry DL models. Finally, most DL models used in medical imaging, including UTI, function as black boxes, providing little insight into their decision mechanisms. This lack of interpretability hinders clinical trust and prevents meaningful adoption in therapeutic contexts. Therefore, the central problem addressed in this thesis is how to design DL frameworks that are robust to imaging variability, efficient under limited annotation, and transparent in their reasoning, thereby enabling accurate and clinically meaningful analysis of tongue movement in children's speech.

1.3.1 Aims and Objectives

To develop and evaluate a comprehensive DL framework that enhances the accuracy, interpretability, and scalability of ultrasound-based systems for diagnosing and monitoring SSDs in children. To achieve this aim, the thesis pursues five interconnected objectives, each corresponding to a major phase of the research:

1. To establish reproducible DL baselines for phonetic classification of child speech from raw UTI, investigating model performance under several speakers' scenarios and quantifying generalisation limitations.
2. To investigate how image representation strategies (raw frames, ROI, and segmentation) and FoV variability influence classification performance and anatomical interpretability in UTI-based models.
3. To develop and evaluate a generative framework for FoV standardisation and ROI enhancement, investigating how two-stage conditional generative adversarial network

(cGAN) architectures can reduce inter-speaker variability and improve cross-session generalisation in UTI-based classification.

4. To determine the minimal data requirements for clinically meaningful performance, investigate an AL strategy that optimises annotation efficiency while maintaining diagnostic accuracy.
5. To evaluate the technical feasibility and computational requirements for deploying DL models for UTI processing, examining the trade-offs between model complexity, inference speed, and image-quality performance, and assessing system behaviour under realistic hardware and operational constraints.

1.4 Research Contributions

The research presented in this thesis directly addresses the three key challenges identified in Section 1.2: data variability and generalisability limitations (**C1**), data scarcity and annotation efficiency (**C2**), and lack of interpretability and clinical usability (**C3**). Each contribution addresses these challenges through methodological innovation, empirical evaluation, and applied system design, forming an end-to-end framework for automated UTI analysis.

The following contributions are organised to mirror the three central challenges outlined earlier. The sequence from R1 to R7 reflects both methodological progression and thematic integration, from establishing reproducible baselines, through representation analysis, generative harmonisation, and cost-efficient learning, to explainable and deployable systems. Together, these contributions operate as a cumulative framework rather than discrete experiments, with advances in one domain, for example, generative harmonisation, directly reinforcing others, such as interpretability and data efficiency.

Addressing C1 – Data Variability and Generalisability Limitations

Heterogeneity in ultrasound acquisition settings, differences in FoV, probe alignment and anatomical coverage, remains one of the major obstacles to reliable UTI-based classification. This thesis contributes a sequence of methodological advances to mitigate these effects and improve model robustness across speakers and sessions.

- **R1. Establishment of reproducible DL baselines:** A CNN and DNN architectures were implemented for phonetic classification using raw UTI, providing reproducible benchmarks for evaluating DL performance. These baselines quantify how model accuracy degrades under speaker-independent evaluation and form a reference for all subsequent enhancements.
- **R2. Development of FusionNet for multi-input learning:** A novel multi-input architecture, FusionNet, was designed to fuse complementary information from raw UTI frames and texture-based Local Binary Pattern (LBP) features to improve generalisability issues. LBP encodes the local texture pattern of an image by comparing the intensity of each pixel with its neighbours, capturing fine-grained micro-structures such as speckle texture and tissue edges that are characteristic of ultrasound. By integrating these local cues with the global tongue shape contained in raw frames, FusionNet achieved superior generalisability across speakers and recording conditions, demonstrating that multimodal fusion can alleviate dataset-specific bias.
- **R3.** Representation analysis and interpretability evaluation through XAI: To investigate how image representation and FoV variability jointly influence model behaviour, a systematic evaluation of raw, ROI, and binary mask inputs was conducted using EfficientNet-B0. Grad-CAM++ visualisation was applied to assess whether models attend to anatomically plausible regions under each representation condition. This analysis demonstrated that harmonised FoV inputs and anatomically focused representations improve both classification robustness and the clinical plausibility of model attention, providing empirical evidence that variability cannot be overcome through architectural innovation alone.

- **R4.** FoV standardisation via a conditional GAN: To resolve geometric inconsistencies arising from differences in probe placement and imaging angle across speakers and sessions, a conditional Pix2Pix GAN was trained to transform wide-angle UTI frames into a standardised 97° field of view. Stage 1 of the generative pipeline achieves near-lossless structural preservation, as measured by SSIM and PSNR, and demonstrably improves downstream classification accuracy by reducing acquisition-level domain shift.
- **R5.** ROI refinement via a conditional GAN: Building on the FoV-standardised outputs of R4, a second Pix2Pix model was trained to refine the region of interest by suppressing background acoustic noise and irrelevant tissue, enhancing the visibility of the tongue surface for downstream analysis. This two-stage generative pipeline together represents a principled solution to inter-speaker imaging variability, delivering the most substantial improvements in cross-speaker classification stability observed across the thesis.

Together, contributions R1–R5 provide a cumulative pathway from conventional baseline modelling through representation-level analysis to generative domain standardisation, directly addressing the limitations of variability and reproducibility in UTI datasets.

Addressing C2 – Data Scarcity and Annotation Efficiency

Obtaining large, well-annotated ultrasound corpora is expensive and time-intensive, as annotation requires expert review of thousands of frames. This thesis introduces a cost-aware methodology that quantifies dataset sufficiency and minimises annotation effort without compromising accuracy.

- **R6.** Cost-aware data efficiency framework combining statistical power-curve modelling and active learning: Adapting methodologies from clinical trial design, power-curve analysis was applied to model the relationship between dataset size and classification accuracy, identifying the point of diminishing returns beyond which further data collection yields marginal performance gains. Building on this, an AL strategy using uncertainty sampling was implemented to prioritise the most informative samples for expert annotation, reducing the labelled dataset requirement by over 50% while maintaining 90% accuracy.

In addition, synthetic data generated by the GAN pipeline was shown to augment training effectively, further mitigating data scarcity and improving model robustness. Together, these methods provide a quantitative framework for balancing annotation effort against performance, offering a scalable and economically viable approach to dataset development in speech ultrasound research.

Addressing C3 – Lack of Interpretability and Clinical Usability

For DL models to be adopted in clinical practice, they must be both interpretable and accessible. This thesis advances interpretability through XAI techniques (R3) and enhances usability through software deployment.

- **R7. Technical prototype deployment of a UTI processing system:** A complete technical prototype deployment pipeline was implemented using FastAPI (for backend inference) and Streamlit (for frontend interaction).
  This operational prototype performs FoV standardisation and ROI refinement, enables rapid processing through a web interface, and provides a foundation for future clinical-facing tools. While not clinically validated, the system demonstrates how the generative models developed in this thesis can be operationalised and used to support research workflows in UTI-based speech analysis.

Together, contributions R6 and R7 bridge the gap between algorithmic development and clinical application, ensuring that the proposed methods are transparent, explainable, and usable beyond the laboratory setting.

Summary of Contributions

By systematically addressing **C1**–**C3**, this thesis advances the field of ultrasound-based speech analysis in several keyways:

1. Establishes a robust empirical baseline and proposes a multi-input fusion model that improves generalisation across speakers.
2. Analyses the effects of image representation and FoV variability on model performance and interpretability using XAI.

3. Introduces a novel two-stage generative framework for FoV standardisation and ROI refinement, producing anatomically consistent and diagnostically relevant data.
4. Provides a quantitative, cost-focused methodology for determining dataset sufficiency and optimising annotation efficiency.
5. Provides a prototype of a deployable real-time processing system for clinical use.

Collectively, these contributions form an integrated pipeline, from data collection and standardisation to classification, cost optimisation, and deployment, laying the groundwork for reliable, interpretable, and scalable AI in speech disorder diagnosis

1.5 Thesis Organisation

This thesis is organised into seven main chapters, each addressing one or more of the research challenges identified in Section 1.2 and contributing to the overarching aim of developing accurate, interpretable, and deployable DL frameworks for UTI in SSD diagnosis. The structure of the work reflects a logical progression, from establishing a contextual understanding and baseline to methodological innovation, evaluation, and practical deployment. The detailed organisation is as follows:

Chapter 1 introduces the clinical and technical background of SSDs and UTI, defines the scope and motivation of the study, articulates the central research problem, and outlines the aims, objectives, and key contributions of the thesis. It establishes the conceptual framework for addressing the three core challenges: Data variability and generalisability limitations (**C1**), data scarcity and annotation efficiency (**C2**), and lack of interpretability and clinical usability (**C3**).

Chapter 2 provides a comprehensive synthesis of prior research relevant to this work. It begins with an overview of medical imaging in clinical diagnostics and narrows the focus to UTI as a tool for visualising articulatory movement. The chapter critically reviews DL applications in medical and speech imaging, identifying the current limitations, data heterogeneity, annotation constraints, and lack of explainability that motivate the present research.

Chapter 3 establishes baseline models for phonetic classification of child speech using raw ultrasound imaging. Addresses **C1** by developing reproducible baseline architectures for phonetic classification from UTI. Custom CNN and DNN models are implemented and

benchmarked against transfer-learning architectures (ResNet-50, Inception-V3). This chapter quantifies model performance under speaker-dependent and speaker-independent scenarios and introduces FusionNet, a multi-input architecture combining raw images and texture-based features to enhance robustness.

Chapter 4 expands upon **C1 and C3** by analysing how image representations and FoV variability affect both model accuracy and interpretability. Grad-CAM++ visualisation is used to reveal model attention patterns, demonstrating that harmonised and ROI-focused representations improve both diagnostic precision and transparency.

Chapter 5 directly tackles **C1** through a generative approach. A two-stage Pix2Pix cGAN pipeline is introduced: Stage 1 performs FoV harmonisation to a standardised 97° view, and Stage 2 refines the region of interest to emphasise the tongue and suppress background noise. Quantitative metrics (SSIM, PSNR) and downstream classification demonstrate that this harmonisation substantially improves model fidelity and generalisability.

Chapter 6 introduces a cost-focused framework for optimising annotation in UTI. Addresses **C2** by presenting a cost-aware methodology that quantifies the trade-off between dataset size, annotation effort, and predictive performance. Statistical power-curve modelling estimates data sufficiency, while AL selectively targets informative samples for labelling, achieving 50 % reduction in annotation requirements without loss of diagnostic accuracy.

Chapter 7 introduces a DL prototype system for UTI processing that responds to **C3** by translating the developed frameworks into a practical, clinician-oriented application. A FastAPI backend and Streamlit frontend are implemented to provide real-time processing for FoV harmonisation and ROI refinement. This chapter details the prototype system architecture, performance evaluation using the SSIM and PSNR, and the challenges encountered during deployment on CPU and GPU environments.

Chapter 8 summarises the key findings across all studies, synthesising how each objective contributes to overcoming the three major challenges. The chapter discusses the clinical and methodological implications of the research, outlines its limitations, and proposes future directions, including larger multi-centre validation, integration of temporal information for dynamic speech analysis, and further enhancements to interpretability.

Chapter 2

2. Literature Review

Medical imaging is an essential component of modern healthcare that supports diagnosis, treatment planning, and disease monitoring. Advances in CT, MRI, and ultrasound have greatly expanded both the quantity and complexity of visual data available to clinicians. Within speech and language therapy, this capability is particularly transformative. Many articulatory movements crucial for speech production, especially those of the tongue, occur within the oral cavity and cannot be observed directly. Visualising these gestures provides clinicians with objective evidence that clarifies the source of a speech disorder, informs therapy, and supports progress monitoring [7], [28]. UTI has emerged as a key technique in speech science and clinical phonetics. It offers real-time, non-invasive observation of tongue motion using safe acoustic energy rather than radiation [29]. This makes it suitable for paediatric use and repeated assessment [30]. Advances in computer vision and DL have revolutionised how ultrasound data can be analysed [31]. DL-based UTI analysis now supports automatic classification, segmentation, and synthesis of tongue patterns, enabling the modelling of both typical and atypical articulatory behaviours [20], [28], [32]. This chapter critically examines the intersection of DL and UTI in the diagnosis of SSDs. It expands upon and contextualises the author's published systematic review [33]. The review's scope and findings are integrated here to provide a comprehensive, thesis-wide perspective that connects current evidence to the methodological developments and experiments presented in later chapters. The discussion begins by situating SSDs within their broader clinical and developmental context, then outlines the principles of UTI and its imaging constraints. Subsequent sections examine DL methods relevant to medical imaging, before synthesising the systematic review's findings on DL in UTI.

The chapter concludes by identifying key research gaps that motivate the research contributions presented in Chapters 3 to 7. These three challenges: data variability and generalisability limitations (**C1**), data scarcity and annotation efficiency (**C2**), and model interpretability and clinical usability (**C3**), were introduced in Chapter 1, Section 1.2, and provide the organising framework for both this literature review and the subsequent experimental work

2.1 Background

2.1.1 Speech Sound Disorders in Childhood

SSDs encompass a range of difficulties in producing speech sounds that limit a child's intelligibility to others [34]. These range from mild articulation errors to severe reductions in comprehensibility that can affect literacy, academic attainment, and social participation [35], [36] ,[37], [38]. In the United Kingdom, SSDs are highly prevalent, with estimates suggesting that nearly a quarter of children experience speech-related difficulties, and approximately 3–4% present with severe and persistent impairments that may extend into adulthood [39]. SSDs may arise from unknown causes or from identifiable structural or functional factors and are frequently accompanied by challenges in related domains such as speech perception and motor control. Clinically, SSDs are commonly classified into articulation disorders, phonological disorders[40], and motor speech disorders, such as childhood apraxia of speech or dysarthria [41]. Because many articulatory movements occur within the oral cavity and are invisible to the naked eye, visualising tongue movement can provide essential diagnostic information that complements auditory–perceptual assessment. This is particularly relevant for identifying atypical lingual gestures, or misarticulated tongue placements that are difficult to infer from the acoustic signal alone.

Children with structural anomalies, including cleft lip and palate (CP±L), frequently present with persistent speech errors, including compensatory articulations, and therefore represent a subgroup in which articulatory imaging is especially informative [42]. These clinical complexities motivate the growing interest in tools that can visualise tongue movement to complement auditory–perceptual assessment.

2.1.2 Speech Sound Disorders and Cleft Lip and Palate

CP±L is the most common congenital craniofacial anomaly, affecting approximately 1 in every 700 live births worldwide [43]. It may present as an isolated cleft lip, a cleft palate, or a combination of both [44]. Children with CP±L often experience a range of functional difficulties, including feeding challenges, recurrent otitis media, velopharyngeal insufficiency, and associated speech disorders [45], [46]. These challenges can have broader psychosocial consequences, affecting both children and their families [47], [48],[49], [50].

Speech production in children with CP±L is characterised not only by passive symptoms associated with structural constraints (e.g., hypernasality, nasal emission) but also by active compensatory articulations, including learned speech behaviours, such as glottal stops, pharyngeal fricatives, or backing tongue placements [51]. These compensations arise as children adapt their articulatory strategy to achieve perceptually acceptable outputs despite anatomical limitations. Although ultrasound cannot visualise the velopharyngeal mechanism directly, it can reveal atypical tongue-based compensations [30]. Such insight is clinically essential for determining whether therapy should focus on articulatory placement, velopharyngeal function, or both. These clinical complexities highlight why articulatory imaging, and UTI in particular, is a powerful complement to traditional perceptual assessment. For children with CP±L, UTI provides objective, real-time visualisation of tongue posture and movement, enabling clinicians to detect otherwise hidden articulatory patterns and to plan targeted, evidence-based therapy.

2.1.3 Current Assessment Practices

Assessment of SSDs relies primarily on auditory–perceptual judgements made by SLTs. While perceptual analysis remains the clinical gold standard, it is subjective and limited by the invisibility of lingual motion. SLTs may disagree on the nature or source of an error, particularly when compensatory strategies or subtle misarticulations are involved [52]. UTI helps overcome this limitation by offering detailed visual information about articulatory movements. This reduces dependence on perceptual judgements, including phonetic transcription, which can be prone to inconsistency and reduced reliability [53].

UTI is portable, affordable, and well-tolerated by children, making it suitable for both research and clinical contexts [7]. It is increasingly used for studying speech production and for visual biofeedback during therapy [52]. Despite these advantages, challenges persist in the analysis of UTI interpretation, including inconsistency in data acquisition, variability in speakers' anatomy, differences in probe placement, and variation in the FoV. These limitations motivate the use of DL to automate interpretation, harmonise variability, and generate clinically meaningful analyses. Section 2.2 systematically reviews how DL has been applied to address these challenges.

2.1.4 Imaging Modalities for Tongue Visualisation

Understanding tongue movement is essential for identifying and treating SSDs; however, visualising the tongue during speech is challenging due to its rapid motion and location within the oral cavity. Imaging techniques address this difficulty by indirectly assessing interior movement without direct contact with the structures. Several imaging modalities have been developed for articulatory research, each offering distinct advantages and limitations:

- X-ray imaging is one of the oldest and most widely known articulatory imaging methods. It varies from other modalities in several important ways. An X-ray projects a beam through the entire head, meaning the resulting image contains overlapping structures from both sides of the vocal tract. As a result, bony structures such as the mandible and teeth tend to obscure the much fainter soft tissues, including the tongue. Moreover, concerns about radiation exposure limit the feasibility of collecting large amounts of speech data using X-ray methods [6].
- MRI uses strong magnetic fields and radiofrequency pulses to visualise internal tissues. It offers excellent soft-tissue contrast and enables full-vocal-tract imaging. Real-time MRI can capture tongue movement at moderate frame rates [54]. However, MRI remains expensive, requires participants to lie supine, and offers limited temporal resolution, making it impractical for routine speech assessment or therapy. The restricted availability further limits widespread clinical use [55], [56].

- Electropalatography (EPG) provides real-time visualisation of tongue–palate contact patterns and has proven useful in addressing articulation problems that have not responded to traditional therapeutic methods [57], [58],[59]. However, it requires a custom palate and records only contact patterns, not tongue shape, restricting its general applicability [60].
- UTI offers a favourable balance of safety, portability, and real-time visual feedback. UTI can be used to image the tongue in either a mid-sagittal or coronal orientation, with the mid-sagittal view being most commonly used for research and clinical assessment [7] and biofeedback [9]. Children tolerate UTI, and it is feasible in both clinical and research environments. Although ultrasound images are affected by acoustic shadowing and speckle noise, UTI remains the most practical and accessible modality for visualising lingual gestures during connected speech.

Table 2.1 summarises the primary differences among these modalities. X-rays are now largely obsolete due to radiation concerns. MRIs provide detailed articulatory information but are invasive, costly, or impractical for routine clinical use. EPG requires a custom palate and records only contact patterns, not tongue shape. By contrast, UTI combines accessibility with sufficient spatial and temporal resolution for functional speech assessment, underpinning all experimental work in this thesis.

Table 2. 1: Comparative Analysis of Major Medical Imaging Modalities.

Modality	Principle	Strengths	Limitations	Clinical Relevance
X-ray	Ionising radiation	Dynamic vocal tract imaging	Radiation risk; obscured tongue; unsuitable for children	Now outdated for speech therapy
MRI	Magnetic fields	Excellent soft-tissue contrast; high spatial resolution	Expensive; supine position; low temporal resolution	Research only
EPG	Custom palate electrodes detect tongue contact	Real-time feedback	Invasive; custom moulds; contact only (no shape)	Therapy use
Ultrasound	Sound waves; real-time surface imaging	Safe; portable; non-invasive; child-friendly; real-time	Artefacts; probe variability; requires training	Most practical for therapy

However, interpreting UTI is not trivial: variability in probe placement, FoV, and anatomical differences introduces inconsistency across recordings, and manual analysis is time-consuming and operator-dependent. These limitations highlight the need for automated, robust, and scalable computational methods, motivating the DL approaches discussed in Section 2.2.

2.1.5 Fundamentals of Ultrasound Tongue Imaging

UTI produces images by exploiting the reflective properties of high-frequency sound waves between 2 and 5 MHz [7]. A piezoelectric crystal within the transducer emits an ultrasonic pulse when stimulated by an electric current. As this pulse travels through the soft tissues beneath the chin, a portion of the acoustic energy is reflected whenever it encounters an interface between materials of different density, for example, between muscle and bone, or between tissue and air [6]. This produces the bright, curved echo corresponding to the tongue dorsum.

Dense structures such as the mandible and hyoid absorb or block the beam, creating characteristic dark regions known as acoustic shadows [7], [61]. An example of a midsagittal UTI frame is shown in Figure 2.1, illustrating typical anatomical landmarks and artefacts.

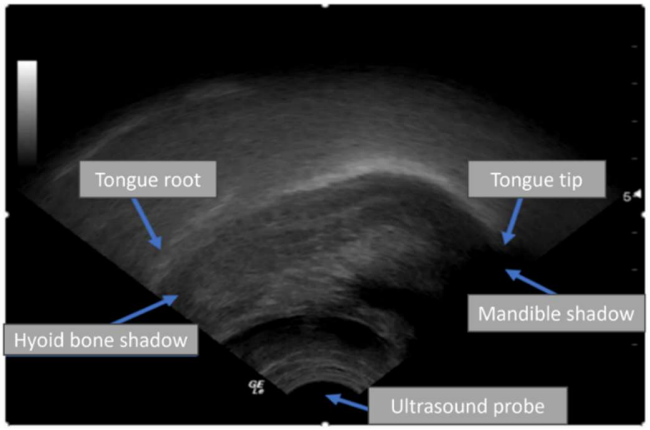

Figure 2. 1: Midsagittal UTI Showing Key Anatomical Landmarks and Characteristic Artefacts.

Image Acquisition and Probe Configuration

In speech research and clinical practice, the ultrasound probe is positioned beneath the chin to capture real-time images of the tongue, generally at 60–100 frames per second [62]. This frame rate is sufficiently rapid to capture articulatory movements, making UTI suitable for visual feedback, therapy, and phonetic analysis [30]. Various probes provide different imaging characteristics. Microconvex probes, because of their compact design, are optimal for imaging the tongue tip in paediatric patients. In contrast, convex probes offer better depth resolution for older children and adults. Regardless of the probe type, stabilisation is crucial to reduce probe movement; however, it can be uncomfortable. In contrast, handheld probes are easier to use but introduce greater variability in probe angle and pressure [16]. These factors contribute to inconsistencies in image appearance across sessions and speakers, one of the major sources of variability addressed later in this thesis.

A typical setup used in phonetics and clinical research is illustrated in Figure 2.2, showing the stabilisation headset, ultrasound probe, and a laptop-based acquisition system commonly used to record and display tongue movement in real time.

Figure 2. 2: Components of a UTI Scanning Setup.

Image Characteristics and Artefacts

UTIs are shaped by both the underlying anatomy and the physical behaviour of sound waves as they propagate through soft tissue. Consequently, UTI frames exhibit several characteristic artefacts that influence clinical interpretation and pose challenges for computational analysis [16]. Speckle noise is the most prominent feature of ultrasound imaging. It arises from the interference of backscattered echoes within the tissue, producing the familiar grainy, mottled texture [63], [64]. Although speckle reflects meaningful acoustic interactions, it can obscure fine anatomical detail and make it difficult to identify the tongue surface with precision. For automated systems, speckle introduces high-frequency variation that may distract learning algorithms from the underlying articulatory structure. A further challenge is acoustic shadowing, which arises when the ultrasound beam encounters a boundary with a large impedance mismatch [65], most notably at the mandible and hyoid bone. At these interfaces, most of the acoustic energy is reflected, producing a bright echo at the bone surface and a dark shadow beneath it where no signal is received.

These shadowed regions obscure portions of the tongue surface and can vary markedly across speakers, depending on jaw size, head posture, probe placement, and probe pressure. This inter-speaker variability is a major source of inconsistency in UTI datasets and contributes to the reduced generalisability of DL models. Taken together, these artefacts create a complex and highly variable visual environment, which increases the difficulty of reliably interpreting tongue posture and movement, reinforcing the need for robust preprocessing, harmonisation of imaging conditions, and carefully chosen feature representations, issues that are addressed in subsequent chapters of this thesis.

Image Interpretation and Clinical Use

UTI analysis is a crucial step in interpreting the features contained within these images. Its purpose is to extract relevant characteristics and patterns that support downstream tasks such as articulatory research and clinical linguistic assessment. For example, alveolar sounds such as /t/ and /d/ typically involve a forward, raised tongue tip, while velar sounds such as /k/ and /g/ involve a high, posterior tongue gesture. Deviations from these typical configurations can signal misarticulation or compensatory strategies [51]. Previous attempts at UTI interpretation remain labour-intensive and can be inconsistent across clinicians. Early computational tools, such as EdgeTrak [66], assisted in tracing the tongue surface; however, it required frequent correction, particularly in noisy or shadowed regions. Since the revolution of DL, CNN-based supervised learning has been successfully applied in UTI processing, achieving near-human levels of accuracy [67], [69], [70], [71]. These methods reduce the burden of manual feature extraction and provide more consistent outputs, supporting both research and therapeutic applications. Automated UTI analysis tools in clinical settings can identify articulatory differences that are often missed by auditory assessment. This detailed technical background, expanding on the brief introduction provided in Chapter 1, Section 1.1, establishes the imaging constraints and artefact challenges that subsequent chapters must address through algorithmic innovation.

2.1.6 Deep Learning Foundations for Medical Imaging

DNNs are computational models composed of multiple layers of artificial neurons that transform input data into increasingly abstract representations [72]. Despite the wide variety of architectures, all DNNs share core components, weights, biases, and nonlinear activation functions, which together determine how information propagates through the network. Training involves iteratively adjusting these parameters so that the model's predictions align with labelled examples. This optimisation relies on a loss function, commonly cross-entropy for classification, and backpropagation, which updates parameters by propagating the gradient of the error through the network. Over successive epochs, the model refines its internal representations and improves in accuracy. The success of this process depends strongly on factors such as architectural design, dataset size and quality, and the degree of variability present in the data [73].

In medical imaging, DNNs have demonstrated exceptional ability to learn spatial structure, identify subtle textural patterns, and generalise across heterogeneous datasets [74]. This progress has been driven primarily by CNNs, which incorporate spatially local filters that enable the hierarchical extraction of edges, contours, and increasingly complex anatomical features [75]. CNNs have played a central role in medical imaging applications such as tumour detection, organ segmentation, and image registration [76], [77].

Several developments have enabled DL to flourish in medical imaging: the availability of large annotated datasets, advances in high-performance general processing unit (GPU) computing, and the emergence of accessible open-source frameworks such as PyTorch [78] and TensorFlow [79]. These tools make it feasible to train sophisticated architectures capable of performing tasks such as lesion detection, segmentation of anatomical structures, image reconstruction, and cross-modality synthesis. Within this broader context, three families of DL models have been particularly influential: CNNs, which underpin most spatial analysis in two-dimensional images; Recurrent neural networks (RNNs) and related sequence models, used for temporal or sequential data; and Generative adversarial networks (GANs), which support image enhancement, synthesis, and domain translation. These architectures form the methodological basis for the DL approaches used in later chapters of this thesis, particularly for improving UTI classification performance, harmonising variability in the FoV and tongue representation, and enhancing clinical interpretability.

Convolutional Neural Networks

CNNs are the most widely used DL architecture for image-based tasks and form the foundation of many advances in medical imaging. Their strength lies in their ability to learn hierarchical visual features directly from raw pixels. Early layers capture simple patterns such as edges or texture, while deeper layers learn higher-level structures relevant to classification or segmentation. A typical CNN contains convolutional layers that extract local spatial features, pooling layers that summarise information and reduce dimensionality [80], and fully connected layers that combine learned features to make predictions [81]. This architecture enables CNNs to model spatial relationships effectively, even in noisy imaging domains such as UTI. An overview of how a CNN processes an ultrasound frame is shown in Figure 2.3, illustrating how local image patches are passed through successive convolutional and pooling layers to extract hierarchical spatial features used for classification.

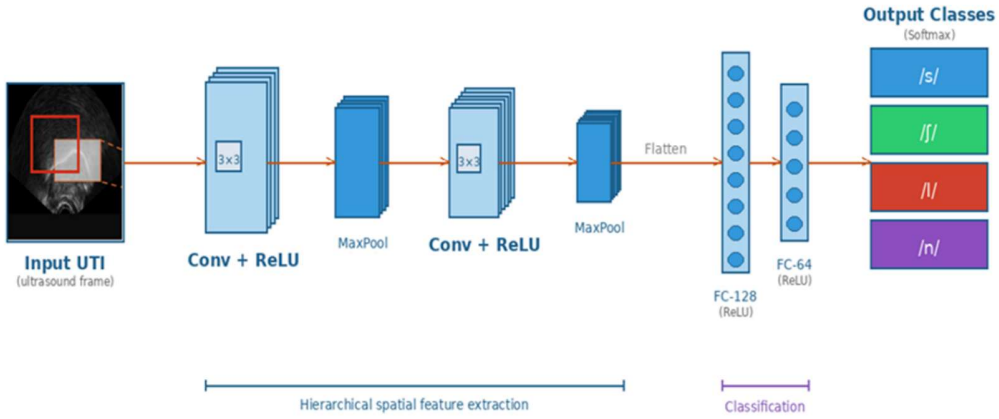

Figure 2. 3: Illustration of a CNN Processing an UTI.

CNNs are particularly well-suited to UTI, where subtle articulatory cues are embedded within noisy, low-contrast images. Their capacity to learn discriminative features directly from raw pixel intensities makes them effective for both classification tasks and generative frameworks, including the harmonisation of variable imaging conditions explored later in this thesis.

This ability to extract stable, informative representations from challenging image data motivates their use in the baseline and generative models developed in Chapters 3–5.

Recurrent Neural Networks

RNNs are DL models designed to process sequential or time-dependent data by maintaining an internal memory of previous inputs [82], [83]. RNNs contain cyclic connections that allow information to persist across time steps, enabling the modelling of temporal dependencies. This makes them particularly well-suited to tasks where the ordering of information is crucial, such as language modelling, speech recognition, or the analysis of connected-speech samples in clinical linguistics and phonetics [84]. However, RNNs are less commonly applied to UTI, which is often analysed on a frame-by-frame basis rather than as long temporal sequences. Because the work in this thesis focuses on spatial pattern extraction, image representation, and generative modelling of static UTI frames, RNNs do not form a central component of the methods developed here. Instead, the thesis relies primarily on convolutional architectures and GANs, which are better suited to modelling spatial structure and performing image-to-image translation.

Generative Adversarial Networks

GANs are a family of DL models designed to learn the underlying distribution of a dataset to generate new, realistic samples [85]. A GAN consists of two components trained in opposition: a generator, which produces synthetic images, and a discriminator, which attempts to distinguish generated images from real ones. Through this adversarial process, the generator progressively improves until its outputs closely resemble the true data distribution [86]. In medical imaging, GANs have demonstrated strong performance in areas such as image synthesis, super-resolution, artefact reduction, and cross-modality translation [87], [88], [89]. Synthetic images can supplement small annotated datasets, helping reduce manual labelling requirements and improving model generalisability [90]. GANs are also routinely used for data augmentation, creating realistic variations of existing images to enhance robustness to noise, occlusion, and anatomical variability. Another important capability of GANs is image-to-image translation, in which the model learns to map one type of image to another, for example, generating synthetic CT images from MRI or enhancing low-quality ultrasound frames [91], [92].

This property is particularly relevant for UTI, where differences in the FoV introduce substantial visual variability. GAN-based translation methods offer a principled approach to harmonising these differences, forming the methodological foundation for the generative standardisation pipeline developed later in this thesis.

2.1.6.1 Training Strategies

DL models can be trained under different learning paradigms depending on the availability of labelled data and the nature of the task.

- **Supervised learning:** Supervised learning remains the most widely used paradigm in medical imaging, relying on labelled datasets to train models to predict specific targets [93]. In UTI, supervised CNNs have been applied to tasks such as tongue-contour segmentation [94] and classification of articulatory gestures. However, supervised learning is heavily constrained by annotation cost, as ground truth labels must be provided by trained clinicians. The limited size of available UTI datasets further increases the risk that models learn speaker- or session-specific patterns rather than generalisable articulatory features, reducing cross-speaker robustness. These limitations correspond directly to Challenge **C1** (variability and generalisation limitations) and Challenge **C2** (data scarcity and annotation efficiency), which motivate the methodological developments explored later in this thesis.
- **Unsupervised learning:** Unsupervised learning focuses on discovering structure within unlabelled data, aiming to identify meaningful patterns or relationships without external guidance [95], [96]. The goal of unsupervised learning is to enable models to autonomously learn feature hierarchies that capture the intrinsic organisation of the data. Such representations support a range of downstream tasks. This approach is particularly valuable in clinical linguistics and phonetics, where large, annotated corpora can be difficult and costly to obtain.

- **Self-supervised learning (SSL):** SSL learning has recently emerged as a powerful paradigm in computer vision and medical imaging [97], [98]. Unlike supervised learning, it does not rely on externally annotated labels. Instead, it exploits the internal structure of the data to create pretext tasks that generate surrogate labels from the images themselves. By solving these tasks, the model learns representations that capture essential visual patterns and anatomical structure. SSL is particularly attractive for UTI because labelled datasets are scarce and expensive to obtain, requiring expert clinicians to annotate tongue contours or articulatory categories [19],[99].
- **Transfer Learning:** Transfer learning is a widely used strategy in DL that adapts models pretrained on one task to new but related problems. Instead of training from scratch, which requires large amounts of labelled data, researchers fine-tune existing models that have already learned useful visual features from large datasets such as ImageNet [92]. Transfer learning can be implemented in different ways: (1) full fine-tuning, where all pretrained weights are updated during training on the new task; (2) partial fine-tuning (or feature extraction), where early layers are frozen to retain low-level features (e.g., edges, textures) while deeper layers are retrained to learn task-specific representations. Or (3) training from scratch, where the model architecture may be borrowed but all weights are randomly initialised and learned entirely from the target dataset. This approach offers significant advantages in medical imaging [25], where it can accelerate convergence, reduce computational costs, and mitigate overfitting, particularly valuable given that annotated medical datasets are often scarce and costly to obtain [72]. Transfer learning has been successfully applied to various medical imaging tasks, including detecting diabetic retinopathy, classifying skin cancer, and segmenting brain tumours [100], [101]. However, transfer learning for ultrasound imaging presents unique challenges due to fundamental domain mismatch: natural image features (edges, textures, RGB colour channels) differ markedly from ultrasound-specific characteristics (speckle noise, acoustic shadows, grayscale intensity patterns, and tongue-specific anatomical structures).

This mismatch can lead to negative transfer, where pretrained knowledge degrades rather than improves performance, as low-level features learned from natural images may not meaningfully represent ultrasound tissue boundaries or tongue contours. While training from scratch would allow networks to learn representations inherently suited to UTI characteristics, this approach is often impractical due to limited annotated ultrasound datasets and high computational demands, particularly problematic for real-time deployment on resource-constrained mobile devices. Partial fine-tuning offers a pragmatic compromise: by freezing early convolutional layers (which may still capture useful low-level visual patterns) and retraining only the deeper layers on ultrasound data, this strategy reduces both the data requirements and computational cost while allowing the model to adapt to domain-specific features. This approach is especially suitable for mobile health applications where model efficiency and inference speed are critical considerations.

Each training strategy offers distinct advantages, but the constraints of UTI, small, labelled datasets, high inter-speaker variability, and the cost of expert annotation make supervised learning, AL, and transfer learning the most practical approaches for this domain. These methods form the foundation of the classification, representation analysis, and data-efficiency frameworks developed in the subsequent chapters of this thesis.

2.1.6.2 Evaluation Metrics

Evaluating DL-based UTI analysis requires metrics that capture both technical performance and clinical relevance. The choice of metric depends on the specific task (classification, segmentation, or image generation) and clinical priorities, such as minimising false negatives in disorder detection or ensuring anatomical accuracy in tongue contour extraction [102]. For segmentation, measures such as the Dice coefficient and the Mean Surface Distance (MSD) quantify how closely a predicted anatomical boundary matches expert annotations [103].

In classification, accuracy is a fundamental measure of a classifier's overall correctness, defined as the proportion of correctly predicted instances relative to the total number of instances:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \qquad (2.1)$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. For medical image analysis, accuracy alone is insufficient, particularly in datasets with class imbalances. Additional metrics such as precision, recall, and F1-score are therefore employed:

$$\text{Precision} = \frac{(\text{TP})}{(\text{TP} + \text{FP})} \qquad (2.2)$$

$$\text{Recall} = \frac{(\text{TP})}{(\text{TP} + \text{FN})} \qquad (2.3)$$

$$\text{F1} = 2 \text{ x } \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \qquad (2.4)$$

For generative models, PSNR and SSIM assess image quality and perceptual fidelity [104]. These metrics enable rigorous evaluation of the models developed throughout this thesis.

2.1.6.3 Interpretability in Deep Learning

DL models are often criticised as “black boxes,” a concern that is especially relevant in medical imaging, where clinicians need to understand why a model makes a particular decision. Interpretability is therefore essential for building trust, identifying failure modes, and ensuring that models focus on clinically meaningful features rather than spurious patterns. Interpretability methods in imaging broadly fall into attribution-based approaches, which highlight the image regions that most influence a prediction, and surrogate-based approaches, which approximate model behaviour through simplified explanations. For UTI, these tools are particularly important.

Because UTI images are noisy and contain shadows and surrounding tissue, clinicians need confidence that models are attending to the tongue surface and articulatory patterns rather than irrelevant background artefacts.

Among attribution methods, Grad-CAM++ are widely used because they provide intuitive heatmaps showing where a model "looks" when forming its prediction [105], [106]. Grad-CAM++ offer finer localisation and have been applied successfully across several clinical imaging tasks [27]. In UTI specifically, Grad-CAM++ has been highly valuable for exposing when models trained on raw images attend to artefact regions, a finding that directly motivated the ROI-based and masked-image representations developed later in this thesis (Chapter 4). These visual explanations help validate model behaviour, guide representation design, and support the broader goal of making DL systems more transparent and clinically interpretable. Although other interpretability approaches exist, such as local interpretable model-agnostic explanation (LIME) [26] and SHapley additive exPlanations (SHAP) [107], their computational overhead makes them less practical for real-time or high-volume ultrasound workflows. Consequently, this thesis focuses on Grad-CAM++ as an accessible and clinically meaningful tool for examining how models process UTI data.

2.2 Deep Learning in Ultrasound Tongue Imaging

This section summarises the methodological framework used to synthesise existing research at the intersection of DL and UTI, as adapted from the author's published systematic review [33]. The original review adhered to the PRISMA 2020 guidelines to ensure transparency, reproducibility, and comprehensive coverage. This synthesis provides not only a summary of prior work but also a critical appraisal of how close existing approaches are to clinical translation, thereby identifying priorities that shape the remainder of this thesis.

The systematic review was guided by three overarching research questions:

Q1. How have DL approaches been applied to UTI to support the detection and assessment of SSDs?

Q2. What key technical and clinical challenges have been identified in these studies that constrain the effectiveness and scalability of DL for ultrasound-based SSD diagnosis?

Q3. What further methodological and translational advancements are required to progress toward clinically deployable, automated UTI systems for diagnostic and therapeutic support?

To complement these questions, Figure 2.4 presents a high-level conceptual framework illustrating how DL can be integrated into the UTI workflow, from data acquisition and preprocessing, through task-specific modelling and interpretation, to clinician-in-the-loop feedback for assessment and therapy.

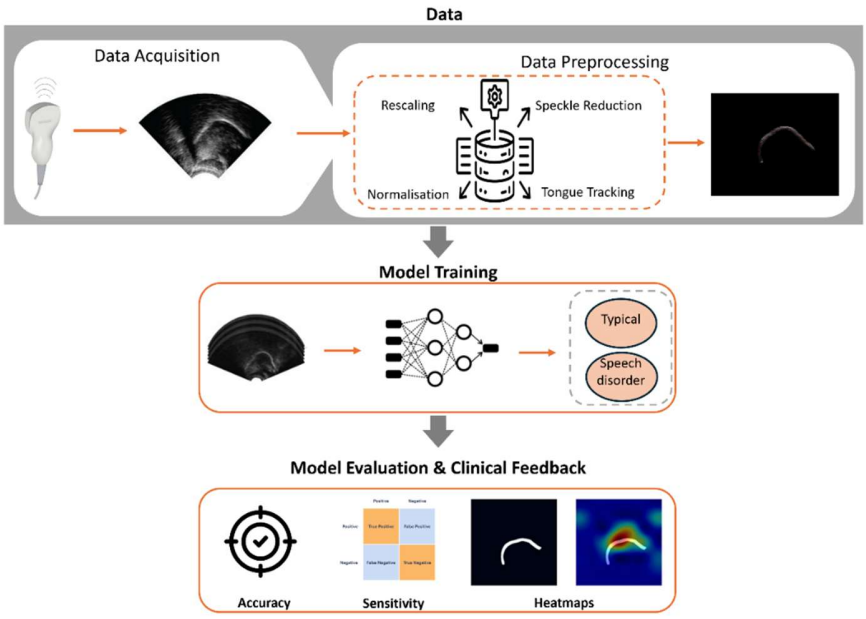

Figure 2. 4: Overview of DL for UTI in SSD Assessment.

Search Strategy

A systematic search was conducted across six major databases, IEEE Xplore, PubMed, ScienceDirect, Scopus, Taylor & Francis Online, and arXiv, to identify studies published between 2010 and 2025. The query design combined terms reflecting three key concepts: *speech disorders*, *ultrasound tongue imaging*, and *deep learning*. Representative Boolean strings included: ("speech sound disorder" OR "articulation disorder") AND ("ultrasound tongue imaging" OR "lingual ultrasound") AND ("deep learning" OR "neural network" OR "CNN" OR "transformer"). Additional synonym clusters ("phoneme classification," "articulatory imaging," "silent speech interface") were incorporated to ensure comprehensive retrieval.

Screening and Eligibility

The search yielded 112 unique records, which were screened in two stages:

1. Title and abstract screening excluded studies unrelated to both UTI and DL.
2. Full-text evaluation applied inclusion and exclusion criteria (Table 2.2), resulting in 11 eligible studies.

Table 2. 2: Inclusion and Exclusion Criteria.

Inclusion Criteria	Exclusion Criteria
Applied a DL method to UTI data.	Focused solely on acoustic or non-ultrasound modalities.
Addressed speech-related tasks (e.g., classification, contour extraction).	Used ultrasound for non-speech applications.
Involved participants (typically developing (TD) or with SSDs).	Grey literature or unverified preprints.
Peer-reviewed, English-language publication (2010–2025).	

Data Extraction and Study Classification

For each included article, key information was extracted, including study aims, DL architecture, input modality, dataset characteristics, evaluation metrics, and clinical relevance. Following PRISMA conventions, the selection process was documented in a flow diagram illustrated in Figure 2.5.

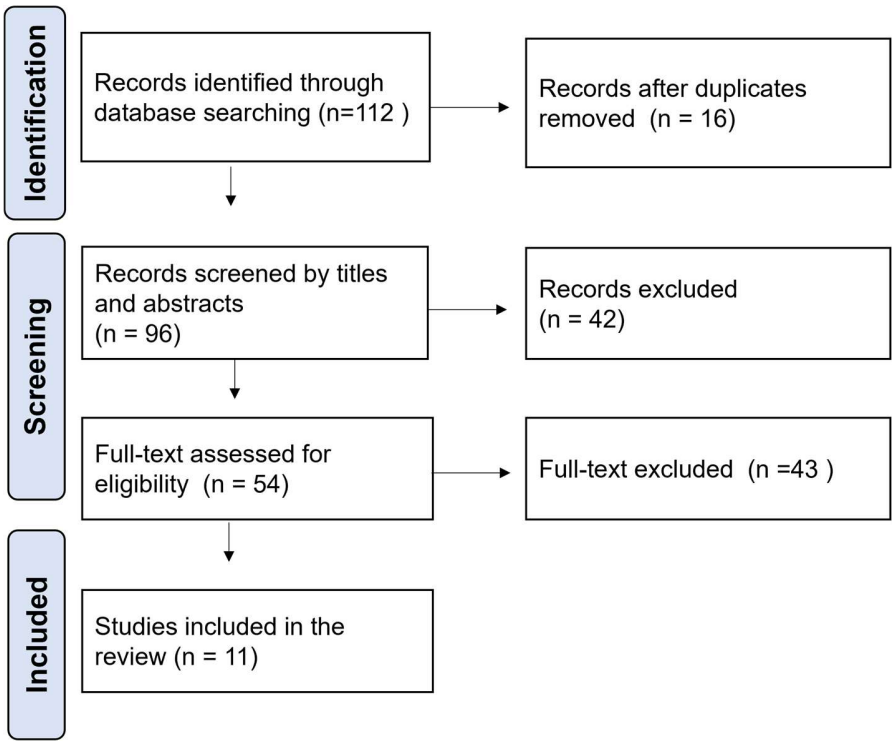

Figure 2. 5: PRISMA Flow Diagram of the Study Selection Process.

Figure 2.5 illustrates the systematic study selection process following PRISMA 2020 guidelines. The diagram traces the four sequential stages of the review: identification, screening, eligibility assessment, and final inclusion. In the identification stage, records were retrieved from electronic databases. Duplicate records were removed before screening, during which titles and abstracts were evaluated against the predefined inclusion and exclusion criteria. Full-text articles were then assessed for eligibility, with reasons for exclusion recorded at this stage.

The final included studies met all criteria: reporting DL-based analysis of UTI or closely related articulatory imaging modalities, involving participants with SSDs or typically developing speech, and published as peer-reviewed English-language articles between 2010 and 2025. The resulting set of included studies forms the evidence base synthesised in the sections that follow. To structure the synthesis, the studies were grouped into three application domains:

1. Direct SSD Detection and Phonetic Classification**:** studies explicitly classifying speech errors or phonetic segments (Section 2.2.2A).
2. Foundational Tools: studies developing core components such as tongue contour segmentation or motion prediction (Section 2.2.2B).
3. Clinical Implementation Context: examination of UTI use in therapy workflows (Section 2.2.2C).

These three domains correspond to the pipeline stages necessary for clinical translation: (1) end-to-end classification of articulatory patterns, (2) foundational techniques enabling robust automated analysis, and (3) integration into real-world therapeutic contexts. This structure mirrors the thesis progression from baseline modelling (Chapter 3) through data representation and harmonisation (Chapters 4-5) to annotation optimisation (Chapter 6).To contextualise these technical contributions within real-world practice, Section 2.3.2C additionally discusses clinical implementation considerations drawn from Cleland (2023) [108], which, while not employing DL methods, identifies the practical barriers that automated systems must address.

1. Direct SSD Detection and Phonetic Classification: studies explicitly classifying speech errors or phonetic segments in disordered or typical speech (Table 2.3).

Table 2. 3: Summary of Direct SSD Detection Studies.

Study (Year)	DL Model	Dataset	Input Type	Task	Metric	Clinical Relevance
Ribeiro et al. [29]	CNN	Ultrax Typically Developing (UXTD), Tongue and Lips (TaL), Ultrax SDD (UXSSD), and Ultraphonix (UPX)	Raw UTI + audio	Error detection	Accuracy	High
Ribeiro et al. [16]	CNN+ DNN	UXTD	Raw UTI	Phoneme classification	Accuracy	Medium
Al Ani et al. [109]	CNN	UXTD	Raw UTI + texture features	Phonetic segment classification	Accuracy	Medium
Xu et al. [19]	Transformer	UXTD	Raw UTI	Phonetic segment classification	Accuracy	Medium
You et al. [110]	Transformer	UXTD	Raw UTI	Phonetic segment classification	Accuracy	Medium
Dan et al. [111]	Transformer	UXTD	Raw UTI	Phonetic segment classification	Accuracy	Medium

Several patterns emerge across these classification studies. First, all models rely heavily on the UltraSuite corpus [112], underscoring both its central importance and the field's dependence on a single paediatric UTI resource. Second, speaker-independent performance consistently lags speaker-dependent accuracy, highlighting persistent generalisation challenges aligned with Challenge **C1**. Third, while most approaches report strong results on TD child speech, validation on disordered speech remains sparse, limiting conclusions about clinical applicability for SSD assessment.

2. Foundational Tools: studies developing core components such as tongue contour segmentation or motion prediction (Table 2.4).

Table 2. 4: Summary of Technical Foundations Studies.

Study (Year)	DL Model	Dataset	Input Type	Task	Metric	Clinical Relevance
Mozaffari & Lee [113]	U-Net	Ottawa UTI Corpus	UTI	Tongue contour segmentation	Dice	Medium
Mozaffari et al. [114]	TonguNet	Ottawa UTI Corpus	UTI	Landmark tracking	MSD	Medium
Li et al. [115]	U-Net	NS, TJU, TIMIT	UTI	Tongue contour extraction	Intersection over Union (IoU)	Medium
Mukai et al. [116]	U-Net	Institutional UTI	UTI	Tongue surface extraction	Accuracy	Low
Zhao et al. [117]	ConvLSTM	Silent Speech Dataset	UTI-Video Sequence	Tongue motion prediction	Mean squared error (MSE)	Low

This categorisation provides a coherent framework linking algorithmic advances to clinical translation. It also reflects the thesis's own trajectory, from baseline modelling (Chapter 3) through data and representation optimisation (Chapters 4–5) to practical deployment (Chapter 7).

2.2.1 Scope and Corpus of the Review

The systematic review encompassed eleven peer-reviewed studies published between 2010 and 2025 that applied DL to UTI for speech analysis or disorder assessment. Although modest in number, detection and span the complete technological spectrum, from segmentation to direct error detection, and together outline the emerging pipeline of automated SSD evaluation.

A central limitation across the reviewed literature is the scarcity of large, annotated paediatric UTI datasets, particularly for disordered speech. The UltraSuite corpora [112] constitute the principal open benchmark for current research: UltraSuite-UXTD contains recordings of TD children, while UltraSuite-UXSSD and UltraSuite-UPX include children with SSDs, and the UltraSuite-cleft dataset contains data gathered from children with CP±L. These datasets have driven much of the methodological progress in the field but remain restricted in scale, diversity, and error coverage, thereby constraining model generalisation. To clarify this landscape, Table 2.5 summarises the key corpora used in the reviewed studies, and Table 2.6 maps each study to its corresponding datasets. As shown, most classification studies rely on UXTD, whereas resources featuring SSDs are smaller and often require additional expert annotation.

Table 2. 5: Datasets Used by the Included Studies.

Dataset	Population	# Speaker	Mean age	Language	Modality	Availability
UXTD	TD children	58	~9y 3m	Scottish English	UTI + audio	Public
Ultrasuite-TaL	Adults	82	N/A	English	UTI + video images of the lips	Public
UXSSD	SSD children	8	~7y 7m	Scottish English	UTI + audio	Public
UPX	SSD children	20	~8y 4m	Scottish English	UTI + audio	Public
Ottawa UTI Corpus	TD (children)	n/a	n/a	n/a	UTI	Private
NS, TJU, TIMIT-UTI	TD (adults and children)	n/a	n/a	n/a	UTI	Private
WSJ0-TJU	TD (adults)	n/a	n/a	n/a	UTI video	Private
Institutional coronal UTI	TD (children)	n/a	n/a	English	UTI	Private

Table 2. 6: Mapping from Included Studies to the Dataset.

Study	Task	Dataset used	Population
Ribeiro et al. [29]	Error detection	UXTD, TaL, UXSSD, UPX	TD + SSD Children
Ribeiro et al. [16]	phoneme classification	UXTD	TD children
Al Ani et al. [109]	phoneme classification	UXTD	TD children
Xu et al., [19]	phoneme classification	UXTD	TD children
You et al., [110]	phoneme classification	UXTD	TD children
Dan et al., [111]	phoneme classification	UXTD	TD children
Mozaffari & Lee, [113]	Tongue segmentation	Ottawa UTI Corpus	Mixed
Mozaffari et al. [114]	Landmark tracking	Ottawa UTI Corpus	Mixed
Li et al., [115]	Tongue segmentation	NS, TJU, TIMIT-UTI	Mixed
Zhao et al., [117]	Motion prediction	WSJ0-TJU	Adults
Mukai et al., [116]	Segmentation for 3D modelling	Institutional coronal UTI	Children

2.2.2 Application Domains and Representative Studies

A. Direct SSD Detection and Phonetic Classification

A central motivation for applying DL to UTI is the automation of articulatory assessment, specifically determining whether a child's speech sound production is typical or contains clinically relevant errors. Automating this process has the potential to support SSD diagnosis, reduce reliance on subjective perceptual judgement, and enable objective monitoring of therapy outcomes. The systematic review underpinning this chapter identified six studies that directly addressed SSD or phonetic segment classification, collectively demonstrating that UTI encodes sufficient articulatory information to distinguish between speech sounds and detect certain misarticulations.

One of the earliest clinically focused studies was conducted by Ribeiro et al. [29], who investigated the automated detection of articulation errors in children using UTI. Their work targeted error patterns of high clinical relevance, including velar fronting and R-errors misarticulation, in Scottish English-speaking children. To enhance model robustness, the training data combined the UltraSuite-UXTD corpus with adult speech from the TaL dataset. Model evaluation incorporated both TD and atypical speech samples, with ground-truth labels provided by experienced SLTs based on synchronised ultrasound and audio recordings. The proposed CNN model fused ultrasound frames with acoustic features, revealing that the ultrasound modality contributed substantially to detecting place-of-articulation errors. The system achieved 86.9% accuracy for phonetic classification in TD speech and correctly identified 86.6% of velar-fronting errors annotated by experts. In contrast, detection of /r/-sound errors was less reliable, likely reflecting low inter-rater agreement among annotators rather than limitations of the imaging modality itself. This study established proof-of-concept for DL-based UTI in clinical error detection, distinguishing it from other work focused solely on typical speech.

Recognising that anatomical variability limits model transferability, Ribeiro et al [16] investigated the classification of tongue shapes from raw ultrasound under different speaker scenarios. They examined how model performance degrades when applied to previously unseen child speakers and evaluated strategies to improve speaker-independent generalisation, which is critical for clinical deployment.

Using the UltraSuite-UXTD dataset, they framed the task as four-class phonetic segment classification based on place of articulation. They investigated various preprocessing methods, including intensity normalisation and dimensionality reduction, along with two classifiers, CNN and DNN, to assess their approach. Results indicated that models performed poorly on unseen speakers, but adding minimal speaker information, such as the mean ultrasound frame, significantly improved generalisation. However, the mean ultrasound frame approach has inherent limitations that constrain its clinical applicability. First, it assumes stable probe positioning throughout the recording session, an assumption that may not hold in paediatric settings where probe movement is common, particularly without stabilisation equipment. Such movement can render the mean frame spatially inaccurate as an anatomical reference. Second, the approach is session-specific: any change in probe placement between sessions requires recalculating a new mean frame, limiting true generalisability. Finally, the mean frame must be computed from a complete or representative sample of speaker data before classification can proceed, which may hinder real-time clinical feedback applications.

Al Ani et al. [109] further investigated automatic phoneme classification in TD child speech using raw UTI, proposing a DL framework that integrates complementary visual representations. Using data from the UltraSuite-UXTD corpus, the study categorised speech sounds into four place-of-articulation classes and extracted texture information from ultrasound frames using LBP, a descriptor well-suited to capturing ultrasound-specific texture variations. Multiple architectures were evaluated, including standard CNNs, DNN, and transfer-learning models (ResNet-50 and Inception-V3). In addition, a novel dual-stream architecture, FusionNet, was introduced. FusionNet processes raw ultrasound frames through a CNN branch to capture global tongue-shape information, while a parallel branch encodes LBP-based texture features using a fully connected network. These streams are subsequently fused and optimised jointly for classification. Evaluation under speaker-dependent (91.9%), multi-speaker (92.1%), and speaker-independent (82.3%) conditions showed that FusionNet consistently outperformed all baseline models. The multi-modal approach demonstrated that texture features complement CNN-learned representations, though the approximately 10pp drop in speaker-independent performance confirmed the persistent generalisation challenge identified in earlier work [16].

More recent work has shifted toward SSL to address annotation scarcity while improving cross-speaker generalisation. Three studies have progressively advanced masked reconstruction frameworks, demonstrating substantial gains in annotation efficiency and speaker-independent robustness. You et al. [110] established masked image modelling as a highly effective paradigm for UTI analysis by pretraining a Vision Transformer (ViT) encoder to reconstruct randomly masked patches from unlabelled ultrasound frames. During pretraining, the model learned robust articulatory representations by inferring masked regions from visible context. When fine-tuned on ~10,700 labelled examples for four-class phonetic classification, the system achieved accuracies of 88.10% (speaker-dependent), 84.82% (multi-speaker), 83.72% (speaker-independent), and 88.94% (speaker-adapted), an average improvement of 13.3% over contrastive SSL baselines (SimSiam).

Building on this foundation, Xu et al. [118] refined the masked modelling approach by systematically evaluating three masking strategies (random patch, vertical, and horizontal) and incorporating hard-example mining during fine-tuning. The framework employed a ViT-Large encoder pretrained at a 75% masking ratio, combined with a dual-loss fine-tuning strategy that prioritised challenging examples, yielding accuracies of 86.60% (speaker-dependent), 85.85% (multi-speaker), 85.18% (speaker-independent), and 90.00% (speaker-adapted) on the UXTD dataset. Notably, vertical masking performed best overall (86.40% mean accuracy), suggesting that preserving horizontal anatomical continuity while forcing reconstruction of vertical structures may be particularly effective for tongue imaging. Hard example mining contributed an additional 2.84% improvement, validating its utility for noisy medical imaging data.

Expanding on these advances, Dan et al. [111] extended masked modelling to the spatio-temporal domain. Their framework introduced two key innovations: (1) spatio-temporal masking, where pixel blocks are masked identically across multiple consecutive frames, forcing the model to leverage temporal context for reconstruction; and (2) a Token Shift Module integrated into the encoder to propagate information across adjacent frames, enabling explicit modelling of frame-to-frame articulatory transitions. Pretrained on unlabelled UTI sequences and fine-tuned on UXTD, the system achieved accuracies of 90.32% (speaker-dependent), 86.45% (multi-speaker), 85.27% (speaker-independent), and 90.11% (speaker-adapted).

The model remained stable at masking ratios up to ~75%, substantially higher than typical image-only masked autoencoders (~50-60%), indicating that temporal redundancy in speech articulation sequences provides rich contextual cues. By explicitly capturing dynamic tongue movements, critical for distinguishing phonetic segments characterised by articulatory trajectories rather than static configurations, this work demonstrated how motion-aware self-supervision can further enhance generalisation and label efficiency. Table 2.7 summarises the included studies used to classify SSDs.

Table 2. 7: Comparative Analysis of DL-Based UTI Studies for Classifying SSDs.

Study	Task & Data	Key Contributions	Limitations / Open Issues
Ribeiro et al. [96]	Error detection; UXTD, TaL, UXSSD, UPX	First clinically focused UTI-based error detection study; showed ultrasound contributes strongly to place-of-articulation error detection	Lower reliability for /r/ errors due to annotation inconsistency; clinician interpretation still required
Ribeiro et al. [16]	phonetic classification; UXTD	Early demonstration of CNN-based phonetic classification from raw UTI; explored PCA/DCT representations and speaker-mean normalisation	Poor speaker-independent performance
Al Ani et al. [109]	Phoneme classification; UXTD	Introduced FusionNet, combining shape and texture features	Still notable performance drop for unseen speakers; focused on TD speech
Xu et al. [19]	Phoneme classification; UXTD	Masked image modelling reduced annotation dependence	Operates on raw inputs; interpretability not addressed; no explicit clinical deployment
You et al. [110]	phoneme classification; UXTD	Demonstrated transformer-based self-supervised pretraining; improved generalisation across speaker scenarios	Computational complexity; still sensitive to acquisition variability; outputs remain a black box
Dan et al. [111]	phoneme classification; UXTD	Integrated temporal dynamics via masked autoencoding; stable performance under high masking ratios	Requires sequential data; clinical applicability and interpretability not evaluated

Overall, these studies demonstrate that DL can extract clinically relevant articulatory information from UTI, but they also reveal three unresolved challenges that motivate this thesis. First (**C1** data variability and generalisability limitations), model performance consistently declines for unseen speakers, indicating sensitivity to anatomical differences and acquisition variability. Second (**C2** data scarcity and annotation efficiency), most approaches rely on limited expert-labelled datasets; although SSL reduces label dependence, it does not address variability in image acquisition. Third (**C3** lack of interpretability and clinical usability), no studies verify model interpretability, limiting clinical trust. Addressing these interlinked challenges requires integrated solutions that combine data harmonisation, annotation-efficient learning, and explainable modelling, which form the focus of the chapters that follow.

B. Foundational Tools – Tongue Segmentation and Motion Modelling

In addition to direct speech-sound classification, several studies have focused on foundational technical components that underpin fully automated ultrasound-based speech analysis pipelines. Two such components are particularly critical: tongue segmentation, which involves identifying the tongue surface in each ultrasound frame, and motion modelling, which characterises how tongue shape evolves. Together, these tasks enable the transformation of raw, noisy ultrasound data into structured and dynamic articulatory representations suitable for downstream analysis, including classification and visual biofeedback.

Several studies have addressed the challenge of automatic tongue contour extraction using DL. Reliable segmentation is a prerequisite for many articulatory analyses, as it converts ultrasound frames into explicit representations of tongue shape. Earlier semi-automatic approaches, such as EdgeTrak, were sensitive to noise and frequently required manual intervention, limiting scalability. DL-based methods offer a data-driven alternative capable of learning robust representations despite the low contrast and speckle artefacts inherent in UTI.

Two related studies from the same research group tackled automatic contour extraction using different CNN architectures. Mozaffari and Lee [119] introduced BowNet and wBowNet, encoder–decoder models built upon VGG-16 with standard and dilated convolutions to capture both local and global context. wBowNet included tighter cross-scale integration to enhance spatial detail. Trained on the University of Ottawa UTI corpus and the Seeing Speech database, both models achieved strong segmentation accuracy (mean Dice ≈ 0.85).

Building on this foundation, Mozaffari et al. [114] proposed TongueNet, a lightweight CNN that tracks discrete landmarks on the tongue surface rather than segmenting the entire contour region. This novel approach eliminates the need for post-processing steps required in traditional segmentation pipelines, enabling faster real-time performance (67 fps). Trained on 2,000 images from the UOttawa database using 10 randomly selected landmarks, TongueNet achieved MSD of 4.87 pixels, comparable to U-Net-based methods while offering superior computational efficiency. Critically, this landmark-based approach enables researchers to leverage legacy databases originally annotated with sparse point sets, expanding the pool of available training data. When tested on the unseen database without additional training, TongueNet demonstrated strong generalisation, confirming its robustness to cross-dataset variation. Together, these models demonstrated efficient, accurate segmentation suitable for research and potential clinical deployment, with TongueNet offering advantages for real-time visual feedback applications.

Li et al. [115] advanced this approach through wUNet, an enhanced U-Net variant incorporating VGG-16-initialised encoding, extra skip connections, and multi-level feature fusion to better integrate low-level edge information with high-level shape representations. Evaluated on three independent datasets (NS, TJU, TIMIT-UTI) spanning different imaging protocols, wUNet achieved IoU ~98% and Dice ~94%, substantially outperforming baseline U-Net IoU ~91% and U-Net++ IoU ~93%. While developed for silent-speech interfaces, wUNet's near-human accuracy and generalisability make it applicable to clinical articulatory analysis, providing a reliable contour extraction module for error detection pipelines.

Mukai et al. [116] contributed a complementary perspective by examining how annotation design influences segmentation performance, particularly in the context of three-dimensional tongue modelling. Using a small institutional dataset of coronal ultrasound images (19 cross-sections, augmented to ~7,700 images), they compared training strategies based on sparse point annotations versus spline-based supervision. A U-Net-based contour-point extractor trained with spline-derived labels achieved substantially higher accuracy (91.7% vs. a lower baseline) and superior subjective acceptability in reconstructed 3D tongue models. This confirms that annotation fidelity, geometric consistency, and boundary smoothness directly impact segmentation performance, with practical implications for dataset construction strategies. High-quality, carefully curated labels are preferable to rapid but imprecise annotations, suggesting that investment in annotation tools and protocols is essential for training reliable clinical systems.

Beyond static segmentation, Zhao et al. [117] explored articulatory motion modelling using convolutional long short-term memory (ConvLSTM) networks. Rather than analysing individual frames independently, their approach predicted future ultrasound frames from preceding sequences, capturing the temporal dynamics of tongue movement. Trained on datasets derived from WSJ0 and TJU corpora, the ConvLSTM model consistently outperformed a 3D-CNN baseline in frame prediction accuracy and structural similarity. While ConvLSTM models exhibited slight smoothing effects when predicting explicit contours, they captured overall motion trajectories with high fidelity. Although developed for silent speech interfaces, this work highlights the importance of temporal modelling for representing coarticulation and articulatory dynamics, which are critical for assessing motor speech disorders and atypical speech patterns. Table. 2.8 summarises the contributions and limitations of the related work.

Table 2. 8: Comparative Analysis of Foundational UTI Studies.

Study	Task	Key Contributions	Limitations / Open Issues
Mozaffari & Lee [119]	Tongue contour segmentation	Real-time automatic contour extraction	Performance degrades across datasets; relies on high-quality annotations; not integrated with downstream classification
Mozaffari et al. [114]	Landmark tracking	Real-time tongue tracking	Only 10 landmarks; limited cross-dataset testing; no SSD validation
Li et al. [115]	Tongue segmentation	The model achieved near-human accuracy, strong robustness across datasets	Focused on silent speech
Mukai et al. [116]	Annotation design for segmentation; institutional dataset	Showed annotation strategy strongly affects learning; spline-based supervision improves contour quality	Small dataset; indirect relevance to real-time clinical workflows
Zhao et al. [117]	Motion prediction	Demonstrated temporal modelling of articulatory dynamics	Motion smoothness may suppress fine detail

These segmentation and motion-modelling techniques provide essential preprocessing capabilities for the classification systems reviewed in Section 2.2.2A. The progression from standard U-Net to enhanced architectures demonstrates that careful architectural design and multi-dataset training can approach human-level accuracy. The emergence of landmark-based alternatives (TongueNet) offers computational efficiency suitable for real-time applications, albeit with sparser spatial representation. Temporal modelling extends analysis beyond individual frames to capture articulatory dynamics.

However, two gaps remain. First, no study systematically evaluated robustness to common UTI variability (e.g. probe angle variation) prevalent in clinical settings. Second, integration with downstream classification has not been empirically validated; it remains unclear whether segmentation errors propagate to classification errors or whether classifiers trained on raw images are robust to contour imprecision. This integration gap motivates the investigation of image representations and their impact on classification performance in Chapter 4.

C. Clinical Implementation: UTI in Practice for SDD Therapy

The eleven studies reviewed above demonstrate substantial technical progress in applying DL to UTI for speech analysis, yet their translation into clinical practice depends on understanding real-world therapy workflows and identifying barriers to adoption. Cleland (2023) [108], though not employing DL methods and therefore not included in the formal systematic review, provides essential insight into how UTI is currently used with children with CP±L and what requirements automated systems must meet to support clinical practice.

Drawing on detailed case analyses, Cleland describes how UTI functions as a visual biofeedback tool during articulation therapy following palate repair. Real-time midsagittal imaging allows clinicians to observe tongue shape and placement directly, revealing atypical or covert articulations that are inaudible in acoustic assessment. For example, children with cleft-related SSDs frequently exhibit posterior or double articulations that may sound correct but involve abnormal lingual postures. Ultrasound enables these gestures to be visualised and explicitly corrected. Therapists can use the live image to guide a child toward a more typical constriction location, effectively linking articulatory intent to visual evidence.

Cleland reported that UTI feedback increased children's awareness of tongue placement and facilitated correction of entrenched misarticulations. Nevertheless, the study highlighted several barriers to clinical adoption. Interpretation of ultrasound images demands substantial expertise: therapists must simultaneously analyse the dynamic display, provide verbal feedback, and maintain the child's engagement. Automation could alleviate this cognitive load; an intelligent system that automatically detects and highlights articulatory features or deviations in real time would support both therapist and client by providing objective, immediate cues. Practical constraints were also noted. Although UTI is non-invasive and child-friendly, it requires specialised equipment and trained clinicians, limiting its availability in routine speech-language therapy. Cleland's findings therefore clarify the kinds of innovations that would enhance clinical viability: simplified user interfaces, automatic annotation or segmentation, and quantitative progress tracking. These requirements align directly with the technical advances explored in the reviewed studies, automated contour extraction (Table 2.8), phonetic classification and error detection (Table 2.7), and the need for robust, speaker-independent models that can operate in variable clinical conditions. In summary, while Cleland's work provides essential clinical context that motivated several design decisions in this thesis, the emphasis on interpretability (Chapter 4), the development of FoV harmonisation to enable cross-session comparisons (Chapter 5), and the deployment of a clinician-facing interface (Chapter 7). Together, the algorithmic advances and clinical insights point toward a unified workflow where automated UTI analysis supports, rather than replaces, expert clinical judgment.

2.3 Challenges, Research Gaps and Link with Challenges

The collective body of evidence, when examined considering the three challenges identified in Chapter 1 (Section 1.2), demonstrates substantial technical progress in applying DL to UTI. However, the pathway toward clinically deployable systems remains fragmented. Despite high within-dataset performance, existing studies reveal persistent obstacles spanning data availability, model generalisability, and clinical integration. These limitations define the research space that this thesis seeks to address.

C1: Data Variability and Generalisability Limitations

A dominant challenge in DL-based UTI analysis is the high degree of variability introduced during data acquisition. Differences in probe placement, head posture, coupling pressure, and ultrasound hardware result in substantial variation in the FoV, scale, and anatomical coverage across speakers and recording sessions. Consequently, identical articulatory gestures may appear at different spatial locations or resolutions, making it difficult for models trained on one dataset or speaker group to generalise to unseen conditions.

Most prior studies attempt to mitigate this issue through basic preprocessing steps such as resizing or intensity normalisation. However, these approaches standardise image dimensions rather than anatomical content, leaving residual heterogeneity that encourages models to learn speaker- or session-specific background patterns instead of linguistically meaningful tongue motion. This limitation is consistently reflected in reduced performance under speaker-independent and cross-dataset evaluation. Image artefacts and noise sensitivity further exacerbate this challenge. UTI is inherently affected by speckle noise, acoustic shadowing from the mandible or hyoid bone. While classical filtering methods can suppress noise, they often degrade fine articulatory detail. Although DL architectures exhibit improved robustness, explainability analyses have shown that models may still attend to high-contrast artefacts rather than the tongue surface itself. These artefact-driven attention patterns introduce additional domain shift, directly undermining model generalisability.

Research Gap 1: There is a need for systematic image harmonisation and domain-robust modelling strategies that explicitly address FoV variability, acquisition artefacts, and anatomical misalignment, enabling reliable generalisation across speakers, sessions, and recording conditions.

C2: Data Scarcity and Annotation Efficiency

A second major challenge is the limited availability of large, well-annotated UTI datasets, particularly for disordered child speech. Although the UltraSuite corpus has been instrumental in advancing the field, the number of participants with SSDs remains small relative to TD speakers. This imbalance biases learning toward typical articulations and reduces sensitivity to clinically relevant error patterns. Dataset expansion is further constrained by ethical considerations surrounding paediatric data collection and the high cost of frame-level expert annotation. Many DL models, therefore, rely on small, highly curated datasets, increasing the

risk of overfitting and limiting reproducibility. While recent work on SSL demonstrates promise, these approaches are rarely combined with principled strategies for selecting the most informative samples for annotation.

Research Gap 2: There is a need for data-efficient learning frameworks, such as AL and cost-aware sampling, that minimise annotation effort while preserving diagnostic accuracy and robustness.

C3: Lack of Interpretability and Clinical Usability

Despite encouraging classification and segmentation performance, most DL-based UTI systems remain opaque, producing categorical predictions without insight into the articulatory features driving those decisions. This lack of interpretability represents a significant barrier to clinical adoption. For SLTs to trust and use automated systems, models must provide transparent, clinically meaningful explanations, such as visualisations that highlight atypical tongue regions or articulatory deviations. Moreover, most published studies evaluate models offline, with limited consideration of real-time performance, user interaction, or integration into clinical workflows. Without interpretable outputs and clinician-oriented interfaces, even highly accurate models are unlikely to transition beyond the laboratory setting.

Research Gap 3: There is a need for explainable and deployable DL frameworks that provide transparent model reasoning, real-time feedback, and interfaces aligned with clinical practice.

The three challenges and associated research gaps identified above form the conceptual foundation of this thesis. Each gap represents an unmet requirement for progress toward clinically viable DL-based UTI systems. Table 2.9 summarises how these limitations are addressed across subsequent chapters, linking observed constraints in prior work to the methodological strategies and innovations introduced in this research.

Table 2. 9: Summary of Research Gaps and Corresponding Thesis Focus.

Challenge	Observed Limitation in Literature	Corresponding Thesis Focus
Data variability and generalisability limitations (C1)	UTI data shows large differences in FoV, probe alignment, and anatomical coverage across speakers and sessions. Models trained on one dataset often fail to generalise due to domain shifts and inconsistent imaging geometry	Development of robust and harmonised models through: (a) baseline and multi-input FusionNet architectures for improved cross-speaker generalisation; and (b) a two-stage generative pipeline for FoV and ROI standardisation (Chapters 3 (objectives 1,2,3)) Chapter 5 (objective 1)
Data scarcity and annotation efficiency (C2)	Annotated UTI datasets are limited and costly to produce; most studies rely on small samples	Formulation of a cost-aware framework combining statistical power-curve modelling with AL to minimise labelling effort while preserving accuracy (Chapter 6 (objectives 1,2))
Lack of interpretability and clinical usability(C3)	Most DL models function as opaque systems, providing no visual explanation or clinical interface. This limits trust, transparency, and real-world adoption	Deployment of a containerised, real-time UTI processing system enabling automated FoV harmonisation and interpretable visual feedback (Chapter 7 (objectives 5)

These gaps collectively motivate the thesis objectives presented in Chapter 1. The overarching goal is to optimise automated diagnosis of SDDs through harmonised, data-efficient, and explainable ultrasound analysis, thereby advancing the field toward practical, AI-assisted articulatory assessment.

2.4 Summary

This chapter has traced the evolution of UTI from a clinical research tool to an emerging platform for automated speech assessment. It reviewed the principles and constraints of ultrasound as an articulatory modality, examined how DL has been applied across phonetic classification, segmentation, and synthesised the barriers that currently limit clinical translation. Three central challenges, **C1** data variability and generalisability limitations, **C2** data scarcity and annotation efficiency, and **C3** lack of interpretability and clinical usability, were identified as the key obstacles to scalability and deployment. Addressing these challenges forms the foundation of this thesis. The next chapter builds directly on these insights by establishing reproducible baseline models for phonetic classification of child speech using raw ultrasound data, providing a benchmark against which later harmonisation and optimisation methods are evaluated.

Chapter 3

3. Establishing Baseline Models for Phonetic Classification from Raw Ultrasound Imaging

This chapter represents the first experimental phase of the thesis and directly addresses the Challenge **C1** data variability and generalisation limitations identified in Chapter 1 (corresponding to Objective 1), which require establishing reproducible baselines and quantifying speaker-condition generalisation. The goal here is to evaluate how well DL models can classify phonetic segments from raw UTI of children, and to measure how speaker variability affects model accuracy. The central hypothesis is that CNNs can learn articulatory patterns from raw UTI, although their performance tends to drop when tested on unseen speakers due to anatomical and acquisition differences [120]. A secondary hypothesis is that combining complementary visual representations, ones that capture both the overall tongue shape and the finer surface details, will help reduce this variability and improve robustness. To test these hypotheses, a set of reproducible baseline experiments was designed. Two conventional models, a CNN and a DNN, were first evaluated, followed by two transfer-learning architectures (ResNet-50 and Inception-V3) to benchmark performance on paediatric UTI data. Building on the observed limitations of these baselines, a new dual-stream model, FusionNet, was introduced. FusionNet integrates raw UTI frames with LBP-derived texture descriptors through late fusion at a shared feature layer before classification. By establishing these baselines and introducing a new model that partially mitigates variability, this chapter lays the diagnostic groundwork and the first methodological step toward harmonised and generalisable UTI analysis. It also prepares the way for the next stages of this thesis, Chapters 4 and 5, which further tackle data heterogeneity through optimised image representations and generative FoV standardisation.

A version of this work has been published as: Al Ani, S., Cleland, J., and Zoha, A. (2024). "Automated Classification of Phonetic Segments in Child Speech Using Raw Ultrasound Imaging." Proceedings of the 17th International Joint Conference on Biomedical Engineering (IJCBME 2024), Electrical Engineering, Springer.

3.1 Introduction

As outlined in Chapter 2, SSDs are a prevalent childhood condition with significant implications for communication, education, and social development [121]. UTI has emerged as a promising tool for capturing real-time articulatory motion in paediatric speech, and it is increasingly employed to support the diagnosis and treatment of SSDs [122], [123]. Compared with other imaging modalities, ultrasound is particularly attractive in clinical linguistics and phonetics due to its non-invasive nature and relatively low cost of data acquisition. However, despite significant advancements in UTI research [16], [94], reliable interpretation remains a technically challenging task, hindered by issues such as high-level speckle noise, low contrast, and high variability, all of which obscure anatomical details and complicate automatic analysis [33].

In recent years, DL has become the leading method for analysing UTI, driving progress in tasks such as phonetic segment classification [16] and tongue-contour extraction [94]. DL models excel at automatically learning complex feature representations. However, their success relies on access to large, annotated datasets, a major limitation in speech research, where data collection and expert labelling are both time-consuming and costly. Furthermore, model accuracy often drops when evaluated on previously unseen speakers, revealing sensitivity to anatomical differences and inconsistencies in image acquisition. These issues align directly with Challenge **C1**, outlined in Chapter 1: the need to design models that can generalise robustly across diverse speakers and recording conditions.

A promising approach to enhance generalisability is transfer learning, where models pre-trained on large natural-image datasets are fine-tuned for new domains with limited data [100]. This strategy has shown strong benefits in medical imaging, helping models converge faster and remain more stable when only small datasets are available [25]. Yet, ultrasound imagery poses a unique challenge: its grainy textures, acoustic shadows, and speckle noise differ greatly from the structured patterns of natural images [124].

As a result, it is unclear how well features learned from everyday visual data transfer to UTI. To explore this, the chapter assesses two widely adopted architectures, ResNet-50 [125] and Inception-V3 [126], to determine whether transferable features can genuinely improve classification accuracy or if effective modelling of UTI still requires domain-specific feature learning.

Earlier studies have taken several different paths to overcome the twin challenges of limited data and high variability in UTI. Initial efforts relied on various preprocessing methods, including intensity normalisation and dimensionality reduction [127], [128]. While these techniques helped reduce data dimensionality, they often did so at the cost of losing fine articulatory detail essential for accurate speech analysis. Subsequent CNN-based approaches showed that tongue gestures could be automatically classified from ultrasound frames [129], marking a shift toward data-driven feature learning. However, most of these studies were trained on typical speakers and experienced sharp declines in accuracy when tested on unseen individuals, highlighting persistent generalisation issues [18], [16]. More recently, SSL methods have been introduced to exploit large pools of unlabelled data [130], [28], achieving notable gains in efficiency. Yet, despite their promise, these methods typically operate on a single imaging modality and continue to struggle with cross-speaker and cross-session variability [131]. Validation on clinical populations with speech disorders remains an open challenge. Additionally, these models are computationally demanding and require substantially larger training datasets. These diverse approaches, comprehensively reviewed in Chapter 2, Section 2.2, demonstrate both the progress and persistent challenges in DL-based UTI analysis.

To overcome these challenges, this chapter focuses on the automated classification of phonetic segments in TD children's speech using raw UTI, establishing a solid experimental baseline for the studies that follow. This directly fulfils Objective 1 from Chapter 1 (Section 1.3.1), which aims to establish reproducible architectures for speaker-dependent, multi-speaker, and speaker-independent phonetic classification. It benchmarks widely used DL architectures for UTI classification and introduces FusionNet, a novel multi-input model developed to improve robustness across different speakers. This phonetic classification offers an objective, data-driven measure of speech production and provides a reproducible framework for analysing how architectural design choices affect both model generalisability and interpretability.

The classification design adopted in this chapter warrants explicit clarification in relation to the broader clinical aim of the thesis. The experiments here frame the task as a four-class phonetic classification based on place of articulation using data from TD children only. This design was chosen as a methodologically controlled setting in which to establish baseline DL performance: by working with a single well-characterised population producing known phonetic contrasts, the chapter evaluates how different CNN architectures handle speaker variability in UTI-based phonetic classification, without the additional confounds introduced by clinical group differences or disordered speech patterns. It should be acknowledged, however, that phonetic segment classification in TD children does not represent the full clinical diagnostic task. Phonetic segment classification refers to the automated identification of which speech sound, defined here by place of articulation, a child is producing at a given moment in an ultrasound recording.

In a clinical context, this capability serves as a prerequisite for detecting misarticulations: a system must first learn to recognise target phonetic categories before it can determine whether a child's realisation of those categories deviates from typical production. However, demonstrating that a model can correctly classify phonetic segments in TD children confirms only that articulatory distinctions are learnable from raw UTI; it does not directly address whether the same framework can distinguish typical from disordered articulation, which is the core clinical diagnostic task. This limitation is addressed progressively across the thesis: Chapter 4 extends the framework to a clinically representative binary classification task contrasting TD children with children with CP±L, a population that presents with systematic, structurally motivated articulatory differences that make them a principled test case for ultrasound-based diagnostic modelling. The binary TD vs. CP±L framing used from Chapter 4 onwards was chosen specifically because it provides a controlled diagnostic contrast with clear clinical relevance, while remaining tractable as an initial evaluation of the framework. It does not reflect the full heterogeneity of SSDs encountered in clinical practice, and this broader applicability is identified as a direction for future work in Chapter 8.

3.1.1 Contributions

This chapter makes three key contributions to the thesis, each directly addressing Challenge **C1** (data variability and generalisability limitations) identified in Chapter 1.

1. Baseline benchmarking: It provides a comprehensive evaluation of standard DL architectures (CNN, DNN, ResNet-50, and Inception-V3) for phonetic classification using raw UTI. Their performance is systematically compared under different speaker conditions to establish reliable reference points for future work.

   Impact on C1: This contribution quantifies how speaker variability affects model performance and establishes reproducible baselines against which improvements in generalisation can be measured.
2. Model innovation: It introduces FusionNet, a new dual-stream architecture that combines raw UTI frames with complementary texture features extracted using LBP [132]. These fusion captures both global articulatory shape and fine-grained surface details, enhancing representational richness.

   Impact on C1: By integrating complementary feature representations, FusionNet partially mitigates speaker-related variability, demonstrating how architectural design can improve robustness without additional data.
3. Generalisation analysis: It examines model robustness across different speaker scenarios, revealing how anatomical and acquisition variability influence classification accuracy and generalisability.

   Impact on C1: This analysis explicitly exposes the limits of raw-UTI modelling under realistic cross-speaker conditions, motivating the need for representation-level and acquisition-level harmonisation strategies explored in later chapters.

The remainder of this chapter is structured as follows. Section 3.2 describes the dataset, preprocessing pipeline, experimental design, and the modelling strategy, covering the baseline architectures, transfer-learning models, and the proposed FusionNet. Section 3.3 presents quantitative results across all evaluation scenarios. Section 3.4 discusses the implications of these findings for data variability and model design, and Section 3.5 summarises the chapter and connects the outcomes to the representation-focused analysis developed in Chapter 4.

3.2 Methods

This section outlines the dataset, preprocessing steps, and experimental setup used to evaluate both the baseline and proposed models for phonetic classification from UTI. The experiments were carefully designed to ensure full reproducibility, allowing for a clear assessment of how different architectures handle cross-speaker generalisation.

3.2.1 Dataset

The experiments were conducted using ultrasound data of TD children's speech from the UltraSuite-UXTD corpus [112] (described comprehensively in Chapter 2, Section 2.2.1, Table 2.5), which provides synchronised ultrasound and audio recordings of TD children producing a range of speech stimuli. All procedures adhered to institutional ethical standards, with approval granted by the University of Strathclyde Department of Psychological Sciences and Health Ethics Committee. Written informed consent was obtained from the parents or guardians of all participants. The UXTD dataset includes recordings from children, each assessed in a single laboratory session led by a qualified SLT using Articulate Assistant Advanced (AAA) software [112]. Both clinicians and children were native speakers of Scottish English. Each recording session comprised various elicitation tasks, such as sentence reading, isolated phoneme production, and picture description, designed to capture a broad range of articulatory gestures. For each utterance, a synchronised acoustic waveform and mid-sagittal ultrasound images were recorded using a micro-convex transducer placed and stabilised with a headset, enabling clear imaging of the tongue surface during speech.

For this study, two categories of speech stimuli were used: Type A, consisting of semantically unrelated real words, and Type B, comprising non-words. Both stimulus types were included to increase the variety of phonetic contexts sampled and to represent the range of elicitation conditions used in clinical UTI assessment. While real words and non-words may elicit slightly different degrees of articulatory precision due to differences in motor planning familiarity, both types require production of the same target phonemes, and the classification task operates at the level of place of articulation rather than lexical identity.

Nine speakers were selected as the maximum feasible subset given the time constraints of manual frame-level annotation, which required expert review of each frame to assign phonetic category labels. This sample size is comparable to or larger than those reported in similar paediatric UTI classification studies. While a larger sample would strengthen generalisation claims, the speaker-independent evaluation with a held-out test speaker provides a meaningful estimate of cross-speaker robustness within the available data. The impact of limited speaker numbers on generalisation is acknowledged as a constraint of the current study and motivates the broader dataset used in Chapter 4. The selected participants represented a diverse mix of ages, genders, task types (A and B), and recording quality, ensuring that the subset remained representative of the larger UXTD cohort while minimising potential bias. Table 3.1 provides an overview of the participants' demographic characteristics.

Table 3.1: Demographic Summary of the Selected Subset.

Measure	Value
Age range	6–11 years
Mean age (SD)	9.1 (1.5) years
Gender	Female: 5; Male: 4

The raw scan-line videos corresponding to the target utterances were exported from AAA and rasterised into 600 × 480-pixel images. Four phonetic categories were defined for classification, each representing a distinct articulatory region:

1. Bilabial sounds are produced with both lips (e.g. /p/, /b/) and labiodental sounds are produced with the lower lip against the upper teeth (e.g. /v/).
2. Dental sounds are produced with the tongue against the upper teeth (e.g. /θ/), alveolar sounds at the ridge behind the top teeth (e.g. /d/, /t/, /z/), and postalveolar sounds just behind the alveolar ridge (e.g. /ʃ/).
3. Velar phones are produced with the back of the tongue against the soft palate (e.g. /g/, /k/).
4. Alveolar approximant /r/ is produced with the tongue close to, but not touching, the alveolar ridge.

3.2.2 Local Binary Patterns

Prior work has demonstrated that explicit texture modelling can be highly informative for ultrasound image classification, where speckle patterns, tissue boundaries, and edge contrast encode articulatory structure beyond raw intensity values [133]. To exploit these properties, an LBP descriptor was computed for each UTI, providing a complementary texture-based representation alongside the raw UTI frames.

LBP is a local texture operator that characterises micro-patterns by analysing intensity relationships within small spatial neighbourhoods. For each pixel, the algorithm compares the grayscale value of the centre pixel g_c with those of P surrounding pixels g_p evenly distributed on a circle of radius R. The LBP code is defined as:

$$\mathrm{LBP}_{P,R} = \sum_{p=0}^{P-1} s(\, g_p - g_c)\, 2^p, s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases} \qquad (3.1)$$

This operation yields a compact binary code for each pixel that encodes local contrast patterns such as flat regions, edges, and corners. In this work, the standard LBP configuration with P = 8 neighbouring pixels and a radius of R = 1 was adopted [134], which has been widely used in texture-based medical image analysis due to its balance between discriminative power and computational efficiency.

When applied densely across the image, these codes form a texture map that highlights fine-grained structures characteristic of UTI, including tongue-surface contours, tissue transitions, and speckle granularity. A key advantage of LBP is that it captures relative intensity differences rather than absolute pixel values, making it inherently invariant to monotonic illumination changes. This property is particularly valuable in UTI, where variations in probe pressure, acoustic coupling, and gain settings can introduce substantial inter-session and inter-speaker variability.

As a result, LBP features provide a stable representation of articulatory texture that complements the global shape and spatial information learned from raw images, directly supporting robustness to acquisition variability and addressing Challenge **C1** (data variability and generalisability limitations) introduced in Chapter 1. Figure 3.1 illustrates an example mid-sagittal UTI frame alongside its corresponding LBP transformation, demonstrating how local texture patterns are emphasised relative to the raw intensity image.

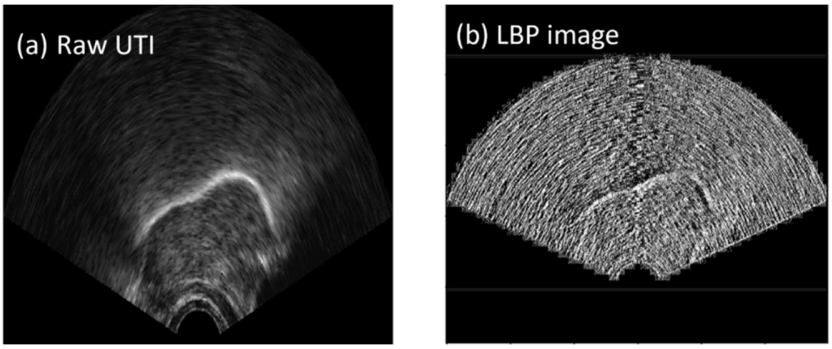

Figure 3. 1**:** Raw UTI Frame and its LBP Texture Representation (a) Example Mid-Sagittal UTI from the UXTD Corpus. (b) LBP Code Image Computed from the Same Frame.

The resulting LBP representations were flattened and standardised using a StandardScaler fitted exclusively on the training split to prevent data leakage. To obtain a compact yet informative texture embedding suitable for late fusion, PCA was applied to reduce dimensionality while preserving dominant variance in the texture space. These compressed LBP feature vectors were used as input to the texture-processing branch of FusionNet, where they were combined with CNN-derived image features at a later stage in the network. Texture features were stored alongside their paired raw UTI frames, ensuring a strict one-to-one correspondence and maintaining identical speaker-locked splits across training, validation, and testing.

3.2.3 Models and Experimental Setup

Multiple DL frameworks were implemented to classify phonetic segments from UTI, covering both single-input baselines and the proposed multi-input FusionNet architecture. The experimental design followed a structured three-stage progression:

1. Baseline benchmarking: task-specific models, a CNN and a DNN were trained on raw UTI data to establish foundational performance levels.
2. Transfer learning: pre-trained architectures (ResNet-50 and Inception-V3) were fine-tuned on the same dataset to evaluate the contribution of transferable visual representations learned from large natural-image corpora.
3. FusionNet evaluation: a dual-stream model combining raw UTI frames with LBP-derived texture features was tested to determine whether multimodal fusion enhances robustness and generalisation.

This staged progression provides a clear framework for comparing domain-specific learning with general-purpose pre-trained feature extraction, enabling a direct investigation into how transferable representations influence performance on ultrasound data. Table 3.2 provides a consolidated overview of all five model architectures evaluated in this chapter, including input size, optimiser, learning rate, batch size, epochs, loss function, dropout or regularisation, and relevant architecture-specific settings.

Table 3. 2: Summary of Hyperparameters for Baseline and FusionNet Models.

Model	Input	Optimiser	LR	Batch size	Epochs	Loss function	Regularisation	Justification
CNN baseline	Raw UTI	SGD	0.001	32	50	Categorical cross-entropy	Dropout = 0.2; early stopping patience = 10	Used as a reproducible spatial baseline. CNN filters are appropriate for UTI because they capture local tongue contours, acoustic shadows, and speckle-related patterns.
DNN baseline	Flattened raw UTI, 224 × 224	SGD	0.001	32	50	Categorical cross-entropy	Dropout = 0.2; early stopping patience = 10	Used as a simple non-convolutional baseline to test whether flattened intensity patterns alone are sufficient for phonetic classification.
ResNet-50	Raw UTI, 224 × 224	SGD	0.001	32	50	Categorical cross-entropy	Early stopping patience = 10	Selected to test whether deep residual ImageNet features can support UTI classification under limited-data conditions.
Inception-V3	Raw UTI, [insert input size, e.g. 299 × 299 if used]	SGD	0.001	32	50	Categorical cross-entropy	Early stopping patience = 10	Selected because its multi-scale filters are relevant to UTI, where articulatory structures and artefacts appear at different spatial scales.
FusionNet	Raw UTI, 224 × 224 + 22-dimensional LBP-PCA vector	SGD	0.001	32	50	Categorical cross-entropy	Dropout = 0.2; early stopping patience = 10	Designed to combine global tongue geometry from raw UTI with local texture cues from LBP, improving robustness under speaker-independent variability.

Baseline Architectures

The baseline CNN and DNN architectures were adapted from Ribeiro et al. [16] (reviewed in Chapter 2, Section 2.3.2), who investigated DL-based phonetic classification from raw UTI. The CNN and DNN baselines were selected to establish reproducible reference points for phonetic classification from raw UTI. The DNN provides a simple, fully connected baseline that tests whether flattened pixel-level intensity patterns contain sufficient discriminative information for phonetic classification. In contrast, the CNN introduces spatially local convolutional filters, making it more appropriate for UTI because ultrasound frames contain local tongue-surface contours, edges, acoustic shadows, and speckle patterns. Comparing the DNN and CNN, therefore, allows the study to isolate the value of spatial feature learning over direct pixel-based classification. These models were also adapted from previous UTI phonetic-classification work, making them suitable for benchmarking and ensuring that subsequent improvements from transfer learning and FusionNet could be interpreted against a reproducible baseline.

Together, these baselines establish a controlled reference for evaluating transfer learning and multimodal fusion, allowing the effect of architectural complexity and feature-transfer mechanisms to be analysed systematically.

Transfer Learning Architectures

Two pretrained models, ResNet-50 and Inception-V3 [136], were employed to evaluate the suitability of transfer learning for small UTI datasets. ResNet-50 and Inception-V3 were selected to evaluate whether transfer learning from large-scale natural-image datasets could mitigate the limited-data problem in UTI. ResNet-50 was chosen because its residual connections allow deeper hierarchical feature learning while reducing vanishing-gradient effects, which is useful for capturing tongue-shape patterns across speakers. Inception-V3 was selected because its multi-scale convolutional modules can process visual structures at different spatial scales, which is relevant to UTI, where articulatory features vary in size and position across children. These architectures therefore provided two complementary tests of transfer learning: depth-based residual feature reuse and multi-scale feature extraction. Their inclusion was appropriate because UTI datasets are relatively small, and transfer learning is a common

strategy for improving convergence and reducing overfitting in limited-data medical imaging. [137], [138].

A partial fine-tuning strategy was adopted for both architectures: early convolutional layers were frozen to preserve low-level visual features learned from ImageNet, such as edge and texture detectors, which retain utility across image domains, while deeper layers were retrained on the ultrasound data to enable adaptation to domain-specific articulatory features. This approach reduces data requirements and computational cost relative to full fine-tuning, while still permitting meaningful domain adaptation in the higher-level feature representations most relevant to phonetic classification.

Proposed Multi-Input Model (FusionNet)

While raw UTI capture the global geometry and overall motion of the tongue, the accompanying LBP features encode local intensity transitions such as edges, fine speckle textures, and contour variations. These two representations offer complementary perspectives: the raw UTI preserves articulatory shape, whereas LBP provides robust micro-textural cues that remain stable across speakers and recording conditions. This distinction is particularly important in UTI, where substantial variability arises from differences in tongue anatomy, probe positioning, and image acquisition settings. Models trained solely on raw images may overemphasise speaker-specific geometric characteristics, whereas texture descriptors provide additional information that is less dependent on absolute tongue shape. Consequently, combining shape-based and texture-based representations offers a principled mechanism for improving robustness to inter-speaker variability and enhancing generalisation. Unlike previous UTI studies, which predominantly rely on single-stream CNNs operating directly on raw ultrasound images or on handcrafted features used in isolation, the proposed FusionNet explicitly integrates both representations within a unified learning framework. This enables the model to exploit complementary information sources simultaneously, capturing both global articulatory structure and local ultrasound texture patterns. Such feature-level fusion has the potential to produce richer and more discriminative representations than either modality alone, particularly in speaker-independent classification scenarios. Figure 3.2 illustrates the proposed FusionNet architecture, which integrates these modalities through a dual-stream design:

1. Image Stream: a three-layer convolutional neural network with ReLU activations and max-pooling layers, followed by flattening to produce a compact spatial embedding.
2. Texture Stream: a multi-layer perceptron (MLP) with two hidden layers that processes the LBP texture vectors extracted as described in Section 3.2.2.

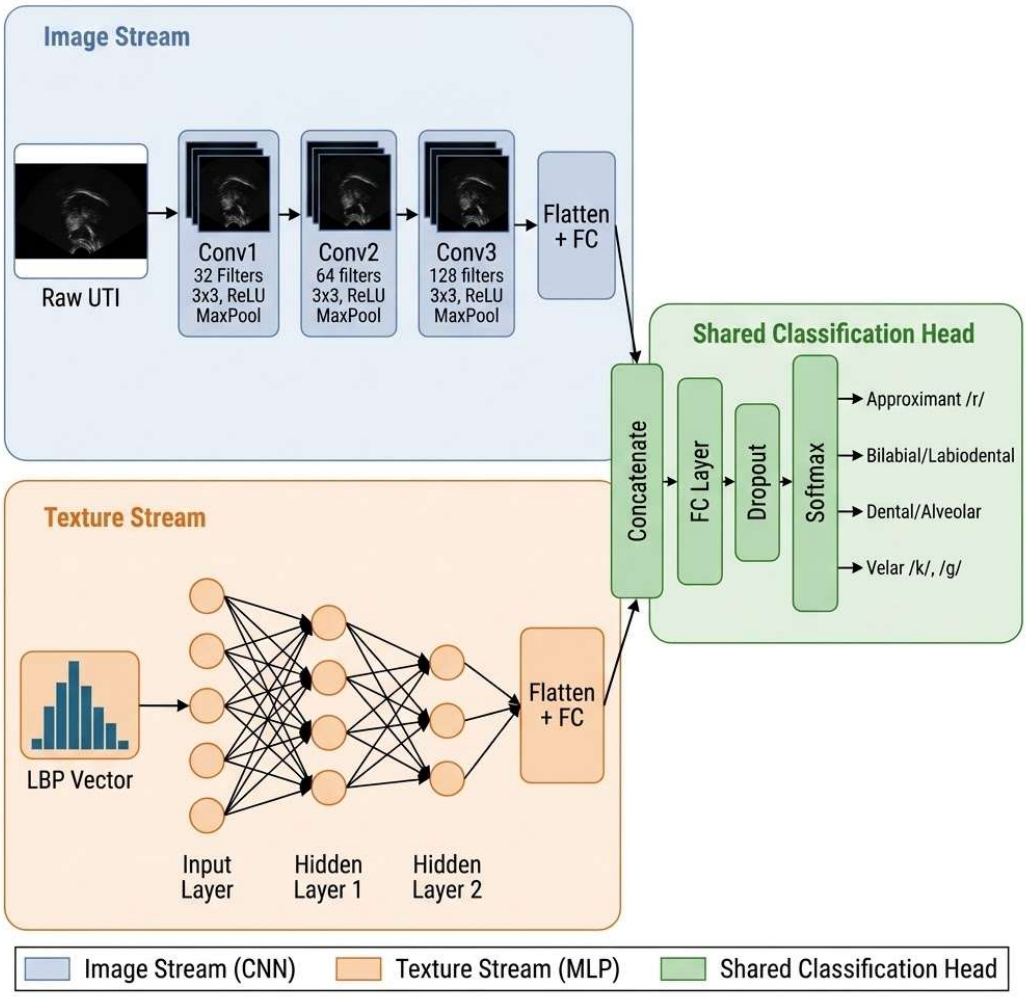

Figure 3. 2: The Proposed FusionNet Architecture.

The flattened feature embeddings from both streams are concatenated and passed to a shared classification head comprising a fully connected layer, dropout ($p = 0.2$), and a 4-unit softmax output layer corresponding to the phonetic classes. This late-fusion strategy allows each input modality to be encoded independently before joint decision-making, enabling the model to exploit both the global articulatory structure and the fine-grained texture patterns characteristic of ultrasound data.

3.2.4 Training Configuration

To ensure balanced class representation across speakers, each of the nine participants contributed 190 frames per phonetic class. These frames were determined by the minimum number of available annotated frames across all speakers and phonetic categories, ensuring a fully balanced dataset. Balancing at this level was prioritised over maximising total frame count because class imbalance in the training data would have introduced confounds into the accuracy and F1-score metrics, making cross-condition comparisons less interpretable. This constraint also meant that speakers with more available frames contributed only a subset of their data, which is acknowledged as a limitation in terms of total training data volume.

All ultrasound frames were resized to 224×224 pixels to ensure compatibility with ImageNet-pretrained architectures; the same input size was applied to all models to ensure consistent experimental conditions. All UTI frames were normalised, and various augmentation techniques were applied to artificially expand the dataset, including rotation, scaling, and edge enhancement, ensuring models were robust and capable of generalising to unseen data. All models were trained for 50 epochs, divided into 70% training, 15% validation, and 15% testing. All experiments were initialised with fixed random seeds and trained under identical conditions to ensure comparability and reproducibility.

3.2.5 Evaluation Strategy

The data partitioning strategy varied across the three evaluation conditions to reflect their distinct purposes. In the speaker-dependent condition, all frames from a single speaker were divided into training, validation, and test subsets within that speaker, allowing the model to learn and be evaluated on the same individual's articulatory patterns. In the multi-speaker condition, frames from all speakers were pooled and split at the frame level, such that each frame was assigned to exactly one partition; speakers could contribute frames to both training and testing in this condition. In the speaker-independent condition, a strict speaker-locked split was applied: all frames from a given speaker were assigned exclusively to either the training or the testing set, ensuring that no speaker seen during training appeared at test time.

This last condition represents the most rigorous test of generalisation and is the primary benchmark for clinical applicability. To quantify uncertainty around point estimates, 95% confidence intervals (CI) were computed for key accuracy figures using the Wilson score interval, which provides reliable coverage for binomial proportions across a range of sample sizes [140]. It should be noted that test frames were not fully independent, as multiple frames were drawn from the same speakers, introducing within-speaker correlation. The reported Wilson score intervals are therefore likely narrower than the true uncertainty; that is, they underestimate the variability and should be interpreted as approximate lower bounds on uncertainty rather than exact intervals. The non-overlapping margins observed between FusionNet and all baseline models are sufficiently large that this limitation does not alter the substantive conclusions. Model performance was evaluated using accuracy and macro F1-score as the primary metrics. All experiments were conducted on the University of Glasgow High Performance Computing (HPC) node.

3.3 Results

This section presents the experimental findings for the baseline, transfer-learning, and multi-input architectures. The results quantify the impact of speaker variability on classification accuracy and evaluate whether architectural design, particularly feature-level fusion, improves robustness across speaker conditions.

3.3.1 Baseline Performance

For a balanced four-class problem with equal class representation, the expected chance-level accuracy is 25%. All models evaluated substantially exceed this baseline, confirming that classification is driven by learned articulatory representations rather than class frequency bias. Table 3.3 summarises the accuracy and F1-scores of the baseline CNN and DNN classifiers across all three speaker conditions. The comparison demonstrates that both architectures in the present study achieved consistent or slightly higher accuracies across all speaker scenarios. In the speaker-dependent setup, accuracy increased by 9% for the CNN and by 8% for the DNN, indicating improved performance over Ribeiro et al.'s reported figures. For the multi-speaker condition, accuracies of 72.42% for the CNN and 69.93% for the DNN closely match those reported in Ribeiro et al. [16] confirming that the models generalised comparably across a limited speaker pool. As expected, the speaker-independent configuration yielded the lowest scores of 54.74% for CNN and 52.16% for DNN, reflecting the well-known challenge of generalising to unseen speakers due to anatomical and articulatory variability.

Overall, CNNs consistently outperformed DNNs across all scenarios, aligning with prior findings that convolutional layers are better suited to capture spatially localised patterns, such as tongue surface contours and speckle distributions. These results validate the reproducibility of the baseline architectures while establishing a reference performance level against which later experiments, transfer-learning and multi-input fusion are compared.

Table 3.3: Baseline Models Accuracy Performance.

Model	Speaker Condition	Accuracy (%)	F1 (%)	95% CI
CNN	Speaker – dependent	74.30	73.90	71.2%,77.4%
	Multi-speaker	72.42	72.10	69.2%,75.6%
	Speaker–independent	54.74	53.09	51.2%,58.3%
DNN	Speaker–dependent	68.70	67.80	65.4%,72.0%
	Multi-speaker	69.93	69.73	66.6%,73.2%
	Speaker–independent	52.16	51.84	48.6%,55.7%

3.3.2 Transfer Learning Models

Following the baseline evaluation, the study examined two pretrained architectures, ResNet-50 and Inception-V3, to test whether transferable natural-image features could enhance performance with limited UTI data. Table 3.4 provides a complete summary of accuracy, F1-score, and CIs for both architectures.

In the speaker-dependent condition, ResNet-50 achieved the highest accuracy of 77.35%, followed by CNN 74.30%, Inception-V3 68.96%, and DNN 68.70%. This suggests that when training and testing occur on the same speaker, deeper residual architectures can effectively exploit speaker-specific structural cues, despite the domain mismatch between ImageNet and ultrasound imagery. Under the multi-speaker setting, ResNet-50 delivered the strongest performance with 80%, outperforming CNN 72.42%, Inception-V3 70.31%, and DNN 69.93%. The residual connections may help preserve hierarchical features when moderate inter-speaker variability is introduced, as they can capture articulatory features at multiple spatial resolutions. In the more challenging speaker-independent scenario, ResNet-50 demonstrated the best generalisation, 72.12%, followed by Inception-V3 58.14%, CNN 54.74%, and DNN 52.16%. The superior performance of ResNet-50 under full domain shift suggests that residual connections may help preserve hierarchical feature representations when anatomical and acquisition variability are maximised. In contrast, the shallow CNN and fully connected DNN showed substantial performance degradation, indicating greater sensitivity to speaker-specific characteristics learned during training.

These results collectively demonstrate that while transfer-learning architectures offer marginal gains in mixed-speaker conditions, their advantage diminishes in fully speaker-independent evaluation. The findings highlight the importance of domain-specific representations rather than generic image priors for UTI classification.

Table 3.4: Transfer Learning Models Accuracy Performance.

Model	Speaker Condition	Accuracy (%)	F1 (%)	95% CI
ResNet-50	Speaker–dependent	77.35	77.10	74.4%,80.4%
	Multi-speaker	80.10	79.82	77.2%,82.8%
	Speaker–independent	72.12	72	68.9%,75.3%
Inception-V3	Speaker–dependent	68.96	68.49	65.7%,72.2%
	Multi-speaker	70.31	70.20	67.0%,73.6%
	Speaker–independent	58.14	57.87	54.6%,61.7%

3.3.3 FusionNet Performance

Building on the performance trends observed in section 3.3.2, the third experimental stage introduced the proposed FusionNet. This dual-stream architecture integrates raw ultrasound frames with complementary texture features derived from LBP. This approach was designed to address the limitations of single-input models, particularly their reduced generalisability under speaker-independent conditions, by enabling the network to learn both global articulatory structures and local intensity transitions. Figure 3.3 presents the accuracy achieved by all models across the three evaluation scenarios, and Table 3.5 shows FusionNet's complete summary of accuracy, F1-score, and CIs. The model attained 91.48 % accuracy in the speaker-dependent condition and 88.68 % in the multi-speaker condition, while maintaining a strong 81.69 % under the speaker-independent setup. These results mark a clear improvement over all previous architectures, demonstrating that combining raw and texture-based features substantially enhances model robustness to inter-speaker variability.

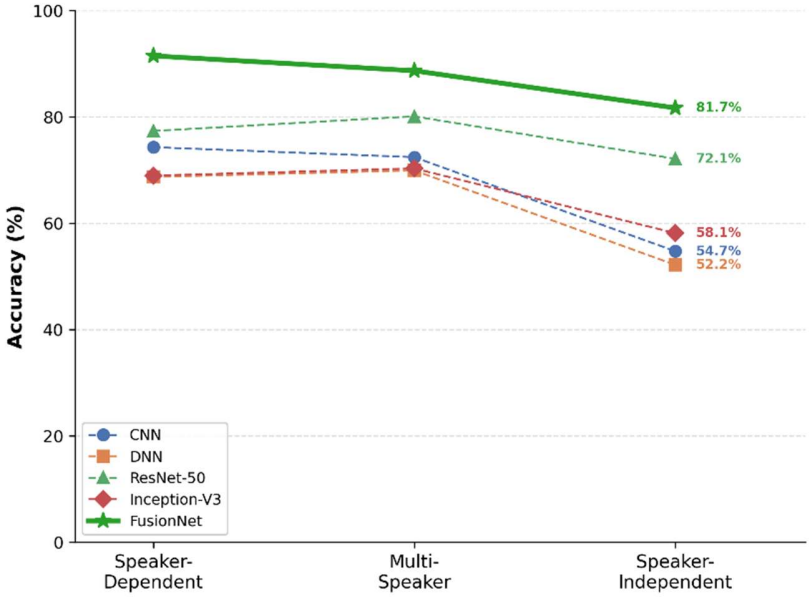

Figure 3. 3: All Models' Accuracy Performance.

To quantify the reliability of FusionNet's performance advantage, 95% CIs were computed for all models. FusionNet achieved 81.69% accuracy [79.0%, 84.4%], compared with the strongest baseline, ResNet-50, at 72.12% [68.9%, 75.3%]. The non-overlapping confidence intervals confirm that this 9.57pp improvement is statistically reliable and not attributable to test set variability. Notably, FusionNet's lower confidence bound (79.0%) exceeds the upper bound of every other model evaluated, ResNet-50 (75.3%), Inception-V3 (61.7%), CNN (58.3%), and DNN (55.7%), providing unambiguous statistical evidence that multimodal fusion delivers a genuine and consistent generalisation advantage under full cross-speaker evaluation.

Table 3. 5: FusionNet Model's Accuracy Performance.

Model	Speaker Condition	Accuracy (%)	F1 (%)	95% CI
FusionNet	Speaker–dependent	91.48	91.11	89.5%, 93.5%
	Multi-speaker	88.68	87.93	86.5%, 90.9%
	Speaker–independent	81.69	81.43	79.0%, 84.4%

To further investigate class-wise behaviour, a confusion matrix for the speaker-independent condition is shown in Figure 3.4. The model exhibits broadly balanced performance across phonetic categories, with recall scores of 0.91 for approximant, 0.78 for bilabial/labiodental, 0.70 for dental/alveolar, and 0.89 for velar sounds, yielding a macro-averaged recall of 0.82. Inspection of the confusion matrix shows that the errors are not uniformly distributed across classes. The dental/alveolar category shows the lowest recall and is most frequently misclassified as approximant (0.18). This error is likely due to anatomical and articulatory similarity between these categories, as both involve anterior tongue configurations in which subtle differences in tongue-tip elevation and constriction location may be difficult to resolve in mid-sagittal ultrasound. Bilabial/labiodental tokens show the second largest error pattern, with 0.13 misclassified as dental/alveolar. This may reflect the limited visibility of lip and labial gestures in UTI, since ultrasound primarily captures the tongue surface rather than external articulators. As a result, the model may rely on indirect tongue-position cues that overlap with anterior coronal sounds.

By contrast, velar sounds show low confusion with all other categories (≤0.04 per cell), which is consistent with the visually distinct posterior tongue-body gesture associated with velar articulation. Approximant /r/ is never confused with velar (0.00), the most articulatorily distant category, further suggesting that the model's errors are linguistically interpretable rather than random. Overall, the dominant errors are most likely explained by anatomical similarity and the limited visibility of some articulatory contrasts in single-frame mid-sagittal UTI, rather than by class imbalance, since the dataset was constructed with balanced class representation. Speaker variability and image quality may also contribute, particularly under the speaker-independent condition, where differences in tongue anatomy, probe position, and ultrasound contrast reduce the consistency of class-specific visual patterns.

These findings indicate that while FusionNet improves cross-speaker robustness, fine-grained discrimination among anterior places of articulation remains a key limitation of frame-level UTI classification.

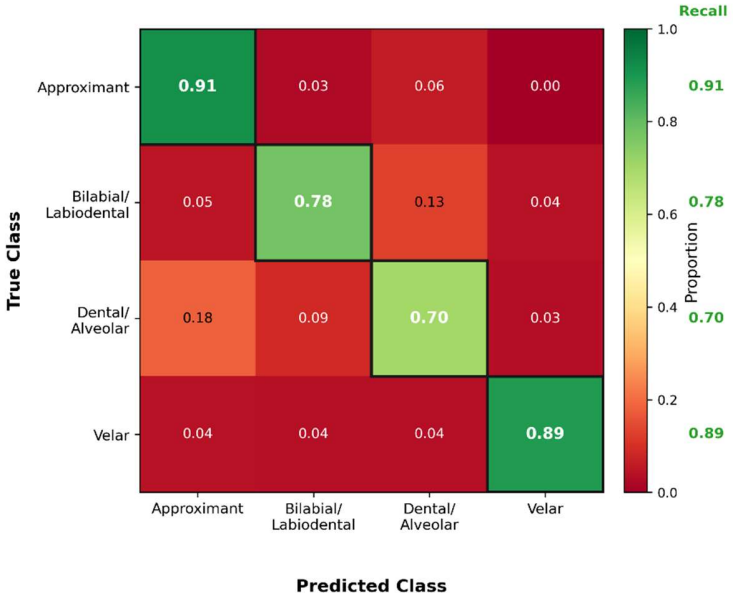

Figure 3. 4: Confusion Matrix for Speaker-Independent on the Testing Dataset.

3.3.4 Contribution of Individual Input Streams to Fusion Performance

To validate the architectural design of FusionNet, an ablation study was conducted comparing three configurations: (a) the UTI stream only, implemented using FusionNet's image branch with three convolutional layers, not the earlier two-layer baseline CNN with the same classification head but no texture input; (b) the LBP texture stream only, implemented using FusionNet's MLP branch with the same classification head but no image input; and (c) the full FusionNet combining both streams through late fusion. It should be noted that the UTI-only configuration uses FusionNet's three-layer image branch rather than the two-layer CNN baseline from Section 3.3.1. This is intentional: the ablation is designed to isolate the contribution of the LBP fusion specifically, and using FusionNet's own image branch ensures that any observed difference between UTI-only and full FusionNet reflects modality contribution rather than architectural depth. All three configurations used identical training procedures,

hyperparameters, data splits, and speaker conditions, ensuring that any observed performance differences reflect modality contribution rather than implementation artefacts.

The results in Figure 3.5 demonstrate a consistent and decisive performance hierarchy across all speaker conditions. The UTI-only configuration achieved 76.25%, 74.38%, and 61.25% accuracy under speaker-dependent, multi-speaker, and speaker-independent conditions, respectively, confirming that raw ultrasound frames alone capture discriminative global articulatory shape but remain highly sensitive to inter-speaker anatomical variability. The 15.00 percentage points (pp) drop from speaker-dependent to speaker-independent testing reflects the challenge of data variability and anatomical variation in paediatric ultrasound imaging.

The LBP-only configuration achieved lower overall accuracy across all three conditions (61.40% speaker-dependent, 60.00% multi-speaker, and 51.88% speaker-independent), showing a gradual degradation of 9.52pp from speaker-dependent to speaker-independent. This more modest decline, compared with the 15.00pp drop observed for the UTI-only configuration, reflects the relative stability of LBP features across speakers; however, the consistently lower absolute accuracy across all conditions demonstrates that texture features alone lack the global spatial context necessary for reliable phonetic classification. A direct statistical test of within-speaker versus between-speaker LBP feature variance was not conducted and represents a direction for future work.

Full FusionNet substantially outperformed both single-stream configurations across all conditions, achieving 91.48%, 88.68%, and 81.69% accuracy. Critically, the gain was largest in the speaker-independent condition (+20.44pp over UTI), precisely the scenario where generalisation is most clinically relevant. This result directly validates the complementary nature of the two streams: the image branch provides discriminative global shape representations, whilst the LBP branch contributes speaker-robust local texture statistics. When fused, these modalities produce representations that are simultaneously discriminative and robust to inter-speaker variability; neither modality achieves this on its own.

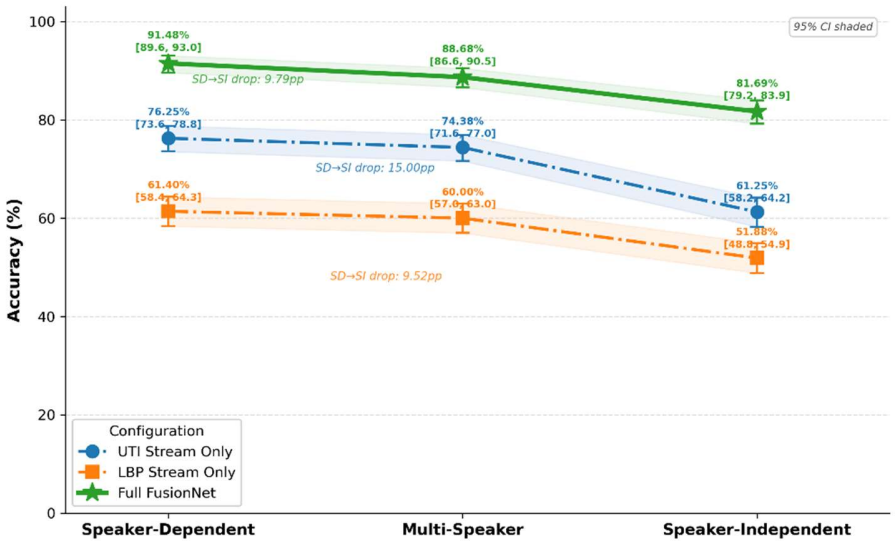

Figure 3. 5: Contribution of Individual Input Streams.

These findings confirm that the architectural design of FusionNet is justified: the performance gains observed throughout Section 3.3 are attributable to the fusion strategy itself, rather than to any single modality or incidental implementation difference.

3.4 Discussion

This chapter examined the automated classification of phonetic segments in children's speech using DL applied to UTI. The experiments followed a structured progression to explore how model architecture and speaker variability influence performance. The baseline experiments showed that CNNs consistently outperformed DNNs across all speaker conditions, reaffirming that spatial feature extraction is essential for ultrasound data. This finding aligns with the results of Ribeiro et al. [16], who first demonstrated CNNs' superiority over fully connected networks for phonetic classification from raw UTI. However, the present work achieved higher accuracies overall, improving speaker-dependent performance, likely stemming from the refined preprocessing and balanced sampling, which minimised intra-speaker variability and improved signal consistency. These methodological refinements illustrate how rigorous data curation can yield substantial gains even without architectural novelty.

Despite these improvements, both models suffered substantial performance declines under speaker-independent testing, mirroring the limitation repeatedly noted in earlier studies: DL systems tend to learn speaker-specific patterns rather than generalisable articulatory representations [16]. This outcome highlights a fundamental obstacle to clinical translation, achieving robust model performance across individuals with different anatomies, probe positions, and acquisition conditions, corresponding to Challenge **C1** data variability and generalisability limitations, as introduced in Chapter 1.

To examine whether transfer learning could mitigate the generalisation limitations observed in the baseline models, two pre-trained architectures, ResNet-50 and Inception-V3, were evaluated. Although transfer learning has demonstrated benefits in other medical imaging domains [141], its impact on UTI classification was mixed. ResNet-50 achieved the highest accuracy in both the speaker-dependent (77.35%) and multi-speaker (80%) conditions and notably demonstrated strong generalisation in the speaker-independent scenario (72.12%), substantially outperforming the task-specific CNN and DNN baselines. This suggests that deeper residual architectures can better preserve hierarchical feature representations under increasing anatomical and acquisition variability. In contrast, Inception-V3 yielded more modest improvements, indicating that multi-scale convolution alone may not sufficiently compensate for the domain gap between natural images and ultrasound data.

Despite the relative success of ResNet-50, transfer learning alone did not fully resolve the variability challenge. ImageNet-pretrained filters are optimised for photographic textures and object boundaries, which differ fundamentally from the speckle-dominated and low-contrast characteristics of ultrasound. This observation is consistent with Xu et al. [19], who reported that domain-specific pretraining offers greater benefit than direct natural-image transfer for articulatory imaging tasks. Collectively, these findings suggest that while architectural depth contributes to robustness, domain-aware representation remains critical for UTI classification.

The proposed FusionNet addressed these limitations through late fusion of raw ultrasound frames and LBP-derived texture features. Integrating LBP with raw ultrasound images in FusionNet reduces speaker variability by shifting the model's focus from pixel intensities to local spatial relationships that define tongue shape. This hybrid approach addresses anatomical variability through LBP's grey-scale invariance, noise robustness by identifying coherent local structures, and resilience to probe placement shifts and minor rotations because texture features describe sequential illumination patterns rather than fixed coordinates. FusionNet achieved

consistent and substantial improvements across all evaluation conditions, reaching 91.48% accuracy in the speaker-dependent setting, 88.68% in the multi-speaker setting, and 81.69% in the speaker-independent setting. These results surpass all single-stream architectures tested in this chapter, including transfer-learning models, demonstrating that multimodal feature fusion yields more discriminative and stable representations under speaker variability. Formal pairwise significance testing was not conducted, as Per-sample prediction outputs were not retained at inference time, precluding formal paired significance tests such as McNemar's test. This is acknowledged as a limitation of the experimental design; future work should retain frame-level predictions to enable formal statistical comparison. However, the non-overlapping 95% CIs between FusionNet and all baseline models in all speaker conditions provide robust statistical evidence that the observed performance differences reflect genuine generalisation gains rather than test set variability. This interpretation is further supported by the ablation analysis in Section 3.3.4, which confirms that neither the image stream nor the texture stream alone achieves comparable performance, validating the architectural rationale for late fusion. Furthermore, FusionNet maintained balanced recall across phonetic categories, an essential requirement for clinical deployment, where false negatives (missed atypical articulations) carry greater diagnostic risk than false positives. These findings reinforce the argument presented in Chapter 2 (Section 2.1.6.2): robust ultrasound-based classification depends not solely on transferring generic visual features, but on designing architectures that explicitly reflect the structural and textural properties of UTI.

The class-wise error analysis reveals that FusionNet's misclassifications follow articulatorily interpretable patterns, providing meaningful insight into the model's representational behaviour. Dental/alveolar sounds showed the lowest recall (0.70), with primary confusion directed toward approximant /r/ (0.18) rather than being distributed uniformly across classes. This asymmetry is consistent with articulatory research demonstrating that both categories involve tongue-tip activity near the anterior vocal tract [122], where the spatial differences in mid-sagittal ultrasound are subtle and easily obscured by inter-speaker anatomical variation.

Bilabial/labiodental sounds showed secondary confusion with dental/alveolar (0.13), again reflecting shared anterior place features. Critically, velar sounds were rarely confused with anterior categories (≤0.04 per cell in all directions), and approximant /r/ was never misclassified as velar (0.00).

This pattern confirms that the model has learned a representational space that broadly respects articulatory distance: errors occur between phonetically adjacent categories and not between maximally distinct ones. This finding underscores the importance of incorporating articulatory knowledge and explainability tools when interpreting model behaviour, themes that are further developed in Chapter 4.

Overall, the results demonstrate that multimodal fusion improves generalisability by leveraging complementary spatial and textural cues. They also highlight two key requirements for achieving clinically robust DL in UTI: (1) architectural adaptation, designing networks that explicitly model the spatial and textural properties of ultrasound; and (2) data standardisation, reducing inter-speaker and inter-session variability through consistent representations and standardised imaging characteristics. Together, these strategies move the field toward reproducible, domain-aware modelling rather than ad-hoc transfer of computer-vision architectures. Although all models in this chapter were trained on raw ultrasound data to provide a consistent evaluation baseline, the persistent gap between speaker-dependent and speaker-independent performance indicates that unstandardised imaging conditions, such as differences in probe alignment, FoV, and tongue visibility, continue to limit generalisation. Because these factors were not controlled in the present experiments, their precise influence remains unquantified. This limitation directly motivates the next stage of research, detailed in Chapter 4, which systematically investigates how image representation and FoV standardisation affect both classification accuracy and model interpretability in ultrasound-based speech disorder diagnosis.

3.5 Summary

This chapter established the methodological and empirical baseline for the thesis. By systematically benchmarking DNN, CNN, transfer-learning (ResNet-50, Inception-V3), and a fusion-based architecture (FusionNet), it quantified the attainable performance for phonetic classification from UTI and identified the main factors constraining generalisability. FusionNet, which integrates raw frames with LBP-based texture features, delivered the most robust results across speaker conditions. Its balanced recall and high accuracy demonstrate that combining complementary representations improves resilience to inter-speaker variability. At the same time, the residual performance gap in speaker-independent testing makes clear that architectural innovation alone cannot fully overcome imaging variability. The evidence indicates that generalisation depends as much on data standardisation, particularly FoV, ROI, and probe-related differences, as on network design. In other words, the variability challenge originates as much from the data as from the model. The next chapters, therefore, pivot from establishing baselines to controlling the data and understanding the model.

Chapter 4

4. Improving Interpretability in Ultrasound Tongue Imaging through Representation and Harmonisation

Building on the architectural baselines established in Chapter 3, this chapter tackles two closely linked challenges that continue to limit the clinical reliability of DL systems for UTI: data variability and model interpretability, specifically addressing Challenge **C1** data variability and generalisability limitations through FoV standardisation and Challenge **C3** lack of interpretability and clinical usability through Grad-CAM++ analysis. The previous chapter showed that even advanced designs such as FusionNet remain sensitive when tested on unseen speakers, leading to noticeable drops in accuracy, which could be due to differences in probe positioning and FoV variation. These findings suggest that the problem of generalisation stems not only from model architecture but also from inconsistencies in how UTIs are represented. To address this, the current study examines how both image representation (including raw, ROI, and masked formats) and FoV variability affect model performance and the anatomical plausibility of model attention. Using a lightweight convolutional backbone (EfficientNet-B0), models were trained on both standardised and non-standardised datasets and interpreted through Grad-CAM++. This combination enabled a dual analysis: a quantitative assessment of classification accuracy and a qualitative evaluation of where the model directs its focus within the tongue image during phonetic classification. The central aim is to understand how representation-level factors influence generalisability and explainability. Specifically, this chapter (i) quantifies the effect of different image representations on classification accuracy across speakers, (ii) evaluates how FoV standardisation contributes to performance consistency, and (iii) determines whether Grad-CAM++ attention patterns correspond to linguistically meaningful articulatory regions.

By linking numerical performance with visual interpretability, this chapter moves beyond architectural optimisation toward representation-level standardisation and XAI, advancing the development of diagnostic frameworks that are both clinically transparent and generalisable. The sections that follow describe the experimental design, results, and interpretive findings, which together lay the groundwork for the generative standardisation framework introduced in Chapter 5.

This work has been submitted to the Scientific Reports and is currently under review following the submission of a revised manuscript.

4.1 Introduction

UTI is a non-invasive technique that captures real-time images of tongue movement during speech, providing articulatory evidence that cannot be inferred from acoustic analysis alone. As detailed in Chapter 1 and comprehensively reviewed in Chapter 2 (Section 2.1.5), UTI has become an increasingly valuable tool for assessing SSDs in children by revealing covert articulatory behaviours that may be perceptually misleading. For example, Bressmann et al. [142] demonstrated that productions transcribed auditorily as glottal stops were frequently accompanied by covert tongue-tip or tongue-dorsum elevation. Such articulatory detail is clinically critical, as early and accurate identification of atypical speech patterns directly informs effective intervention planning.

UTI recordings are typically captured in either a mid-sagittal view, which shows tongue movement from tip to root, or a coronal view, which reveals lateral elevation and grooving [143]. These perspectives enable clinicians and researchers to observe subtle articulatory behaviours, such as double articulations, covert contrasts, and compensatory gestures, that are otherwise difficult to infer from the acoustic signal. Despite its clinical value, automated analysis of UTI remains difficult. UTIs are subject to imaging artefacts and substantial inter-speaker anatomical variability, factors that collectively constrain the ability of DL models to generalise across children when trained under heterogeneous acquisition conditions [144], [145]. These limitations align directly with Challenge **C1** (data variability and generalisability limitations) articulated in Chapter 1 and systematically reviewed in Chapter 2 (Section 2.3).

This challenge becomes even more pronounced in children with CP±L. Many children with CP±L develop compensatory articulations, glottal stops, pharyngeal fricatives, or posterior tongue placements that deviate from typical articulatory targets [146], [147]. As a result, their ultrasound frames often fall outside the articulatory patterns observed in TD peers. Clinically, this variability is informative; computationally, it introduces substantial heterogeneity that can destabilise model predictions. A second source of variability arises from the way ultrasound data are represented. Raw frames contain rich articulatory detail but also include background tissue, probe artefacts, and acoustic clutter, elements that may dominate the learned representation. Restricting analysis to an ROI or using a binary mask that isolates the tongue surface can reduce this noise and yield inputs more relevant to clinical interpretation [148], [149]. However, the effects of these different representations on model accuracy and interpretability have not been systematically compared in UTI research. A further, largely unaddressed, source of variation lies in the FoV, the angular extent of the ultrasound frame. FoV can vary between sessions, equipment, and operators, often spanning 90° to 150° [150]. Narrow FoVs may exclude the tongue root, obscuring compensatory gestures such as backing. Wider FoVs, by contrast, capture more anatomy but at the cost of reduced temporal resolution and greater background interference. These inconsistencies determine how much of the tongue the model can "see" and directly influence both classification and generalisation. While FoV standardisation is a known requirement in imaging modalities such as MRI and CT [151], [152], its impact on ultrasound-based speech analysis has not been quantified. In this chapter, FoV is therefore treated as a key experimental variable; models trained on mixed-FoV data are compared with those trained on standardised-FoV datasets to assess how acquisition differences propagate through model performance.

Another obstacle to clinical adoption is the lack of interpretability in DL models. The "black box" nature of these systems limits clinician trust and complicates integration into speech therapy workflows, where understanding the rationale behind a diagnostic decision is essential for treatment planning [27]. This opacity undermines clinician trust, particularly in speech and language therapy, where treatment decisions must be transparent and explainable to both therapists and caregivers. To address this, the present study incorporates Grad-CAM++, which generates heatmaps indicating the image regions that most significantly influence a model's prediction.

Grad-CAM++ allows evaluation not only of how well a model performs but also of where it focuses its attention, whether on the tongue surface or on irrelevant background artefacts. Although Grad-CAM++ has been widely used in other medical imaging domains to verify anatomically meaningful model attention [153], its application to UTI, and particularly to cleft-related articulation, remains limited. Addressing Challenge **C3** lack of interpretability and clinical usability identified in Chapter 1, and confirmed as under-explored in Chapter 2's systematic review.

The working hypothesis is that controlling visual input through refined image representations (raw, ROI, mask) and standardised FoV will improve both generalisability and interpretability. Specifically, (i) models trained on standardised FoV data will exhibit greater robustness across speakers, and (ii) models trained on anatomically focused representations (ROI or mask) will show Grad-CAM++ attention concentrated on linguistically relevant regions, whereas models trained on raw frames will display more diffuse or artefactual focus. To test these hypotheses, a lightweight convolutional architecture (EfficientNet-B0) was trained on UTI data from TD and CP±L children under multiple controlled conditions. Three image representations, raw, ROI, and binary mask, were evaluated across both mixed-FoV and standardised-FoV datasets. Performance was measured using standard classification metrics (accuracy, precision, recall, F1-score), while Grad-CAM++ visualisations provided qualitative insight into the anatomical plausibility of model attention. Additionally, quantitative analysis of Grad-CAM++ activation features was performed through unsupervised clustering and statistical comparison to assess whether model attention patterns encode diagnostically meaningful articulatory differences.

4.1.1 Contributions

This chapter makes three key contributions to the thesis, each directly addressing Challenge **C1** data variability and generalisability limitations and Challenge **C3** lack of interpretability and clinical usability identified in Chapter 1.

1. Image representation analysis: This chapter provides a systematic evaluation of how different UTI representations, raw frames, tongue ROI, and binary tongue masks affect both classification performance and model attention behaviour in UTI-based speech disorder diagnosis.

Impact on C1 and C3: This contribution shows that representation choice directly influences model robustness and generalisability (**C1**), while also revealing that accuracy alone is insufficient for clinical deployment unless model attention is anatomically plausible and interpretable (**C3**).

2. Quantitative evaluation of FoV standardisation: This chapter presents the first quantitative investigation of how FoV harmonisation influences model behaviour across different image representations.

   Impact on C1: This contribution establishes FoV standardisation as a necessary condition for cross-speaker and cross-site generalisability, showing that acquisition-level variability propagates directly into model performance. These findings motivate the need for automated FoV harmonisation strategies developed in Chapter 5.

3. Explainability-driven validation of articulatory learning: This chapter extends explainable AI beyond qualitative inspection by introducing a quantitative analysis of Grad-CAM++ activation features.

   Impact on C3: This contribution demonstrates that explainability methods can function as analytic tools rather than purely visual aids, supporting transparent, clinically interpretable decision-making and strengthening trust in DL-based UTI systems.

The remainder of this chapter is organised as follows. Section 4.2 outlines the datasets, preprocessing, FoV standardisation, model architecture, and explainability methods. Section 4.3 presents classification results alongside qualitative and quantitative Grad-CAM++ analyses. Section 4.4 discusses the implications for clinical deployment, interpretability–performance trade-offs, and acquisition standardisation. Section 4.5 concludes the chapter, summarising key contributions and limitations and motivating the generative standardisation framework introduced in Chapter 5.

4.2. Method

4.2.1. Datasets

This study drew on two complementary datasets: (i) a clinical dataset of children with CP±L obtained from the publicly available UltraSuite repository [112] and (ii) a newly collected dataset of TD children [154]. Both datasets comprised UTI frames paired with synchronised audio recordings of English consonant productions. Participants were aged between five and nine years, and all used the same elicitation protocol comprising consonant productions in a controlled non-word /aCa/ context. All ultrasound data were acquired exclusively in the mid-sagittal plane. Each ultrasound image corresponds to a single frame extracted at the point of maximal lingual gesture for the target consonant. This frame was identified manually using synchronised audio, waveform, and spectrogram displays. Labels were assigned at the speaker-group level (CP±L vs TD) to examine group-related articulatory patterns.

UTI data for the TD children were collected using a micro machine controlled by a Windows laptop running AAA v.2.20.01 [155]. Echo return data were captured at approximately 100 frames per second (fps) using a 5–8 MHz, 10 mm radius micro-convex probe, held in place with an Ultrafit lightweight stabilisation headset to minimise probe displacement [156]. The imaging depth was fixed at 70 mm across all participants. Audio was synchronised using an Audio Technica 3350 microphone mounted to the headset.

The CP±L recordings from the UltraSuite dataset were acquired using the same probe type, frequency range, and a matching depth setting, ensuring consistency in key imaging parameters across datasets. However, recordings were obtained in a clinical environment by a different operator, resulting in variation in the ultrasound FoV, which ranged from 119° to 133°. By contrast, the TD dataset was acquired with a uniform 97° FoV. Because FoV directly determines the angular extent of visible tongue anatomy and influences the geometric projection of articulatory structures, FoV variation represents the primary acquisition difference between datasets. These differences reflect realistic variability in probe calibration and acquisition practice, forming the basis for the FoV standardisation analysis described in Section 4.3.3.

All data were collected in accordance with institutional guidelines and ethical regulations. Ethical approval for the TD dataset was granted by the Department of Psychological Sciences and Health Ethics Committee at the University of Strathclyde, and parents/legal guardians provided written informed consent for data collection and sharing. The UltraSuite dataset was collected and disseminated with ethical approval and explicit permission for research use by the original authors [112].

4.2.1.1 Speaker Selection Criteria

Speakers were selected from both datasets based on strict data-completeness criteria to ensure sufficient articulatory sampling and annotation quality. To be included, each speaker required: (i) complete ultrasound recordings across all target consonant categories (alveolar, palatal, velar), (ii) synchronised audio with clear productions suitable for temporal alignment, and (iii) reliable tongue contour annotations verified by an experienced SLT. While the UltraSuite repository contains additional speakers, the selection was limited to 14 due to the substantial time required for expert annotation by a qualified SLT. Each speaker's data required approximately 2-3 hours of manual review, including frame-by-frame inspection, synchronisation verification, quality control of automated tongue tracking, and identification of maximal lingual gestures. Given the need for clinical expertise and the rigorous quality standards required for model validation, this sample size struck a balance between data quality and the feasibility of annotation effort. For the TD dataset, 14 speakers were selected to match the CP±L group. Matching criteria included: (i) age range (all TD speakers were within ±12 months of the CP±L group mean age of 7.05 years), (ii) comparable data completeness, and (iii) similar image quality characteristics. This matched design ensured that group differences reflected articulatory variation rather than confounds due to demographics or data quality. The specific UltraSuite speaker IDs are listed in Appendix A for replicability.

This design prioritised annotation quality and internal validity over sample size. While the resulting cohort is modest (n=14 per group), all included data met rigorous quality standards, with verified tongue surface tracking and expert-validated frame selection. The resulting dataset comprised 2,660 ultrasound frames in total (1,400 CP±L, 1,260 TD), with an average of approximately 95 frames per speaker representing productions across multiple consonant categories.

Table 4.1: Overview of Datasets Used in this Study.

Dataset	No. of Speakers	Mean Age	No. of Images	Speech Task	Annotation Method
CP±L	14	7.05 ± 1.3 years	1400	non-word utterances	Expert-annotated tongue contours
TD children	14	7.05 ± 1.3 years	1260	non-word utterances	Automatically annotated tongue contours

This cohort size is comparable to prior UTI studies with similar data collection constraints (e.g., Eshky et al., 2019: n = 8 speakers [157]; Xu et al., 2017: n = 3 speakers [129]). Accordingly, the results should be interpreted as preliminary findings that warrant validation on larger, independent cohorts before clinical deployment. Together, these datasets provide complementary perspectives: the TD data capture relatively consistent articulatory behaviour under uniform imaging conditions, while the CP±L data introduce clinically relevant articulatory variability and acquisition differences. This combination enables a controlled investigation of both anatomical heterogeneity and imaging variability, key factors influencing the performance and generalisability of DL models for UTI.

4.2.2. Image Representation Techniques

To examine how image representation affects model performance and interpretability, three distinct versions of each ultrasound frame were generated: raw, ROI, and binary mask representations. These representations capture progressively higher levels of visual abstraction, from the full acoustic field to a simplified depiction of the tongue surface and were produced using a structured preprocessing pipeline designed to preserve spatial alignment across representations. For each consonant token, the frame corresponding to the maximal lingual gesture was identified manually by an experienced SLT using synchronised audio, waveform, and spectrogram displays.

To support tongue localisation and quality control (QC), automated tongue surface tracking was performed using AAA's integrated DeepLabCut (DLC) module, which incorporates a ResNet-50 backbone [158]. The DLC model used in this study was specifically designed for tracking oral anatomy in mid-sagittal UTI and predicts 14 anatomically defined landmarks, including 11 points along the tongue surface spanning from the vallecula to the tongue tip. Prior validation of this model reported a mean localisation error of approximately 0.93 mm relative to human-labelled ground truth and demonstrated robustness across a range of probe FoV settings (approximately 60°–120°), probe depths, and subject sizes. While these properties indicate that the model is well-suited to ultrasound tongue tracking, additional QC and dataset-specific validation were performed to verify its behaviour on the present data, as described in Section 4.2.2.1.

The DLC module was trained directly within AAA, and all automated tracking outputs were visually inspected within AAA's interface. These contours were used exclusively for QC and validation of tongue localisation. Importantly, they were not used to generate ROI or mask representations, nor were they provided as inputs to the classification models. Figure 4.1 illustrates an example raw UTI frame and its corresponding AAA-generated DLC tongue contour. Automated tongue contours were subjected to structured QC within AAA. For each sequence, contours were overlaid on the corresponding B-mode frames and inspected frame by frame. A frame was flagged for exclusion if the predicted contour (i) deviated from the visible tongue–air interface across a continuous region, (ii) exhibited anatomically implausible frame-to-frame displacement inconsistent with smooth lingual motion, or (iii) corresponded to frames

in which the tongue surface was poorly visible due to acoustic shadowing or low contrast. Flagged frames were excluded from subsequent preprocessing.

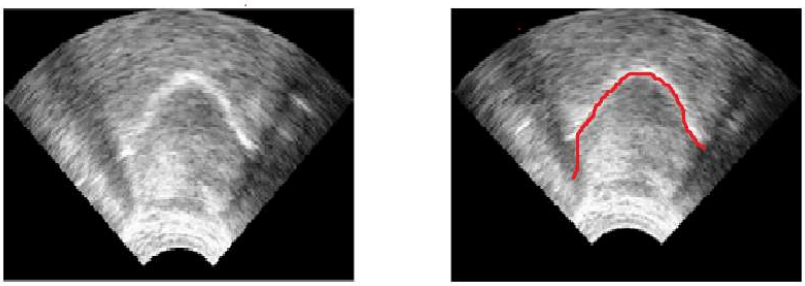

Figure 4. 1: Original UTI frame (left) and automatically tracked tongue surface AAA's built-in DLC module (right, red overlay), with the tongue tip on the right.

Image Representations

Raw B-Mode Images: Raw ultrasound images were extracted directly from the recordings by AAA. Frames corresponding to relevant consonant productions were exported as individual PNG images. No additional preprocessing was applied at this stage.

ROI Images: The ROI images isolate the tongue region and suppress irrelevant background features using a consistent preprocessing pipeline. Each B-mode frame was first converted from blue-green-red (BGR) to hue-saturation-value (HSV) colour space to enhance the separation between tongue tissue and darker surrounding regions. A targeted HSV threshold was applied, with lower and upper bounds of [H=0, S=20, V=70] and [H=20, S=255, V=255], respectively, generating an initial binary estimate of the tongue region. This range was selected to capture the characteristic mid-intensity, low-saturation appearance of tongue tissue in mid-sagittal UTI while excluding the darker acoustic shadow regions and brighter specular reflections. The initial mask was refined using morphological dilation with a 5×5 elliptical structuring element to ensure continuity along the tongue surface and bridge minor gaps introduced by speckle noise [159]. Contour detection was then performed, and the largest connected component, corresponding to the tongue and its acoustic shadow, was selected. This mask was applied to the raw image to yield the ROI frame, emphasising the articulatory region while suppressing unrelated background features.

Mask Images: The binary segmentation mask was derived from the ROI image by applying thresholding. A threshold value of 1 was applied, such that foreground tongue pixels with intensity values above this threshold were assigned a value of 255 and background pixels were assigned 0. This near-zero threshold ensures that all non-zero-pixel intensities retained in the ROI image are classified as foreground, producing a clean binary map that preserves the full extent of the tongue's shape while removing all speckle patterns and intensity-based artefacts. Figure 2 shows an example ROI image and its corresponding binary segmentation mask.

Figure 4. 2: ROI image (left) showing the isolated tongue region, and binary segmentation mask (right) created via thresholding, with the tongue tip on the right.

4.2.2.1 Tongue Contour Validation

Automated tongue surface tracking was validated to ensure that tongue localisation was sufficiently accurate and consistent to support the interpretation of image representation effects and Grad-CAM++ visualisations. Importantly, tongue contours were not used to generate ROI or mask representations, nor were they provided as inputs to the classification models. Instead, automated tracking was employed exclusively for QC and validation of tongue localisation in the frames selected for analysis.

Validation was performed by comparing automated tongue contours generated by AAA's integrated DLC module with manually accepted tongue splines. For each validated frame, the tongue surface was represented by 11 anatomically ordered points spanning from the vallecula to the tongue tip. Point-wise localisation error was computed as the Euclidean distance between corresponding manual and automated points in pixel space. Errors were aggregated at the frame level and summarised separately for TD and CP±L data.

As shown in Table 4.2, the mean localisation error was low for both groups, with average point-wise errors of 1.15 ± 1.11 pixels for TD data and 1.53 ± 0.84 pixels for CP±L data. Maximum observed errors remained below 4.1 pixels in all cases. Errors were smallest at the tongue tip and tongue root and slightly larger in the mid-tongue region, consistent with greater articulatory curvature and variability in this area. No frames exhibited gross tracking failure, implausible tongue geometry, or drift outside the acoustic fan. For a subset of TD frames, automated DLC contours required no manual correction following visual inspection and were therefore identical to the manually accepted splines, resulting in zero measured error for those cases.

These instances reflect agreement between automated tracking and expert judgement rather than circular validation. In contrast, CP±L frames more frequently required minor manual adjustment, consistent with the increased articulatory variability associated with compensatory speech patterns. Although spline coordinates were exported in pixel units due to the absence of physical depth calibration, the observed localisation errors were small relative to the spatial extent of the tongue surface, which typically spanned approximately 30–40 pixels in the vertical dimension. Consequently, the magnitude of tracking error represents only a small fraction of tongue height. It is insufficient to account for the substantial performance differences observed between raw, ROI, and masked image representations. Overall, this validation confirms that automated tongue localisation was reliable across both TD and CP±L datasets and that gross tracking errors did not confound subsequent analyses.

Table 4.2: Tongue Contour Localisation Validation.

Group	Mean Error (px)	SD (px)	Max Error (px)
TD	1.15	1.11	4.04
CLP	1.53	0.84	3.55

Together, these three representations, raw, ROI, and binary mask, enable a controlled investigation of how varying degrees of anatomical abstraction influence model learning. Because all representations were maintained at the same spatial resolution and coordinate alignment, any differences in performance or Grad-CAM++ attention can be attributed directly to the representation design, rather than geometric inconsistencies.

4.2.3. Field of View Normalisation

In the CP±L dataset, variation in the FoV was observed across recording sessions, ranging from approximately 119° to 133°. These differences arose from routine variation in probe calibration, participant positioning, and acquisition settings during clinical data collection. By contrast, the TD dataset was recorded under controlled laboratory conditions with a fixed FoV of 97°, offering a narrower but more consistent view of the tongue surface. Because other imaging parameters (probe type, depth, and frequency) were consistent across datasets, FoV constituted the primary systematic acquisition difference, which, if not addressed, introduces potential confounding factors by allowing models to capture acquisition-related geometric differences rather than genuine articulatory variation. Wider FoVs distribute identical anatomical structures across a broader angular span, whereas narrower FoVs compress these structures into a reduced projection. Such discrepancies are technical in nature rather than articulatory; however, they may inadvertently function as surrogate labels when one group is consistently recorded under different FoV conditions. To mitigate this risk, FoV harmonisation was implemented to ensure that classifier decisions reflect articulatory characteristics rather than acquisition-specific geometry.

FoV harmonisation was implemented by remapping all CP±L frames to the TD dataset's 97° FoV. This process does not reconstruct or infer anatomical structures absent from the original frame; rather, it standardises the angular spacing of ultrasound scanlines so that comparable portions of the tongue are projected consistently across participants and datasets. Minor probe-positioning differences are inherent to UTI; however, both datasets were recorded using the same stabilisation headset and probe type, reducing the likelihood of systematic group-linked variation. FoV was therefore the only acquisition parameter that differed consistently between CP±L and TD data. In practice, the original B-mode ultrasound frames were (i) mapped to a common angular domain corresponding to 97°, (ii) resampled using interpolation to adjust the angular spacing of pixels, and (iii) re-expressed on a fixed Cartesian grid while preserving imaging depth.

Two commonly used interpolation-based remapping methods were implemented: bicubic and spline interpolation [30]. These methods were selected in preference to simpler alternatives such as nearest-neighbour interpolation, which introduces blocky artefacts, and more computationally demanding methods such as Lanczos resampling, as they offer a practical balance between reconstruction quality and computational efficiency suitable for large-scale UTI preprocessing. Bicubic interpolation estimates pixel intensities using a weighted average of the surrounding neighbourhood, providing smooth transitions with minimal artefacts. Spline interpolation uses smooth polynomial fitting, further preserving structural continuity. Including both methods also allows sensitivity testing of whether the choice of interpolation algorithm materially affects downstream classification performance and model interpretability.

Both methods remap the original fan-shaped data into a 97° Cartesian grid while preserving depth. No cropping was performed, and no anatomical structures were added or removed; only the angular sampling domain was standardised. Because the exported B-mode images are already in Cartesian form, the visual appearance of the harmonised frames remains very similar to the originals. Interpolation modifies the underlying geometric mapping but produces subtle visible changes due to smooth resampling.

To examine the impact of FoV variation and the sensitivity to interpolation method, we constructed three dataset versions. An overview of these datasets is presented in Table 4.3. All three datasets were processed identically, enabling a controlled evaluation of classification performance and Grad-CAM++ attention patterns. By comparing harmonised and unmodified conditions, we isolate the effect of acquisition geometry from true articulatory differences and assess whether image representation interacts with FoV in influencing downstream model behaviour.

Table 4.3: Overview of Dataset Variations Used for FoV Experiments.

Dataset	FoV	Interpolation	Contents
Original Dataset	Mixed	N/A	Raw, ROI, Mask
Bicubic Dataset	Fixed	Bicubic	Raw, ROI, Mask
Spline Dataset	Fixed	Spline	Raw, ROI, Mask

Through this normalisation process, geometric inconsistencies between datasets were substantially reduced, creating a controlled framework for evaluating how FoV alignment affects both quantitative model performance and qualitative interpretability. This standardised foundation underpins the explainable AI analyses presented later in Section 4.3.3.

4.2.4. Model Architecture and Training Strategy

For the binary classification task, EfficientNet-B0 was employed as the classification backbone. EfficientNet-B0 is a lightweight CNN architecture designed using a compound scaling strategy that balances network depth, width, and input resolution. This architecture has demonstrated strong performance in medical image analysis because it provides a favourable trade-off between accuracy, model size, and computational efficiency [160]. This trade-off is particularly relevant to UTI, where datasets are relatively small, images are noisy and low-contrast, and the long-term aim is to support clinically usable, potentially real-time systems. A very large architecture could increase the risk of overfitting and would be less suitable for deployment, whereas EfficientNet-B0 provides sufficient representational capacity while remaining computationally efficient.

Although Chapter 3 showed that ImageNet-pretrained filters alone are insufficient to fully capture the characteristic speckle patterns, acoustic artefacts, and contrast properties of UTI, transfer learning was used here as an initialisation strategy rather than as a fixed feature extractor. To balance domain adaptation with training stability, a partial fine-tuning strategy was employed: the early convolutional layers of EfficientNet-B0 were frozen to preserve generic low-level feature detectors learned during ImageNet pretraining, while later layers were fine-tuned to adapt higher-level representations to the statistical properties of ultrasound data.

This approach is grounded in the established observation that early CNN layers encode relatively generic features, such as edges and texture primitives, whereas deeper layers encode more task-specific representations that benefit from domain-specific adaptation.

The choice of EfficientNet-B0 was also motivated by the primary aim of this chapter. The objective was not to propose a new classification architecture, but to systematically isolate the effects of image representation and FoV variability on classification performance and interpretability. Using a single, well-characterised and computationally efficient backbone with a consistent fine-tuning protocol across all experimental conditions ensured that observed performance differences could be attributed primarily to the input representation and FoV configuration rather than to architectural variation. EfficientNet-B0, therefore, provided a controlled and deployment-relevant model for evaluating how raw, ROI, and mask-based UTI representations influence both predictive performance and Grad-CAM-based interpretability. Table 4.4 shows the model summary of hyperparameters. The dataset comprised a slightly imbalanced class distribution. To address this, a weighted random sampler was applied during training, assigning each sample a weight inversely proportional to its class frequency, ensuring that both classes were represented equally across training batches.

Table 4. 4: Summary of Hyperparameters for EfficientNet-B0.

Component	Setting	Justification
Model	EfficientNet-B0	Lightweight architecture suitable for small UTI datasets and future real-time deployment.
Input representations	Raw, ROI, and binary mask UTI images	Enables controlled comparison of how representation affects performance and interpretability.
Input size	224 × 224	Standard EfficientNet-compatible resolution; balances anatomical detail with computational efficiency.
Pretraining	ImageNet pretrained	Used as an initialisation strategy to improve convergence on a small dataset.
Fine-tuning strategy	Early layers frozen; later layers fine-tuned	Preserves generic low-level features while adapting higher-level representations to ultrasound-specific patterns.
Optimiser	Adam	Provides stable optimisation for fine-tuning under limited-data conditions.
Learning rate	1×10^{-4}	Lower learning rate appropriate for partial fine-tuning to avoid disrupting pretrained weights.
Batch size	32	Chosen based on GPU memory and stable gradient estimation.
Epochs	30	Sufficient for convergence while limiting overfitting.
Loss function	Binary cross-entropy	Appropriate for TD vs CP±L binary classification.
Regularisation	Batch normalisation and dropout	Used to improve stability and reduce overfitting.
Evaluation metrics	Accuracy, precision, recall, F1-score, inference time, Grad-CAM	Captures predictive performance, clinical sensitivity, computational feasibility, and interpretability.

Overall, the combination of partial fine-tuning, controlled use of pretrained initialisation, and a lightweight architecture provided a robust and computationally efficient model for analysing representation-driven effects in UTI. The same architectural choice also facilitates future integration into real-time diagnostic systems, as explored in Chapter 7.

4.2.5 Evaluation Metrics and Explainability

To systematically assess model performance across the different image representations, each dataset was divided into three subsets corresponding to the preprocessing types: raw, ROI, and masked images. Separate models were trained on each subset under identical configurations, allowing for a controlled comparison of how representation design influences performance within a consistent experimental framework. Model accuracy was quantified using standard classification metrics: accuracy, precision, recall, and F1 score. These measures provide complementary perspectives on predictive reliability, accuracy reflecting overall correctness, precision capturing resistance to false positives, recall quantifying sensitivity to true articulatory events, and F1 score balancing both dimensions. To quantify uncertainty around point estimates, 95% CIs were computed for key accuracy figures using the Wilson score interval. In addition, inference time was recorded to evaluate computational efficiency and the feasibility of future real-time deployment.

Beyond quantitative accuracy metrics, interpretability was assessed using Grad-CAM++, which generates class-specific saliency maps indicating the regions that most strongly contributed to each prediction. These visualisations allowed us to qualitatively examine whether the model relied on articulatory structures (e.g., tongue surface) or acquisition-related cues. Grad-CAM++ was selected due to its improved localisation and sensitivity to multiple relevant regions compared to standard Grad-CAM. To analyse group-level patterns in model attention, we performed an unsupervised clustering analysis on the Grad-CAM++ activation features. For each image, the Grad-CAM++ heatmap was summarised into a 5-dimensional feature vector consisting of total activation intensity, mean activation intensity, activation variability (standard deviation (SD) of intensity), and the activation-centre coordinates (X, Y). These features capture both the overall strength of the model's attention and its spatial distribution within the ultrasound frame, providing quantitative descriptors of where and how strongly the model focuses during classification.

K-means clustering was applied to explore whether distinct patterns of model attention emerged beyond the binary diagnostic classification [161]. Although the classification task itself was binary (CP±L vs TD), we tested k = 2, 3, and 4 clusters to investigate whether subgroups within diagnostic categories or intermediate attention patterns might exist. Clustering was performed on the combined dataset, without access to diagnostic labels or phoneme categories, ensuring that any structure emerged solely from activation pattern similarity. Clustering quality was assessed using silhouette scores, which measure cluster cohesion and separation by comparing within-cluster distances to between-cluster distances [162]. Silhouette scores range from -1 to +1, with higher values indicating better-defined clusters. Values above 0.5 generally indicate good separation, while values between 0.25 and 0.5 suggest weak but detectable structure. The optimal number of clusters was selected based on the highest silhouette score, combined with assessment of cluster interpretability and stability. To validate the robustness of the clustering solution, hierarchical clustering (Ward linkage) was applied as an alternative method [163]. Ward linkage builds clusters by iteratively merging pairs that minimise within-cluster variance, using a bottom-up agglomerative approach. Comparison between k-means and hierarchical clustering allows assessment of whether the identified structure is robust across different algorithmic assumptions: k-means assumes spherical clusters and uses iterative centroid optimisation, while Ward hierarchical clustering uses a deterministic merging process based on variance minimisation. Consistency across both methods provides evidence that clustering patterns reflect genuine data structure rather than algorithm-specific artefacts.

For phoneme-specific analyses, k-means clustering (k=2) was applied separately to alveolar, palatal, and velar subsets to examine whether diagnostic group separation varied systematically by place of articulation. This stratified approach allows assessment of whether attention pattern differences are consistent across different articulatory regions or whether certain phoneme categories show stronger diagnostic separation than others, as would be expected from clinical observations that anterior consonants are most affected in cleft speech. All clustering analyses were visualised using PCA, which projects the 5-dimensional activation features onto the two principal components capturing the greatest variance. While PCA provides useful low-dimensional visualisation, it should be noted that 2D projections may not fully represent subtle group differences present in the full feature space.

Consequently, quantitative metrics (silhouette scores, cluster composition cross-tabulations) were used alongside visual inspection to evaluate clustering quality. Combining quantitative evaluation with visual explainability provided a holistic understanding of model behaviour. This dual analysis is crucial for clinical translation: high accuracy alone is insufficient without transparent reasoning that clinicians can interpret and trust. The integration of standardised datasets, structured image representations, and tightly controlled training conditions created a rigorous foundation for analysing how acquisition parameters and preprocessing choices shape both model performance and interpretability. The next section presents the corresponding results, including quantitative metrics and Grad-CAM++ visualisations, which demonstrate how FoV normalisation and image representation influence classification outcomes and model attention patterns.

4.3 Results

4.3.1 Model Performance Across Image Representations

To examine how image representation influences classification performance, three separate models were trained using raw, ROI, and masked versions of the ultrasound data. The objective was to quantify how progressively restricting visual information from full raw frames to tongue-focused representations affects predictive accuracy and computational efficiency.

Models were evaluated using accuracy, precision, recall, and F1-score, and inference time was measured to assess suitability for real-time clinical use. The results, summarised in Table 4.5, show that models trained on raw ultrasound frames consistently achieved the highest overall classification performance across all FoV conditions. In the original mixed-FoV dataset, the raw representation achieved an accuracy of 97.44%, compared with 72.77% for ROI and 72.14% for masked images. This performance gap persisted after FoV standardisation: in the bicubic-rescaled dataset, accuracies were 94.53%, 80.55%, and 87.13%, respectively, while in the spline-rescaled dataset they were 94.45%, 82.24%, and 86.50%. The CIs confirm that the performance differences between representations under standardised FoV conditions are statistically reliable. In the spline-rescaled dataset, masked accuracy of 86.50% [83.1%, 89.9%] does not overlap with ROI accuracy of 82.24% [78.5%, 86.0%], supporting the conclusion that anatomically focused masking provides a genuine performance benefit beyond ROI alone.

Raw-input models also exhibited the highest recall across all datasets. However, this improved predictive performance was accompanied by slightly longer inference times (≈ 0.082–0.087 s per frame). In contrast, ROI representations yielded the fastest inference times (≈ 0.063–0.069 s), reflecting the reduced spatial complexity of the input, while masked representations provided an intermediate trade-off between accuracy and computational cost (≈ 0.072 s). To assess whether the observed performance differences reflect stable generalisation rather than optimisation artefacts, additional diagnostics were examined. Training and validation learning curves, confusion matrices for all representations and datasets are provided in Appendix B.

Table 4.5: Model Performance Across Image Representations and Datasets.

Dataset	**Representation**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Inference Time (s)**	**95% CIs**
Original	Raw	97.44%	94.48%	100%	97.16%	0.087	95.9%,98.9%
	ROI	72.77%	70.61%	78.39%	74.30%	0.069	68.4%,77.2%
	Masked	72.14%	67.74%	82.89%	74.56%	0.074	67.7%,76.6%
BiCubic	Raw	94.53%	87.76%	100%	93.48%	0.082	92.3%,96.8%
	ROI	80.55%	82.17%	78.75%	80.43%	0.063	76.7%,84.4%
	Masked	87.13%	87.71%	86.61%	87.16%	0.072	83.8%,90.4%
Spline	Raw	94.45%	94.51%	100%	97.18%	0.082	92.2%,96.7%
	ROI	82.24%	82.23%	82.92%	82.57%	0.064	78.5%,86.0%
	Masked	86.50%	87.55%	85.36%	86.44%	0.072	83.1%,89.9%

Taken together, these results indicate that retaining the full ultrasound frame provides the model with additional discriminative information that improves classification performance, whereas restricting the input to tongue-focused representations reduces accuracy but improves computational efficiency. Importantly, the relative performance ordering (raw > masked > ROI) remained consistent across all FoV conditions, suggesting that representation choice exerts a stable influence on classification outcomes independent of acquisition harmonisation. However, high numerical performance does not necessarily imply anatomically grounded decision-making, as subsequent Grad-CAM++ analysis reveals systematic differences in the regions

driving model predictions. The implications of this performance–efficiency trade-off, and its relationship to model interpretability, are examined in subsequent sections.

4.3.2 Impact of FoV Alignment

To evaluate the influence of FoV variability on classification performance, results were compared between the original mixed-FoV CP±L dataset and two FoV-standardised variants rescaled to a uniform 97° FoV using bicubic and spline interpolation. This analysis isolates the effect of acquisition geometry alignment while holding model architecture, training protocol, and image representation constant.

FoV standardisation produced consistent performance improvements for anatomically constrained representations. In the original mixed-FoV dataset, ROI and masked inputs achieved accuracies of approximately 72%. Following FoV alignment, ROI accuracy increased to 80.55% with bicubic rescaling and 82.24% with spline rescaling, while masked representations improved to 87.13% and 86.50%, respectively. These gains indicate that geometric alignment enhances the model's ability to learn articulatory patterns when input representations are explicitly focused on the tongue region. In contrast, models trained on raw ultrasound frames exhibited a modest reduction in accuracy following FoV standardisation, decreasing from 97.44% in the mixed-FoV dataset to approximately 94.5% in the rescaled datasets. Rather than indicating degraded articulatory modelling, this differential effect reveals important distinctions in what each representation enables the model to learn. Specifically, the slight decline in raw-input performance suggests that these models partially exploited FoV-specific acquisition characteristics and peripheral image structure that vary systematically across datasets, in addition to true articulatory information. Standardising the FoV reduces these acquisition-dependent cues, thereby improving anatomical validity even when numerical accuracy decreases slightly.

Qualitative Grad-CAM++ analysis supports this interpretation. Models trained on FoV-standardised datasets exhibited more spatially consistent and anatomically coherent activation patterns, particularly for ROI and masked representations, with attention concentrated along the tongue surface. In contrast, models trained on mixed-FoV raw data more frequently displayed peripheral or dispersed activations, consistent with partial reliance on acquisition-related artefacts.

Overall, the differential impact of FoV alignment highlights a trade-off between numerical accuracy and anatomical validity. While raw models achieve the highest absolute accuracy, their performance is more sensitive to acquisition heterogeneity. In contrast, FoV-standardised ROI and masked representations yield more robust and anatomically grounded behaviour, which is preferable for clinically interpretable and trustworthy ultrasound-based speech analysis. Consequently, FoV standardisation plays a critical role in promoting representations that generalise based on articulatory structure rather than dataset-specific artefacts.

4.3.3 Grad-CAM Visualisation of Model Attention

To assess model interpretability and identify which image regions most strongly influenced classification decisions, Grad-CAM++ was applied to the final convolutional layers of each trained network. The resulting heatmaps highlight areas contributing most significantly to the model's predictions, providing a qualitative measure of anatomical focus and offering insight into how the network interprets articulatory structure. Figure 4.3 presents representative Grad-CAM++ heatmaps for models trained on raw ultrasound images. These visualisations reveal broad, diffuse attention patterns that extend across both articulatory and non-articulatory regions. Activation frequently spread into background tissue and peripheral artefacts, indicating that the network partially relied on non-tongue cues when distinguishing between TD and CP±L samples. This behaviour suggests that, while raw frames contain the richest articulatory detail, they also expose the model to irrelevant visual features that can weaken generalisation.

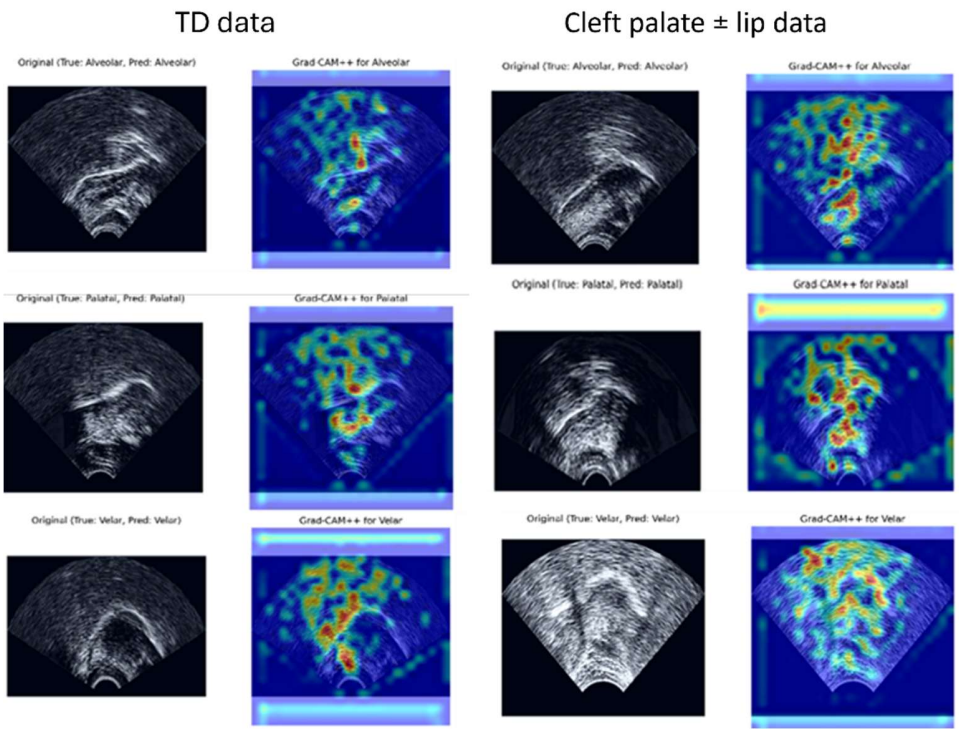

Figure 4. 3: Grad-CAM Heatmaps from a Model Trained on Raw UTI.

By contrast, Figure 4.4 shows Grad-CAM++ outputs for models trained on ROI and masked representations. The ROI-based models showed more consistent attention across the tongue surface, with activations concentrated in midline articulatory regions. The masked representation produced the most anatomically coherent focus: heatmaps were sharply confined to the tongue contour with minimal spillover into the surrounding background. Across TD samples, these attention maps aligned closely with expected articulatory gestures, for example, focused activation near the tongue tip for alveolar consonants and posterior dorsum emphasis for velar sounds, indicating that the model captured linguistically meaningful spatial patterns. In CP±L samples, Grad-CAM++ revealed more diffuse or posteriorly displaced activations, corresponding to the atypical or compensatory tongue postures characteristic of this population.

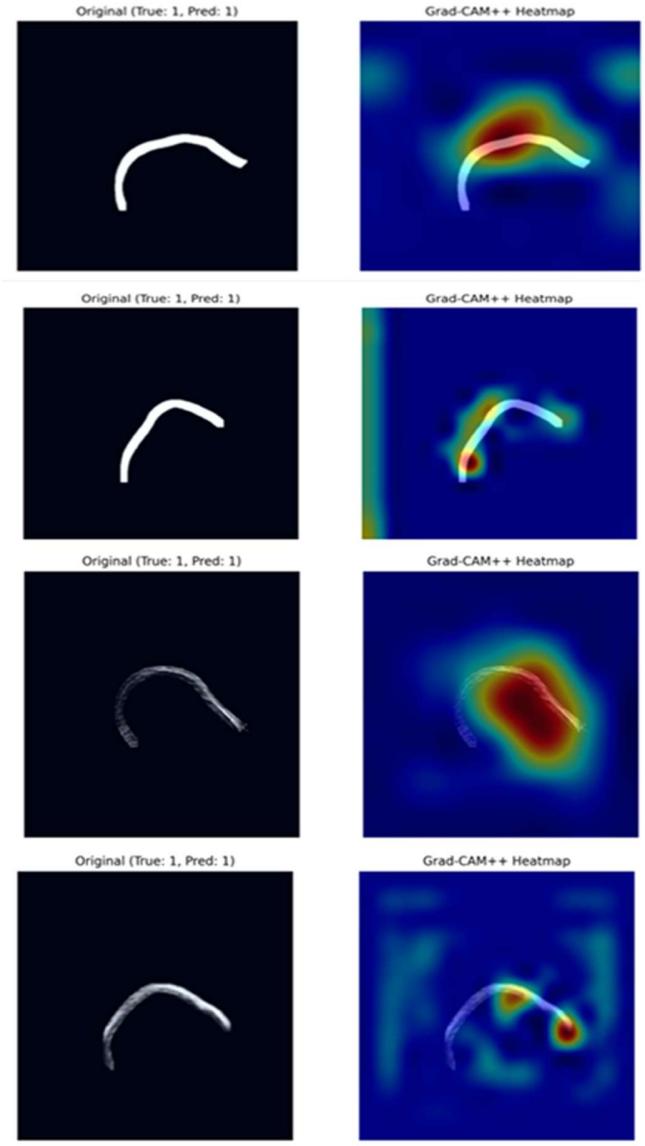

Figure 4. 4: Grad-CAM Heatmaps for Models Trained on ROI and Masked UTI.

Collectively, these results confirm that representation design directly influences model interpretability. Constraining the network's input to anatomically relevant regions (as in ROI or mask representations) encourages attention to clinically meaningful features, whereas models trained on unfiltered raw inputs are more susceptible to spurious correlations from background noise.

4.3.4 Statistical Comparison of Grad-CAM Activations

To quantitatively assess whether model attention differed systematically between TD and CP±L speech, five summary statistics were extracted from each Grad-CAM++ heatmap: total activation intensity (sum of all pixel weights), mean activation intensity (average weight), activation dispersion (SD), and activation centroid coordinates (X, Y). These metrics capture both the strength and spatial distribution of model attention.

Independent-samples t-tests compared TD and CP±L groups separately for each phoneme category (alveolar, palatal, velar). Given the multiple comparisons performed (5 features × 3 phoneme categories = 15 tests), Bonferroni correction was applied to control the family-wise error rate at $\alpha = 0.05$ [164]. The corrected significance threshold was $\alpha = 0.05/15 = 0.00333$. Table 4.6 reports only features demonstrating statistical significance. Effect sizes were quantified using Cohen's d, calculated as the difference in means divided by the pooled SD [165]. Following conventional guidelines, $|d| < 0.2$ indicates a negligible effect, 0.2-0.5 a small effect, 0.5-0.8 a medium effect, 0.8-1.5 a large effect, and $|d| > 1.5$ a very large effect. Given the sample sizes, the study had adequate power (>80%) to detect large effects ($d > 0.8$) but limited power for detecting moderate effects ($d \approx 0.5$), as is typical for pilot studies with modest sample sizes.

After Bonferroni correction ($\alpha = 0.00333$), 10 of 15 features showed statistically significant differences between TD and CP±L speech (Table 4.5). All five features tested for alveolar consonants reached significance, with very large effect sizes observed for SD Intensity (Cohen's $d = 4.30$, $p = 1.27\times10^{-13}$) and Centre Y ($d = 2.48$, $p = 1.47\times10^{-7}$), and large effects for Total Intensity, Mean Intensity (both $d = 2.05$, $p = 3.41\times10^{-6}$), and Centre X ($d = 1.46$, $p = 7.36\times10^{-4}$). These results indicate that CP±L samples showed significantly higher and more dispersed activation patterns with more posterior and superior spatial positioning compared to TD controls.

For palatal consonants, three of five features remained significant after correction: Total Intensity ($d = 1.33$, $p = 1.91\times10^{-3}$), Mean Intensity ($d = 1.33$, $p = 1.91\times10^{-3}$), and SD Intensity ($d = 3.55$, $p = 6.01\times10^{-9}$). Spatial features showed weaker effects: Centre Y showed a nominally significant difference ($p = 0.0317$) that did not survive multiple comparison correction, while Centre X showed no significant difference ($p = 0.180$). This pattern suggests that palatal consonants show robust intensity-based differences but more subtle spatial shifts compared to

alveolars. For velar consonants, three features remained significant: Total Intensity (d = 1.69, p = 1.14×10^{-4}), Mean Intensity (d = 1.69, p = 1.14×10^{-4}), and SD Intensity (d = 3.03, p =9.83×10^{-9}). Neither spatial feature showed significant differences (Centre X: p = 0.119; Centre Y: p = 0.937), consistent with clinical observations that both TD and CP±L speakers use similar posterior tongue dorsum raising for velar production, limiting the magnitude of spatial differences.

The most consistent discriminator across all three phoneme categories was SD Intensity (activation dispersion), which showed very large effect sizes for alveolars (d = 4.30), palatals (d = 3.55), and velars (d = 3.03), all with $p < 10^{-8}$. This indicates that increased variability and spatial dispersion of model attention are robust characteristics of CP±L speech across all places of articulation. In contrast, spatial displacement (Centre X, Centre Y) was significant only for alveolar consonants, where compensatory backing and abnormal tongue-tip control are most pronounced clinically.

Table 4.6: Summary of Significant Grad-CAM Activation Differences between TD and CP±L Speech.

Consonant Category	Feature	*t*-Statistic	*p*-value	Cohen's *d*	Effect Magnitude
Alveolar	Total Intensity	−5.22	3.41×10^{-6}*	2.05	Large
	Mean Intensity	−5.22	3.41×10^{-6}*	2.05	Large
	SD Intensity	−10.98	1.27×10^{-13}*	4.30	Very large
	Centre X	3.72	7.36×10^{-4}*	1.46	Large
	Centre Y	−6.33	1.47×10^{-7}*	2.48	Large
Palatal	Total Intensity	−3.39	1.91×10^{-3}*	1.33	Large
	Mean Intensity	−3.39	1.91×10^{-3}*	1.33	Large
	SD Intensity	−9.07	6.01×10^{-9}*	3.55	Very large
Velar	Total Intensity	−4.30	1.14×10^{-4}*	1.69	Large
	Mean Intensity	−4.30	1.14×10^{-4}*	1.69	Large
	SD Intensity	−7.73	9.83×10^{-9}*	3.03	Very large

Note: * indicates significance after Bonferroni correction. Only features showing significant differences after correction are displayed. Five comparisons did not reach significance after correction: Palatal Centre Y (p = 0.0317), Palatal Centre X (p = 0.180), Velar Centre X (p = 0.119), Velar Centre Y (p = 0.937), all with p > 0.00333.

Overall, the statistical analysis confirms that Grad-CAM++ features capture systematic and anatomically meaningful differences in model attention between TD and CP±L speech. Importantly, the strongest effects were observed for features describing activation strength and spatial concentration, rather than absolute position alone, suggesting that variability and instability of articulatory control, rather than uniform displacement, characterise cleft-affected speech. These quantitative findings reinforce the qualitative Grad-CAM++ observations presented earlier and demonstrate that explainability-driven features can reveal clinically relevant distinctions that are not evident from classification accuracy alone.

4.3.5 Clustering of Grad-CAM Features

K-means clustering was applied to the Grad-CAM++ activation features to examine whether the model's attention patterns formed distinct groups. Clustering was performed in an unsupervised manner on the combined dataset of TD and CP±L, representing all speakers and phoneme categories across the three places of articulation. Initial exploration tested k = 2, 3, and 4 clusters to assess whether substructure existed beyond the binary diagnostic classification. Silhouette scores were k=2: 0.423, k=3: 0.350, k=4: 0.366. The highest silhouette score for k=2 confirmed that a two-cluster solution provided optimal partitioning quality. For k=3 and k=4, lower silhouette scores indicated that additional clusters fragmented the data without revealing meaningful structure. For k=2, clustering produced two groups; this pattern demonstrates that Grad-CAM++ activation features capture diagnostically meaningful differences while allowing for realistic clinical overlap. To validate clustering stability, hierarchical clustering was applied to the same feature space, achieving a silhouette score of 0.399. Although slightly lower than k-means (0.423), this confirms the existence of a two-group structure and demonstrates that the pattern is robust across different clustering algorithms. The consistency of findings across both methods provides evidence that the two-cluster structure reflects genuine organisation in the data rather than algorithm-specific artefacts. Figure 4.5 shows the PCA projection of k=2 clustering across all speakers and phonemes.

TD samples formed a relatively cohesive cluster predominantly in the left and central regions of the PCA space, reflecting more consistent articulatory patterns and stable model attention across typical productions. In contrast, the CP±L samples showed greater dispersion across both principal components, with substantial spread, indicating higher articulatory variability and diverse compensatory strategies in cleft-affected speech.

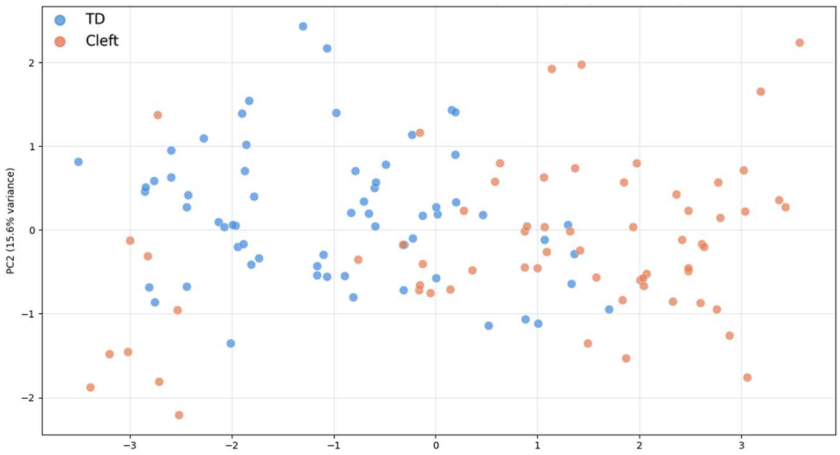

Figure 4. 5: PCA Projection of Grad-CAM++ Activation Features Across all Speakers.

To examine whether diagnostic group separation varied systematically by place of articulation, k=2 clustering was applied separately to alveolar, palatal, and velar subsets. All three phoneme categories showed comparable clustering quality, with silhouette scores of 0.445 (alveolar), 0.435 (palatal), and 0.465 (velar), indicating consistent separation between TD and CP±L attention patterns across different articulatory regions.

Visual inspection of the PCA projections revealed distinct patterns across phoneme categories. Alveolar consonants (left panel) showed clear horizontal separation, with TD samples predominantly clustering in the central-right region and CP±L samples distributed more broadly across the left side. Some overlap was visible in the central region, consistent with the overall pattern observed in Figure 4.6. Palatal consonants (middle panel) exhibited similar separation, though with greater vertical spread for both groups, suggesting higher variability in activation patterns for mid-tongue articulations. The separation remained primarily along PC1, with TD samples concentrated in the right-central region and CP±L samples more dispersed on

the left. Velar consonants (right panel) showed the highest silhouette score (0.465), indicating the most distinct clustering quality.

However, visual inspection revealed substantial overlap between groups, particularly in the central and right regions. This apparent contradiction high silhouette score but visible overlap, reflects the fact that silhouette scores measure within-cluster cohesion relative to between-cluster separation. For velars, both groups formed relatively tight, internally consistent clusters despite some spatial proximity in the 2D projection.

Across all three phoneme categories, the 2D PCA projections showed consistent separation trends, with TD samples generally clustering toward the central-left regions and CP±L samples showing broader dispersion. However, the degree of overlap and the tightness of clustering varied, with alveolar consonants showing the clearest visual separation and palatal/velar consonants showing more intermixing. These patterns align with clinical observations that cleft-related articulatory deviations are most pronounced for anterior consonants, where compensatory backing and abnormal tongue-tip control are commonly observed [30], [166], while posterior consonants (velars) may show more subtle differences as both groups use posterior tongue dorsum raising.

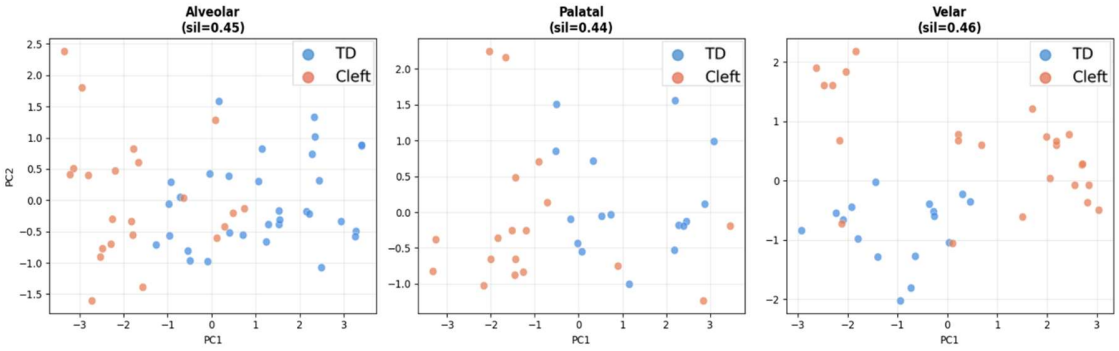

Figure 4. 6: Phoneme-Specific PCA projections of Grad-CAM++ Activation Features for Alveolar, Palatal, and Velar Consonants Across all Speakers.

Together, these findings demonstrate that Grad-CAM++ features encode articulatory structure sufficiently to distinguish between typical and cleft-affected speech across multiple phoneme categories.

4.4 Discussion

This study provides new insight into how image representation and FoV variability influence both model performance and interpretability in UTI for speech disorder diagnosis. The results demonstrate that image representation plays a decisive role not only in classification accuracy but also in the reliability and transparency of DL models applied to technical speech assessment.

Across all datasets, models trained on raw ultrasound frames consistently achieved the highest classification scores (94–97% accuracy). However, Grad-CAM++ analysis revealed that these models frequently directed attention toward peripheral artefacts and background regions rather than focusing exclusively on the tongue surface. This discrepancy between numerical performance and anatomical validity is consistent with what has been termed "shortcut learning" in medical AI, where models exploit spurious correlations in training data rather than learning clinically meaningful features [167]. Similar phenomena have been documented across medical imaging domains. In chest radiography, Zech et al. [168] demonstrated that pneumonia classifiers achieved high accuracy by detecting hospital-specific image markers rather than pathological features, failing catastrophically when deployed at new sites. Winkler et al. [169] showed that skin lesion classifiers relied on ruler presence and image backgrounds rather than lesion morphology.

In the present study, the high raw-input accuracy may not arise solely from anatomically grounded articulatory modelling. Instead, raw frames expose the model to acquisition-related cues that may correlate systematically with class labels, including FoV geometry, probe shadows, and intensity distributions in peripheral regions. The differential impact of FoV standardisation is consistent with this interpretation: when acquisition geometry was harmonised, performance improved for anatomically constrained representations (ROI: +7–10 pp; Masked: +14–15 pp) but declined slightly for raw inputs (−3 pp). This pattern suggests that part of the raw model's apparent advantage may derive from FoV-specific acquisition characteristics rather than invariant articulatory structure.

However, this interpretation should be treated with caution. FoV standardisation involves interpolation and resampling, which can alter local pixel intensities, smooth fine speckle texture, and slightly modify edge sharpness. These effects may influence CNN feature extraction independently of shortcut learning, particularly in raw ultrasound frames where speckle patterns and acoustic texture contribute strongly to the learned representation. Therefore, the observed reduction in raw-input performance may reflect both the removal of acquisition-specific shortcut cues and the effect of interpolation-related changes introduced during standardisation.

A useful control experiment would be to apply an equivalent interpolation/resampling procedure without changing the FoV or anatomical coverage. For example, raw images could be resampled to an intermediate resolution and then mapped back to their original geometry, preserving the same visible anatomy while introducing comparable interpolation effects. If performance decreased under this resampling-only condition, this would indicate that interpolation artefacts contribute to the accuracy reduction. Conversely, if performance remained stable under resampling alone but decreased only after FoV standardisation, this would provide stronger evidence that the original raw-input model was exploiting FoV-related shortcut cues.

Critically, this effect cannot be attributed solely to tongue-tracking error. ROI and masked representations were generated using intensity-based preprocessing, including thresholding and morphological operations, rather than DLC-derived contours, and independent validation (Section 4.2.2.1) confirmed that tongue-localisation errors were small (1.15–1.53 pixels) relative to tongue dimensions (~30–40 pixels vertically). The performance differences, therefore, most likely reflect representation-level effects, although future controlled experiments are needed to fully separate shortcut learning from interpolation-related artefacts.

In contrast to raw inputs, models trained on ROI and masked representations achieved lower numerical accuracy (72-87%) but exhibited markedly improved anatomical focus in their Grad-CAM++ attention maps. By constraining the model's input to the tongue region, these representations suppress peripheral acquisition artefacts and encourage attention to clinically meaningful articulatory features. This finding aligns with principles of inductive bias in ML: when task-relevant constraints are encoded in the input representation, models learn more robust and generalisable features [170].

In medical imaging, similar benefits have been observed when anatomical priors are incorporated through segmentation [171]. The reduced performance observed for ROI and masked inputs is therefore best understood not as methodological failure but as the removal of contextual cues that aided dataset-specific discrimination. Additional contributing factors may include: (i) loss of global spatial reference information (e.g., relative tongue position within the ultrasound fan), and (ii) use of a fixed threshold that may not optimally segment the tongue across all speakers and acquisition conditions. Despite these limitations, the resulting attention patterns were substantially more consistent with known speech motor control mechanisms.

The importance of anatomically focused attention is particularly relevant to the assessment of speech disorders. Studies using UTI to analyse articulation have consistently demonstrated that specific tongue regions are critical for different sound categories and that deviations from typical patterns serve as key diagnostic markers. Preston et al. [143] outlined characteristic tongue configurations across places of articulation, emphasising that accurate identification of the active articulator region is essential for distinguishing typical from disordered patterns and guiding treatment decisions.

In children with CP±L, instrumental studies have revealed systematic deviations from typical spatial patterns. Cleland et al. [30] documented increased tongue dorsum activity during alveolar consonant production in children with CP±L, with posteriorly displaced tongue configurations reflecting compensatory backing strategies. Similarly, Hashemi and Xing [166] found that children with velopharyngeal insufficiency exhibited increased and more variable tongue dorsum raising during velar productions compared to TD peers. Wyatt et al.'s [51], a comprehensive review further emphasised that cleft-related speech errors frequently involve place of articulation shifts, with children substituting posterior (pharyngeal, glottal) for anterior oral targets.

Our Grad-CAM++ analyses captured these expected spatial distinctions in TD speech, with attention concentrated anteriorly for alveolar consonants and posteriorly for velar consonants. More importantly, attention patterns in CP±L speech revealed posteriorly shifted and spatially diffuse activations consistent with the compensatory strategies documented by Cleland et al. [30]. Statistical analysis confirmed these deviations: CP±L alveolars showed significantly more posterior activation centres (Cohen's d = 1.46) and greater spatial dispersion (Cohen's d = 4.30) compared to TD controls (Table 4.5). These quantitative metrics demonstrate that explainability

methods can automatically extract clinically relevant spatial deviations that instrumental phonetic studies have identified through expert visual analysis.

Although these Grad-CAM++ findings are consistent with established articulatory patterns in TD and CP±L speech, their anatomical interpretation could be strengthened through direct validation against expert-labelled tongue contours. A future analysis could extract spatial descriptors from each Grad-CAM++ heatmap, such as the activation centroid, vertical centre of activation, posterior–anterior activation position, or proportion of heatmap energy overlapping the tongue region.

These features could then be compared with expert-derived anatomical measurements, including tongue-contour position, tongue-tip elevation, tongue-dorsum height, or the location of maximal tongue raising. Strong agreement between heatmap-derived spatial features and expert-labelled tongue measures would provide evidence that the model is attending to clinically meaningful articulatory structures rather than peripheral artefacts or correlated image statistics. Conversely, weak overlap or systematic displacement of activation away from the traced tongue contour would indicate that the model's decision-making is less anatomically grounded, even if classification performance remains high. Such contour-based validation would provide a more objective bridge between Grad-CAM++ explanations and clinically interpretable measures of speech production.

The clustering analysis of Grad-CAM++ features revealed that model attention patterns naturally separate into two primary modes that strongly align with diagnostic categories. K-means clustering with k=2 achieved a silhouette score of 0.423, indicating moderate-to-good cluster definition. This level of separation is substantial but not absolute, reflecting the clinical reality that speech disorders exist on a continuum rather than as discrete categories. The observed overlap (~20%) is clinically expected: speech disorders demonstrate substantial individual variability, with some children with CP±L producing near-typical articulations for certain phonemes, particularly when structural deficits are mild, or compensatory strategies are inconsistently applied. To validate clustering stability, hierarchical clustering was applied to the same feature space, achieving a silhouette score of 0.399. Although slightly lower than k-means (0.423), this confirms the existence of a two-group structure and demonstrates that the pattern is robust across different clustering algorithms.

The consistency of findings across both methods, despite their different optimisation approaches (k-means uses iterative centroid refinement, Ward uses variance-minimisation merging), provides evidence that the two-cluster structure reflects genuine organisation in the data rather than algorithm-specific artefacts. The predominant horizontal separation of groups in PCA projections suggests that the model's attention mechanism encodes a primary diagnostic signal alongside secondary variations in individual articulatory strategies, with the greater dispersion of CP±L samples aligning with established clinical understanding of cleft-affected speech. Wyatt et al. [51] documented substantial individual variability in compensatory articulation patterns, with children adopting different strategies (backing, glottal replacement, pharyngeal constriction) depending on structural constraints and learned motor patterns [51].

Our clustering results provide quantitative evidence that this clinical heterogeneity is reflected in the spatial characteristics of model attention, with CP±L samples occupying a broader region of the activation feature space than the more tightly clustered TD samples. Phoneme-specific clustering revealed consistent separation quality across places of articulation, with silhouette scores ranging from 0.435 to 0.465, indicating that Grad-CAM++ features capture diagnostic differences reliably regardless of tongue region. However, the visual characteristics of separation varied systematically by phoneme category in ways that align with established phonetic knowledge of cleft speech. Alveolar consonants showed the clearest visual separation in PCA space, consistent with extensive clinical and instrumental evidence that anterior sounds are most affected by cleft-related structural constraints.

Cleland et al. [30] documented increased tongue dorsum activity during alveolar production in children with CP±L, reflecting compensatory backing strategies where posterior articulation substitutes for anterior tongue-tip gestures. Our clustering results provide quantitative support for this pattern: the clear horizontal separation along PC1 for alveolar sounds suggests that the model's attention to anterior versus posterior tongue regions differs substantially between groups. For velar consonants, the highest silhouette score (0.465) might initially suggest the strongest diagnostic separation. However, this metric reflects within-cluster cohesion rather than between-cluster distance. Both TD and CP±L speakers produce velars using posterior tongue dorsum raising, inherently limiting the magnitude of spatial differences in tongue positioning. The high silhouette score, therefore, indicates that both groups form tight, consistent clusters for velar production, even though those clusters are spatially closer together than for alveolar sounds.

This interpretation is supported by clinical observations that velar consonants are often relatively preserved in cleft speech, with compensatory patterns more prominent for anterior targets. Palatal consonants showed intermediate characteristics, with moderate silhouette scores and greater vertical spread compared to other categories. This may reflect the articulatory complexity of palato-alveolar sounds, which require coordinated control of both anterior tongue blade and mid-tongue body, potentially introducing higher variability in activation patterns within both diagnostic groups. The phoneme-specific variation in clustering patterns, strongest visual separation for alveolars, intermediate for palatals, and high cohesion but closer proximity for velars, aligns precisely with what would be expected from clinical phonetic knowledge. This convergence between computational clustering results and established articulatory phonetics provides face validity that the model's attention patterns capture genuine articulatory structure rather than arbitrary image statistics or acquisition artefacts. These findings underscore an important tension between numerical performance and anatomical validity. While raw ultrasound inputs maximise classification accuracy, they risk unstable decision-making driven by acquisition artefacts rather than articulatory evidence. In contrast, FoV-standardised ROI and masked representations yield more anatomically grounded, interpretable, and reproducible behaviour, even when numerical accuracy is modestly reduced.

This trade-off has direct implications for clinical deployment. Recent work on XAI in healthcare emphasises that model transparency and clinician trust are prerequisites for adoption. Adeniran et al. [172] discussed how XAI can enhance trust and transparency in critical medical decision-making, while Abgrall et al. [173] examined whether AI models should be explainable to clinicians, concluding that interpretability is essential for clinical integration [173]. A system that achieves 97% accuracy by exploiting probe shadows and FoV geometry is fundamentally unreliable: it will fail when deployed with different equipment, operators, or acquisition protocols. Conversely, a system achieving 87% accuracy based on anatomically plausible tongue-shape features is better positioned for cross-site generalisation and longitudinal stability, as its decisions are grounded in articulatory structure rather than acquisition-specific artefacts, though empirical validation across multiple sites and time points remains necessary to confirm it.

For paediatric speech assessment specifically, where decisions guide treatment planning and parental counselling, the ability to visualise and verify what the model "sees" is essential. Grad-CAM++ visualisations provide this transparency, enabling clinicians to assess whether classifications align with their own visual interpretation of tongue posture. This explainability is particularly valuable for borderline cases or children with mild impairments, where perceptual judgments may be ambiguous. The clustering analysis extends this transparency by providing a quantitative, data-driven characterisation of articulatory patterns that can support objective monitoring of compensatory behaviours and response to therapy.

Several factors limit the strength of claims about what the model has learned and constrain the generalisability of the findings. First, the datasets differ in acquisition parameters, introducing potential confounds that could enable the model to discriminate based on dataset source rather than articulation. While FoV harmonisation was applied to reduce geometric differences, residual systematic variations in image quality, contrast, or operator technique cannot be fully excluded. The finding that raw-input models showed reduced performance after FoV standardisation suggests that these models partially exploited acquisition-related cues, though the phoneme-specific clustering patterns provide counterevidence that genuine articulatory structure is also captured.

Second, the sample size (n=14 speakers per group) is modest and raises the possibility that clustering reflects individual speaker characteristics rather than diagnostic group patterns. With only 28 speakers, speaker-level variance could potentially dominate group-level patterns. However, several findings argue against this interpretation: (i) phoneme-specific separation patterns align with clinical expectations systematically, (ii) statistical tests of Grad-CAM++ features show large, consistent effect sizes (Cohen's $d > 1.5$) across multiple features, and (iii) clustering stability across both k-means and hierarchical methods suggests genuine structure rather than overfitting to individual speakers. Nonetheless, validation on larger, multi-site cohorts where speakers from both diagnostic groups are recorded under identical acquisition conditions remains essential. Third, while the phoneme-specific separation pattern aligns with clinical expectations, this provides only indirect evidence that Grad-CAM++ features reflect true articulatory differences. Direct validation would require correlating Grad-CAM++ spatial features (Centre X, Centre Y) with quantitative tongue shape measurements (e.g., tongue tip height, dorsum curvature) extracted from expert-annotated contours.

Such analysis would confirm whether activation patterns genuinely track anatomical displacement rather than correlated image statistics. The tongue contour validation was performed (Section 4.2.2.1), which established that automated tracking was accurate (mean error 1.15-1.53 pixels), but did not directly correlate tracked positions with Grad-CAM++ features. Fourth, the study focused on binary classification (CP±L vs. TD). Extending the framework to multi-class diagnostic tasks, distinguishing specific compensatory error types (backing, glottal replacement, pharyngeal articulation), severity levels, or predicting treatment response, would better reflect the complexity of clinical decision-making. Such extensions require careful consideration of class imbalance, ordinal relationships between severity categories, and integration with perceptual rating scales used in clinical practice. Fifth, all data were acquired using a single ultrasound system (AAA with matching probe specifications), limiting assessment of generalisability across different equipment. While this consistency strengthens internal validity, variations in probe design, image quality, and system-specific processing algorithms are inevitable in real-world clinical settings and may affect model performance and interpretability.

Finally, the ROI and mask generation processes employed in this study were semi-automated. The use of a fixed threshold may not be optimal across all speakers and acquisition conditions, potentially introducing systematic bias. Fully automated tongue segmentation using DL architectures trained on annotated UTI data would improve scalability, reduce operator dependence, and enable more consistent preprocessing across diverse datasets. Future work should prioritise multi-centre data collection with broader demographic representation, which would also enable evaluation of model robustness across different clinical settings, equipment configurations, and operator expertise levels. Also, the study focused on binary classification. Extending the framework to multi-class diagnostic tasks would better reflect the spectrum of articulatory patterns encountered in clinical populations. Further, real-time integration of explainability tools should be developed to support clinical workflows, enabling verification of model reasoning and identification of potential errors or acquisition issues.

4.5 Summary

This chapter examined how image representation and FoV normalisation influence both the performance and interpretability of DL models applied to UTI for speech disorder diagnosis. Through a systematic comparison of raw, ROI, and masked image representations across both original and FoV standardised datasets, the results showed that while raw images achieve the highest classification accuracy, they often capture irrelevant background features and acquisition-related cues that compromise generalisability and anatomical plausibility. In contrast, models trained on ROI and masked inputs demonstrated more anatomically grounded behaviour, with Grad-CAM++ visualisations revealing focused attention along the tongue surface rather than in peripheral artefacts.

Among these, the masked representation produced the clearest and most clinically interpretable attention patterns. Further analysis of articulatory differences in TD and CP±L speech confirmed that Grad-CAM++ attention maps capture meaningful articulatory behaviour beyond what is reflected by accuracy metrics alone. Statistical comparison and clustering of Grad-CAM++ features revealed clear separation between TD and CP±L groups, alongside greater within-group variability for the cleft cohort, patterns consistent with compensatory articulation strategies and reduced stability of tongue motor control reported in the clinical literature. In practical terms, these explainability-driven features could support clinician-facing tools for visualising and comparing articulatory strategies across patients, enabling objective monitoring of compensatory behaviours and response to therapy rather than relying solely on categorical classification outcomes. Taken together, these findings highlight the importance of integrating representation design, acquisition standardisation, and XAI to enhance both the robustness and transparency of DL systems for clinical speech assessment. Standardising the FoV across datasets improved anatomical consistency and interpretability, providing a stronger foundation for subsequent model development.

Building on these insights, Chapter 5 introduces a generative standardisation framework that automates FoV alignment and tongue-region refinement using a two-stage Pix2Pix model, advancing toward anatomically faithful, robust, and clinically deployable ultrasound-based speech analysis systems. These findings fulfil Objective 2 from Chapter 1 by demonstrating how representation and FoV standardisation affect both performance and interpretability, motivating the automated generative standardisation framework in Chapter 5.

Chapter 5

5. Generative Harmonisation of Ultrasound Tongue Imaging via a Two-Stage Conditional GAN

Building on the representation-level analysis established in Chapter 4, this chapter advances the response to Challenge **C1** data variability and generalisability limitations by introducing a generative standardisation framework based on cGANs. The previous chapter demonstrated that inconsistencies in FoV and image representation produce substantial variability in model accuracy and interpretability, even when lightweight architectures are used. While interpolation-based FoV standardisation achieved 94.53% and 94.45% classification accuracy, respectively, these deterministic geometric transformations remained post-hoc corrections rather than data-driven adaptations capable of learning how heterogeneous acquisitions map into a unified anatomical reference frame. To address this, the current study employs a two-stage generative framework designed to harmonise acquisition differences and enhance tongue visibility in UTIs. In Stage 1, a cGAN is trained to translate UTI frames acquired under varying FoVs into a standardised 97° FoV, reducing geometric variability caused by probe angle and transducer design. In Stage 2, a second cGAN refines the ROI, enhancing the clarity of the tongue surface while suppressing irrelevant background artefacts. Together, these generative stages form a unified preprocessing pipeline that transforms heterogeneous UTI data into anatomically consistent, diagnostically focused images. The central aim is to determine whether generative standardisation can improve both image fidelity and downstream model performance.

Specifically, the chapter (i) quantifies image quality improvements using perceptual metrics such as SSIM and PSNR, (ii) evaluates the contribution of standardised images to downstream classification accuracy, and (iii) assesses cross-domain generalisation by testing whether models trained on real images transfer to GAN-generated counterparts. By embedding generative modelling within the UTI workflow, this chapter moves beyond representation optimisation toward domain standardisation. These claims are made under the assumption that paired reference representations are available and that the learned mappings are constrained by interpolation-derived targets. The methods and findings presented here establish the generative foundation for the subsequent cost-optimisation framework in Chapter 6, where data efficiency and annotation reduction are explored under standardised imaging conditions.

A version of this work has been published as: Al Ani, Saja, Cleland, Joanne and Zoha, Ahmed (2025). Two-Stage GAN for Field-of-View Standardisation and Tongue Region Enhancement in Ultrasound for Cleft Palate Speech Pattern Analysis. In: 6th International Conference on Medical Imaging and Computer-Aided Diagnosis (MICAD 2025), London, UK, 19-21 Nov 2025, (Accepted for Publication).

5.1 Introduction

Acquisition-induced variability in UTI is not unique to speech imaging. Across radiology and clinical ultrasound, domain shifts arising from differences in scanner hardware, operator technique, and imaging protocol have long been identified as barriers to reproducible automated analysis [174], [175]. In UTI specifically, the consequences of this variability are particularly acute: unlike MRI or CT, ultrasound images lack standardised intensity scales, and the placement of a hand-held transducer introduces geometric inconsistencies in FoV and anatomical coverage that vary between sessions and operators. The interpolation-based standardisation evaluated in Chapter 4 reduced geometric variance but was constrained by its deterministic nature; it applied a fixed mathematical rescaling without any capacity to adapt to the local texture, boundary, or intensity patterns that distinguish clinically meaningful articulatory regions from background noise. This motivates a fundamentally different approach: learning the standardisation mapping directly from paired data.

GANs have emerged as powerful tools for modelling complex image distributions and transformations. GAN is a class of DL models that comprises two networks: a generator that synthesises images and a discriminator that learns to distinguish real from generated data [176], [177]. In medical imaging, GAN-based methods have been successfully applied to several clinically relevant problems.

Costa et al.[178] generated realistic retinal images from vessel-tree segmentations, while Dai et al. [179] proposed the Structure Correcting Adversarial Network (SCAN) to segment lung and heart regions from chest X-rays. In neuroimaging, Xue et al. [180] employed dual GANs for brain tumour segmentation, and Nie et al. [181] translated MR images into CT with a patch-based GAN. These studies show that GANs can learn complex mappings between imaging domains while preserving fine anatomical detail.

cGANs have extended this framework by conditioning the generator on an input image, enabling direct pixel-to-pixel translation between domains [182], [183]. This conditional formulation underpins a wide range of image-to-image translation tasks, including modality conversion, denoising, super-resolution, segmentation, and inpainting [184], [185]. Rather than hand-designing preprocessing pipelines, cGANs learn transformations that preserve underlying structure while adapting appearance or geometry. Beyond segmentation and modality translation, cGANs and CycleGAN-based frameworks have been used for data harmonisation and restoration: Ben-Cohen et al. [186] generated CT-like images from positron emission tomography (PET) scans to improve lesion localisation, and Zhao et al. [187] standardised ultrasound radiomics features across acquisition settings for cervical cancer lymph-node prediction. Inpainting-style models such as AnoGAN [188] have reconstructed missing or corrupted regions by learning the manifold of normal anatomy.

More broadly, generative frameworks have been used to support clinical workflows indirectly by improving image interpretability, enhancing rare-class representation, and generating standardised visuals that facilitate collaboration between clinicians and AI systems [189], [190]. These properties are highly relevant to UTI. Consistent and interpretable ultrasound images can enhance automated analysis, support the visual feedback used in therapy sessions, and make DL decisions easier to inspect and trust. Despite this progress in CT, MRI, and conventional ultrasound, the application of GAN-based harmonisation to speech-specific UTI remains underdeveloped.

The closest prior work, deterministic cropping and masking pipelines [61], [191], [192], addresses spatial inconsistency through fixed geometric operations rather than learned mappings, and no published study has applied paired cGAN translation to standardise articulatory ultrasound geometry or to jointly harmonise FoV and refine the tongue ROI within a unified framework. This gap motivates the two-stage Pix2Pix approach introduced in this chapter.

Existing UTI preprocessing methods typically tackle spatial variability or noise reduction in isolation, relying on deterministic geometric operations rather than data-driven mappings learned from paired examples. This limits adaptability across probes, operators, and sites, and places a heavy burden on manual preprocessing. Furthermore, prior work has not combined geometric standardisation with targeted enhancement of the tongue region within a unified framework tailored to articulatory analysis.

This chapter addresses these gaps by introducing a two-stage Pix2Pix-based cGAN pipeline for UTI standardisation. Building on the interpolation baseline established in Chapter 4, this work investigates whether learned generative standardisation can surpass deterministic geometric methods. Stage 1 performs FoV translation, converting ultrasound frames captured at varying angular extents into a standardised 97° reference geometry that aligns anatomical coverage across datasets. Stage 2 refines the ROI, isolating and enhancing the tongue surface while suppressing background structures and probe artefacts. Together, these stages form a generative preprocessing framework that standardises both imaging geometry and anatomical focus, yielding spatially consistent and diagnostically meaningful inputs for downstream models.

The central hypothesis is that generative standardisation can reduce acquisition-induced variability and improve both the fidelity and interpretability of UTI data beyond interpolation-only methods. Quantitatively, the framework is evaluated using perceptual similarity metrics and by measuring the impact of generated images on phonetic-classification accuracy. In doing so, this chapter contributes directly to Challenge **C1** data variability and generalisability limitations by replacing hand-engineered standardisation with a learned solution, and advances R3 generative standardisation via a two-stage Pix2Pix framework. The resulting representations provide a consistent and interpretable foundation for the cost-efficiency and deployment frameworks developed in Chapters 6 and 7.

5.1.1 Contributions

This chapter makes two main contributions:

1. This chapter introduces a paired image-to-image translation framework based on Pix2Pix cGANs to harmonise FoV and refine the tongue ROI. The proposed two-stage generative preprocessing strategy is tailored specifically to articulatory UTI and advances beyond the interpolation-based FoV standardisation methods evaluated in Chapter 4 by learning data-driven mappings between heterogeneous acquisition geometries and a unified anatomical reference.

   Impact on C1: This contribution directly addresses acquisition-induced variability by enforcing geometric and representational consistency across ultrasound recordings, improving cross-speaker and cross-session generalisability. By replacing fixed rescaling with learned harmonisation, the framework mitigates domain shifts caused by probe placement and FoV differences.
2. Establishing diagnostic and cross-domain improvements through empirical validation. This chapter demonstrates that the proposed two-stage Pix2Pix framework enhances downstream classification accuracy and robustness beyond deterministic preprocessing baselines. Stage 1 improves stability and performance by standardising imaging geometry across acquisition protocols, while Stage 2 reinforces anatomical salience by producing tongue-centred representations that suppress irrelevant background content.

   Impact on: C1: FoV harmonisation improves cross-domain robustness and enables consistent performance across heterogeneous datasets. **C2:** The ability to mix real and synthetic standardised images without performance degradation demonstrates the potential of generative preprocessing to stabilise learning in data-limited settings. **C3:** ROI refinement encourages attention to linguistically meaningful articulatory regions, improving interpretability and sensitivity to tongue-shape variation relevant to SSDs.

The remainder of this chapter is organised as follows. Section 5.2 describes the method, including dataset preparation, Pix2Pix architecture design, and evaluation metrics. Section 5.3 presents quantitative and qualitative results for both generative stages, while Section 5.4 discusses their implications for diagnostic reliability and clinical integration, including explicit comparison with the interpolation baselines from Chapter 4. Section 5.5 provides a summary of the chapter, integrating a discussion of limitations and directions for future work, and outlines how this generative framework supports the cost-efficiency and deployment strategies developed in Chapters 6 and 7.

5.2 Method

This chapter introduces a two-stage generative framework based on cGAN, specifically the Pix2Pix model, to enhance spatial and anatomical consistency in UTI data. The proposed two-stage Pix2Pix pipeline standardises UTI data through supervised image-to-image translation. In Stage 1, the model learns to map raw UTI frames with varying FoVs to a standardised 97° reference geometry using paired training data constructed from bicubic interpolation. In Stage 2, the model learns to refine the ROI by mapping FoV-standardised images to tongue-focused representations. The target images for Stage 2 were the ROI images created using the preprocessing pipeline described in Chapter 4, Section 4.2.2. This sequential approach addresses two major sources of variability that limit automated SSD diagnosis: differences in FoV across ultrasound acquisitions and inconsistencies in ROI representation. Figure 5.1 illustrates the overall design of the proposed framework. The two generative stages are applied sequentially, with Stage 1 performing FoV standardisation and Stage 2 refining the anatomical focus by generating tongue-centred representations that suppress background artefacts and probe noise. The resulting standardised and tongue-centred images form high-quality inputs for subsequent classification tasks.

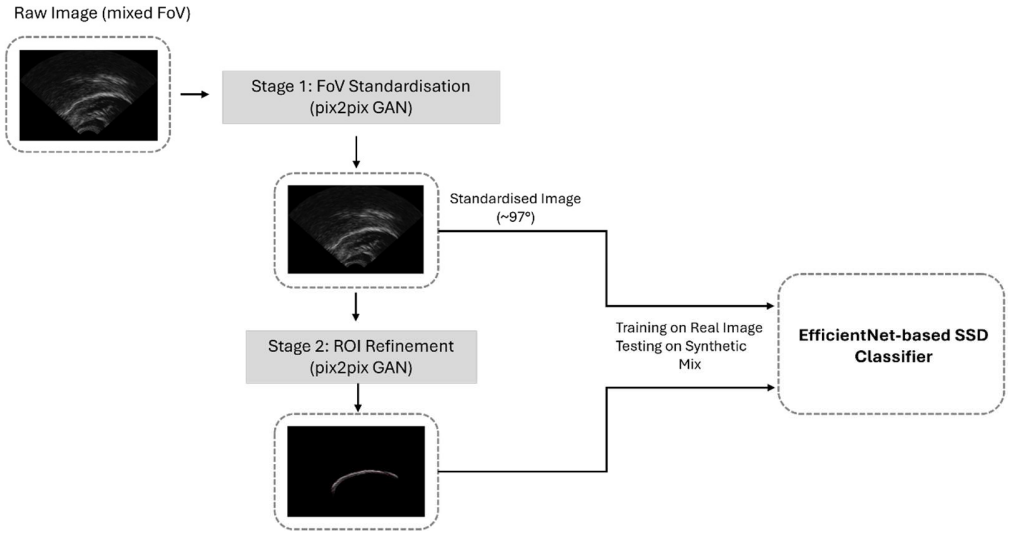

Figure 5. 1: Overview of the proposed system, comprising the two-stage Pix2Pix generative preprocessing pipeline (Stage 1: FoV standardisation; Stage 2: ROI refinement) followed by the EfficientNet-B0 downstream classifier. The generative stages operate at the image level to standardise inputs; the classifier operates on the standardised outputs to perform binary SSD classification.

By generating spatially consistent and anatomically relevant UTI frames, the framework supports the training of more robust and generalisable downstream models for SSD classification. The same Pix2Pix architecture was employed in both stages to maintain design consistency and minimise implementation complexity. Each stage was trained, validated, and evaluated independently using SSIM, PSNR, and qualitative perceptual analysis. In addition, the synthesised outputs were assessed through a binary classification task and cross-domain generalisation experiments comparing performance on real versus generated images.

5.2.1 Datasets

The datasets used in this chapter are identical to those described in Chapter 4, Section 4.2.1, and are summarised here for completeness. The combined dataset comprised 2,660 mid-sagittal UTI frames from two participant groups: children with CP±L and TD children. The CP±L group contributed 1,400 frames across 14 speakers, and the TD group contributed 1,260 frames across 14 speakers. All frames were captured during the production of consonant targets spanning alveolar, palatal, and velar places of articulation. To ensure that classification results reflect genuine speaker-generalised patterns rather than artefacts of identity overlap, data partitioning was performed at the speaker level: all frames from any given speaker were assigned exclusively to training, validation, or test subsets, with no speaker appearing in more than one partition. The 70/15/15 split was applied after this speaker-level stratification. For generative training, paired images were constructed in the (A, B) format required by Pix2Pix, where A represents the raw wide-FoV image, and B represents the corresponding target, either the 97°-standardised version (Stage 1) or the ROI-refined frame (Stage 2). All data collection and handling procedures complied with institutional ethical standards and data-sharing agreements, as described in Section 4.2.1.

5.2.2 Data Preprocessing

Before training both the generative and classification models, all UTI frames underwent a structured preprocessing pipeline designed to enhance visual quality, suppress artefacts, and ensure compatibility with DL architectures. Each grayscale ultrasound image was first processed using Contrast-Limited Adaptive Histogram Equalisation (CLAHE; clip limit = 2.0, tile grid size = 8×8 pixels) to enhance local contrast and improve the visibility of the tongue surface and surrounding tissue boundaries, without over-amplifying noise in homogeneous regions. To mitigate speckle noise inherent to ultrasound acquisition, a Gaussian blur filter (kernel size = 3×3, $\sigma = 1.0$) was subsequently applied, smoothing high-frequency noise while preserving key structural features such as tongue contour and dorsal surface curvature. This preprocessing ensured that both generative and discriminative models operated on images with improved contrast, reduced noise, and anatomically consistent representation, providing a stable foundation for the standardisation and classification experiments that follow.

5.2.3 Pix2Pix cGAN Architecture

The Pix2Pix framework was selected for this study due to its proven ability to translate input images from one domain to another while preserving spatial and structural integrity [193]. UTI frames vary substantially in FoV, anatomical coverage, probe positioning, and background artefacts. These differences create a domain-shift problem that cannot be fully solved by changing the classifier alone. Pix2Pix is appropriate because it learns a paired image-to-image mapping from an input ultrasound representation to a desired target representation while preserving spatial correspondence between the two. This makes it well-suited to FoV harmonisation, where the output must remain anatomically faithful to the original tongue image while conforming to a standardised 97° view. It is also appropriate for ROI refinement, where the model must suppress irrelevant background tissue and acoustic clutter while preserving the tongue surface and diagnostically relevant articulatory structure.

The proposed Pix2Pix cGAN adopts a conditional U-Net generator and a PatchGAN discriminator, following the standard Pix2Pix configuration. The generator consists of an encoder–decoder network with skip connections between corresponding layers. The encoder progressively down-samples the input image through convolutional layers, capturing multi-scale geometric and textural features essential for the translation task. The decoder then up-samples these latent representations using transposed convolutions to reconstruct an output image aligned with the target domain. The skip connections directly transfer fine-grained spatial information from the encoder to the decoder, preserving low-level structural details, such as tongue boundaries and surface curvature, during generation. The discriminator employs a PatchGAN architecture that classifies image authenticity at the patch level rather than the entire image. A standard GAN employs a discriminator that outputs a single value indicating the likelihood that an image is real. While this is adequate for simple generation tasks, it is too coarse for image-to-image translation, where spatial structure matters. A single probability cannot capture whether different regions of the image look realistic. PatchGAN reframes the discriminator's role by producing a grid of probabilities rather than a single global score. Each value in the grid reflects the realism of a small spatial region or patch of the image. The discriminator, therefore, judges many local areas independently, giving the generator fine-grained, spatially aware feedback that supports more detailed and coherent image translation.

This design encourages the generator to produce high-frequency, locally coherent details by penalising unrealistic textures within each image patch. Through this adversarial dynamic, the generator learns to synthesise anatomically realistic and spatially consistent outputs. Figure 5.2 illustrates the structure of the proposed Pix2Pix cGAN model. The generator receives a raw UTI frame as input and produces a standardised or ROI-refined version, conditioned on the paired target. The discriminator jointly evaluates the input-output pair to determine whether the generated image is realistic. This conditioning mechanism enables the network to generate outputs that are contextually appropriate and structurally aligned with the source anatomy.

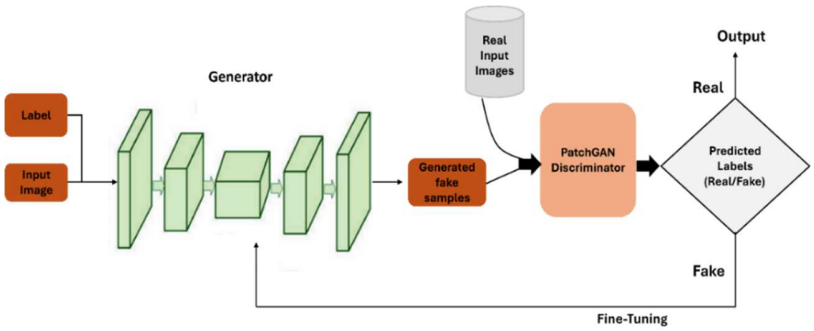

Figure 5. 2: Structure of the Proposed cGAN (Pix2Pix).

5.2.3.1 Model Training

The generative model G in the Pix2Pix framework aims to produce an output image y conditioned on an input image x. Formally, this process can be expressed as:

$$G: x \rightarrow y \qquad (5.1)$$

where x represents the input ultrasound frame (e.g., wide-FoV image), and y denotes the generated target image (e.g., standardised FoV or refined ROI). Unlike some GAN variants, standard Pix2Pix does not rely on explicit noise injection; deterministic mapping from input to output is learned directly through paired supervision. The discriminative model D operates concurrently, attempting to distinguish between real target images and those produced by G. Through this adversarial interplay, the two networks continually refine one another: the

generator learns to create increasingly realistic, anatomically consistent outputs, while the discriminator becomes more adept at identifying synthetic images. This competitive learning process allows the system to capture complex spatial and textural mappings between domains, ensuring that the generated ultrasound frames preserve articulatory structures while standardising imaging geometry or focus.

In a cGAN framework, the generator G and discriminator D are optimised in opposing directions. The discriminator seeks to maximise the adversarial loss by correctly identifying real versus generated images. At the same time, the generator attempts to minimise the same loss by producing outputs that are increasingly indistinguishable from real targets. This adversarial interaction is typically expressed as a minimax objective:

$$G^* = \arg \min_G \max_D \; L_{\text{cGAN}}(G, D) \qquad (5.2)$$

where L_{cGAN} denotes the conditional adversarial loss, defined over paired (input, target) examples. However, previous research [104] has shown that combining the adversarial loss with a pixel-wise reconstruction loss, typically an $\ell_2$or $\ell_1$term, reduces blurriness and improves structural fidelity in the generated outputs. In Pix2Pix, the conditional adversarial loss is therefore paired with an $\ell_1$reconstruction loss to encourage the generator to produce images that are not only perceptually realistic but also anatomically aligned with the ground-truth targets. The full objective becomes:

$$L_{\text{total}} = L_{\text{cGAN}} + \lambda L_{\text{L1}} \qquad (5.3)$$

where λ controls the relative weighting of pixel-level reconstruction relative to adversarial realism. Following the original Pix2Pix formulation [193], λ was set to 100, prioritising structural fidelity over purely adversarial sharpness, a choice appropriate for the paired, anatomy-preserving objectives of both generative stages. Table 5.1 summarises the hyperparameters for the Two-Stage Pix2Pix cGAN Pipeline.

Table 5. 1: Summary of Hyperparameters for the Two-Stage Pix2Pix cGAN Pipeline.

Model / Stage	Input / Output	Architecture settings	Optimiser	Learning rate	Batch size	Epochs	Loss function
Pix2Pix Stage 1: FoV standardisation	Input: mixed/wide FoV UTI; target: 97° standardised FoV UTI	U-Net generator with skip connections; PatchGAN discriminator	Adam	0.0002	1	200	Conditional adversarial loss + L1
Pix2Pix Stage 2: ROI refinement	Input: FoV-standardised UTI; target: ROI-refined UTI	Same U-Net generator and PatchGAN discriminator as Stage 1	Adam	0.0002	1	200	Conditional adversarial loss + L1 reconstruction loss

Generator outputs were saved after each epoch to support qualitative inspection and selection of the best-performing checkpoint for downstream evaluation. This combined training objective ensures that the generated images are both perceptually convincing and anatomically accurate, capturing the detailed spatial features required for reliable downstream analysis. Qualitative assessments involved visual comparison between real and synthetic images, while quantitative evaluation employed classification metrics and image-fidelity scores, as described in later sections.

5.2.4 Generative Image Standardisation Pipeline

A generative adversarial approach was adopted to address variability in UTI data, specifically targeting inconsistencies in FoV and ROI localisation. The proposed two-stage Pix2Pix pipeline was designed to progressively harmonise image geometry and enhance anatomical focus. In Stage 1, the model performed FoV standardisation, mapping ultrasound frames acquired under mixed geometries to a unified 97° reference view. In Stage 2, the same architecture was reused for ROI refinement, learning to highlight the tongue region while suppressing background artefacts. Both stages employed identical network configurations, loss functions, and optimisation settings to ensure architectural consistency and comparability across experiments.

5.2.4.1 Stage 1: Field-of-View Standardisation

The first stage of the pipeline focused on standardising spatial geometry across datasets. The Pix2Pix model was trained to translate UTI with varying FoVs (domain A) into a consistent, standardised view (domain B). The paired training data were constructed by aligning each original image (real_A) with its corresponding rescaled version (real_B) produced via bicubic interpolation to a 97° angular extent. This pairing provided explicit supervision, allowing the generator to learn both spatial alignment and intensity mapping between heterogeneous acquisition settings. Once trained, the generator was deployed to convert all raw UTIs into standardised synthetic counterparts (fake_B), each conforming to the same geometric reference. The resulting images were inspected visually to ensure anatomical plausibility and spatial uniformity across speakers. The standardised dataset was then used in downstream classification experiments to evaluate the impact of spatial consistency on diagnostic performance.

Unlike traditional interpolation-based rescaling, which applies a fixed mathematical transformation, the Stage 1 Pix2Pix model learns a data-driven approximation of the mapping from wide-FoV images to the 97° reference geometry. Although the paired targets were created using bicubic interpolation, the generator does not merely reproduce this operation; rather, the adversarial and reconstruction losses enable it to capture subtle, non-linear geometric and textural corrections that may better preserve articulatory detail, particularly around the tongue dorsum and root. As a result, the Stage 1 generator aims to provide a standardised and anatomically consistent foundation for inter-speaker comparability and for the subsequent ROI refinement stage, with performance evaluated against the interpolation baseline established in Chapter 4.

5.2.4.2 Stage 2: Region-of-Interest Refinement

Following FoV standardisation, the second stage of the pipeline aimed to refine the ROI, ensuring that subsequent models focus on articulatory structures most relevant for phonetic and diagnostic interpretation. UTI frames often include substantial background noise, such as the submental region, hyoid shadow, and speckle artefacts, that can mislead classifiers and reduce generalisation. ROI refinement mitigates this by enhancing the spatial salience of clinically meaningful regions while suppressing irrelevant visual information. In this stage, the same Pix2Pix configuration was retained, but the training task was redefined.

The inputs (domain A) were the standardised images from Stage 1, while the targets (domain B) were the ROI images created using the preprocessing methodology described in Chapter 4. The pairing of FoV-standardised inputs with ROI-focused targets enabled the generator to learn a mapping from full-frame ultrasound images to refined, tongue-centred representations.

After training, the generator was applied to the entire standardised dataset, producing a refined synthetic set (fake_B_ROI) characterised by clear articulatory contours and reduced background interference. This refined dataset was subsequently used in classification experiments to assess whether tongue-focused representations improved model discriminability, particularly under data-limited conditions. Beyond improving visual clarity, ROI refinement functions as a soft attention mechanism, guiding downstream networks to concentrate on the most diagnostically relevant anatomical regions.

Together, the FoV standardisation and ROI refinement stages form a cohesive generative pipeline that produces spatially consistent, interpretable, and clinically meaningful UTI representations for automated speech disorder analysis.

5.2.5 Evaluation of Generated Images

A comprehensive evaluation framework was implemented to assess the quality, diagnostic utility, and generalisability of the generated ultrasound images. The assessment combined quantitative image-quality metrics, classification performance measures, and cross-domain generalisation tests, ensuring that the proposed generative framework was evaluated both perceptually and functionally. Table 5.2 summarises the key metrics used in this study.

Table 5.2: Summary of Evaluation Metrics used to Assess Generated Image Quality, Classification Performance, and Generalisation.

Category	Metric	Purpose / Description
Image Quality	SSIM	Measures structural similarity between generated and target images.
	PSNR	Measures how far the generated image pixels are from the ground truth.
Classification Performance	Accuracy	Proportion of correctly predicted samples.
	Precision	Ratio of true positives to all predicted positives.
	Recall	Ratio of true positives to the total number of actual positives.
	F1 Score	Harmonic mean of precision and recall, reflecting overall balance.
	Average Inference Time	Mean processing time per image, indicating deployment feasibility.
	95% CI	A range of values that expresses the degree of uncertainty about an estimate.
Generalisation Testing	Train: real_B → Test: fake_B	Evaluates the capacity of models trained on real data to generalise to synthetic images.
	Train: fake_B → Test: real_B	Tests whether GAN-generated images can substitute for real data during training.

5.2.5.1 Image Quality Assessment

Quantitative assessment of image quality was conducted using SSIM and PSNR, comparing each generated image (fake_B) to its corresponding ground-truth target (real_B). SSIM quantifies perceptual similarity in terms of luminance, contrast, and structural alignment, providing an indicator of anatomical fidelity. PSNR complements this by measuring pixel-level reconstruction accuracy, where higher values reflect lower distortion [194]. Together, these metrics provide an objective measure of how well the generative model preserves fine articulatory detail while reducing acquisition-induced variability.

5.2.5.2 Classification Performance

To evaluate the diagnostic relevance of the generated datasets, a binary classification experiment was conducted using the EfficientNet-B0 architecture. The goal was to distinguish between UTIs of children with CP ± L and TD children. Evaluation was performed separately for two generative outputs:

1. FoV-standardised images and
2. ROI-refined images.

EfficientNet-B0 was selected for its balance between performance and parameter efficiency on small medical imaging datasets. The model's lower convolutional layers were partially frozen to retain pretrained visual representations, while deeper layers were fine-tuned to adapt to UTI-specific features. The data were stratified and split into 70% training, 15% validation, and 15% testing subsets. To mitigate class imbalance, a weighted random sampler was applied during training. Optimisation was performed using Adam with a ReduceLROnPlateau scheduler (patience = 5 epochs, reduction factor = 0.1) to stabilise convergence by reducing the learning rate when validation loss ceased to improve. Classification performance was reported using accuracy, precision, recall, F1-score, and CIs. Additionally, the average inference time per image was measured to assess computational feasibility for future real-time deployment.

5.2.5.3 Generalisation Testing

Cross-domain generalisation was assessed to determine whether synthetic data could substitute or complement real data in model training. Two primary experiments were conducted for each generative stage:

1. Stage 1 (FoV standardisation): Train on real_B, test on fake_B.
2. Stage 2 (ROI refinement): Train on real_B, test on fake_B_ROI.

To evaluate the robustness of synthetic data, additional experiments compared single-source training (real-only or synthetic-only) with mixed training (Real and Synthetic) conditions. Performance stability across these configurations served as an indicator of how well the generative process captured domain-invariant articulatory features.

Through these evaluations, the two-stage generative preprocessing framework was shown to mitigate key challenges in ultrasound-based SSD diagnosis, namely, anatomical variability, low contrast, and data scarcity, by producing standardised, tongue-centred representations that improved interpretability, generalisability, and computational efficiency across diagnostic tasks.

5.3 Results

5.3.1 Image quality

The proposed two-stage Pix2Pix pipeline produced high image fidelity across both generative stages, although the interpretation of the similarity metrics differs between Stage 1 and Stage 2 because the two stages perform different transformation tasks. Quantitative image-quality results are summarised in Table 5.3. These metrics reflect per-image variability across the test set from a training run over 200 epochs. It should be noted that image-quality metrics were computed from a single training run; while the results were consistent with visual inspection across saved epoch checkpoints, formal reproducibility across independent runs remains a direction for future work, as discussed in Section 5.5.

In Stage 1, which performs FoV standardisation, the generated outputs achieved a mean SSIM of 0.96 ± 0.02 and a PSNR of 88.50 ± 3.12 dB. The high SSIM indicates that the generated images retained the structural organisation of the target frames, including the fan-shaped ultrasound geometry, tongue contour, and surrounding anatomical context. The very high PSNR indicates extremely low pixel-level error between the generated and target images. This result is expected because the Stage 1 target images were derived from the corresponding inputs using bicubic interpolation to create a standardised 97° FoV. The task is therefore primarily a geometric harmonisation problem rather than a fully unconstrained image-generation problem. In other words, the generator is learning a deterministic spatial transformation in which the anatomical content of the input and target remains closely aligned. The high PSNR should therefore be interpreted as evidence that the model accurately reproduced the standardised FoV mapping with minimal pixel-level distortion, rather than as evidence of broad generative realism across unrelated domains.

The Stage 1 results are important because FoV variation is a major source of acquisition-related domain shift in UTI. A high SSIM in this stage suggests that the model preserves the clinically relevant tongue structure while reducing geometric inconsistency across recordings. This is essential for downstream classification, where differences in image scale, angular coverage, and anatomical visibility could otherwise be mistaken for diagnostic or phonetic variation. The U-Net skip connections likely contributed to this preservation by allowing low-level spatial information to pass directly from encoder to decoder, maintaining fine articulatory detail while the network learned the required geometric standardisation.

In Stage 2, which performs ROI refinement, the generated outputs achieved a mean SSIM of 0.91 ± 0.03 and a PSNR of 33.26 ± 2.45 dB. These values are lower than those observed in Stage 1, but this reduction is expected and reflects the more transformative nature of the ROI-refinement task. Unlike Stage 1, Stage 2 is not intended to reproduce the full ultrasound frame as closely as possible. Instead, its purpose is to suppress background regions, reduce irrelevant acoustic artefacts, and emphasise the tongue region. As a result, pixels corresponding to background tissue, speckle noise, and non-diagnostic regions are intentionally modified or removed. This inevitably reduces pixel-level similarity and therefore lowers PSNR, even when the transformation is clinically desirable.

The lower Stage 2 SSIM should therefore be interpreted critically rather than simply as reduced image quality. SSIM measures global structural similarity between the generated image and the target, but ROI refinement deliberately changes the structural composition of the image by focusing attention on the tongue region and reducing the surrounding context. A moderate reduction in SSIM is consequently consistent with the intended function of the model. Importantly, the Stage 2 SSIM remains high enough to indicate preservation of the main tongue structure, while the lower PSNR reflects the deliberate suppression of background pixels rather than a failure to reconstruct the articulatory region. Visual inspection confirmed that the refined outputs retained clear tongue boundaries and improved contrast between the articulatory surface and surrounding tissue.

The difference between Stage 1 and Stage 2, therefore, reflects the different objectives of the two generative stages. Stage 1 prioritises geometric consistency and minimal distortion, so very high SSIM and PSNR are desirable and expected. Stage 2 prioritises anatomical focus and background suppression, so a lower PSNR and slightly lower SSIM are acceptable because the model is intentionally altering non-essential image regions. This distinction is important because standard image-similarity metrics can penalise clinically useful transformations when those transformations involve removing irrelevant content. For this reason, the Stage 2 metrics should be interpreted alongside visual inspection and downstream classification performance rather than as standalone indicators of quality.

Table 5.3: Image Quality Assessment of Generated Outputs Across both Pix2Pix Stages.

Stage	Metric	Mean ± SD
Stage 1	SSIM	0.96 ± 0.02
	PSNR (dB)	88.50 ± 3.12
Stage 2	SSIM	0.91 ± 0.03
	PSNR (dB)	33.26 ± 2.45

Overall, these results demonstrate that the two-stage cGAN framework performs two complementary functions. Stage 1 provides spatially standardised images with minimal loss of anatomical fidelity, reducing FoV-related variability across acquisitions. Stage 2 produces more tongue-centred representations that improve anatomical focus and interpretability, even though this necessarily reduces global pixel-level similarity. Together, the two stages generate standardised and anatomically meaningful UTI inputs suitable for downstream classification and clinical analysis.

5.3.2 Classification Performance Analysis

The classification performance of the proposed two-stage generative pipeline was evaluated across both FoV-standardised (Stage 1) and ROI-refined (Stage 2) datasets. Quantitative results are summarised in Table 5.4.

In Stage 1, where images were standardised to a uniform FoV, the classifier achieved outstanding performance. When trained on a mixed dataset combining real and synthesised (Fake_B) images, the model reached a maximum accuracy of 98.80%. Even when trained on single-source data, both Real_B and Fake_B models performed comparably well, achieving accuracies of 95.60% and 94.51%, respectively. These consistently high values indicate that FoV harmonisation substantially improved the learnability of articulatory patterns by eliminating scale and context discrepancies across images. The negligible performance gap between real and generated data also suggests that the synthetic images retained sufficient anatomical realism to serve as effective training material, confirming the reliability of the Stage 1 generator. It is notable that the mixed training condition produced perfect precision, indicating that all positive predictions were correct. This reflects the classifier's conservative threshold behaviour on this test set: with augmented training data providing a broader coverage of the feature space, the model achieved a very high decision boundary for positive labelling, eliminating false positives. Whether this generalises beyond the current test set would require evaluation on held-out multi-centre data and is therefore appropriately treated as a finding specific to this experimental configuration rather than a general performance guarantee.

In Stage 2, classification accuracy decreased slightly across all datasets, reflecting the more challenging nature of prediction from ROI-focused inputs where background context is reduced. Nevertheless, the mixed training configuration again yielded the strongest results, achieving an accuracy of 91.21%. Notably, classifiers trained solely on synthetic ROI-refined data (Fake_B) slightly outperformed those trained on real data (Real_B), reaching 89.58% versus 85.42% accuracy. This improvement suggests that the generative ROI refinement process may have enhanced tongue visibility and reduced noise, thereby enabling the model to focus more effectively on discriminative articulatory features. To quantify the reliability of the observed classification improvements, 95% CIs were computed. For Stage 1, the mixed training condition achieved 98.80% accuracy [97.7%, 99.9%], with non-overlapping intervals compared to Real_B alone at 95.60% [93.6%, 97.6%], confirming that incorporating GAN-generated FoV-

standardised images provides a statistically reliable performance gain. For Stage 2, the Mix condition [88.4%, 94.0%] does not overlap with Real_B [81.9%, 88.9%], supporting the conclusion that synthetic ROI-refined data contributes genuine diagnostic value beyond real data alone.

Table 5.4: Classification Performance of the Proposed Two-Stage Generative Pipeline.

Stage	Dataset	Accuracy (%)	F1 Score (%)	Precision (%)	Recall (%)	95% CI
Stage 1– FoV Standardisation	Real_B	95.60	95.83	93.88	97.87	93.6,97.6
	Fake_B	94.51	94.74	91.84	97.83	92.3,96.7
	Mix	98.80	97.47	100.00	98.72	97.7,99.9
Stage 2– ROI Refinement	Real_B	85.42	85.71	87.50	84.00	81.9,88.9
	Fake_B	89.58	89.80	91.67	88.00	86.6,92.6
	Mix	91.21	91.67	86.27	97.78	88.4,94.0

Collectively, these results demonstrate complementary strengths across the two stages. Stage 1 FoV standardisation maximised classification accuracy and precision by ensuring geometric consistency and reducing inter-speaker variability, while Stage 2 ROI refinement improved sensitivity and recall, particularly beneficial for detecting positive (CP ± L) cases where subtle articulatory cues are critical. The modest reduction in global accuracy following ROI masking reflects the trade-off between spatial context and targeted anatomical focus. Overall, the findings confirm that the generative harmonisation pipeline enhances both data quality and diagnostic reliability across heterogeneous ultrasound datasets.

5.3.3 Generalisation Testing

Model generalisation and robustness were evaluated through cross-domain experiments, in which classifiers trained on real images were tested on GAN-generated images at the same stage. The results of this evaluation are summarised in Table 5.5. In Stage 1 (FoV standardisation), cross-domain generalisation was virtually seamless. A classifier trained on real, standardised images achieved 98.68% accuracy when tested on generated (Fake_B) images, maintaining equally high F1-score (98.69%) and precision (99.21%).

These results confirm that the synthetic images generated by the Pix2Pix model were almost indistinguishable from their real counterparts in terms of the discriminative features used by the classifier. The GAN successfully preserved the distribution of articulatory structures while harmonising geometry, producing realistic, domain-stable outputs. This finding demonstrates that FoV standardisation does not introduce a measurable domain shift and that real and synthetic FoV-standardised images can be used interchangeably in training or diagnostic pipelines without degrading performance.

In Stage 2 (ROI refinement), the domain gap was more pronounced. A model trained on real ROI images achieved 80.82% accuracy when tested on GAN-generated ROI images, with a relatively high recall (93.41%) but a lower precision (75.73%), indicating an increased false-positive rate on synthetic inputs. These results suggest that while the refined synthetic images closely resemble real ROIs qualitatively, subtle discrepancies in texture and boundary sharpness can mislead classifiers trained exclusively on real data. The observed performance decline suggests that ROI refinement causes a slight distribution shift, which the existing generative model fails to fully address.

Importantly, incorporating synthetic data during training substantially alleviated this gap. Mixed training (real + synthetic) improved generalisation and stabilised recall, confirming the value of data augmentation through generative synthesis. This highlights a practical benefit of the two-stage pipeline: synthetic data can effectively supplement real UTI frames to increase robustness in small or imbalanced datasets. Another noteworthy outcome concerns computational efficiency. Inference on ROI-refined images was approximately twice as fast, averaging 0.065 s per image compared to 0.135 s for full FoV inputs. The improvement reflects the reduced image area and simplified content after background suppression, allowing the classifier to process fewer, more relevant features. Although the total pipeline introduces additional preprocessing time, the ROI-focused representations offer tangible benefits for real-time diagnostic workflows by enabling faster downstream analysis.

Table 5.5: Stage 1 Cross-Domain Generalisation Results.

Stage	Training Set	Testing Set	Accuracy (%)	F1 Score (%)	Precision (%)	Recall (%)	Inference Time (s)	95% CI
Stage 1 FoV Standardisation	Real_B	Fake_B	98.68	98.69	99.21	97.41	0.1354	[97.6,99.8]
Stage 2 ROI Refinement	Real_B	Fake_B	80.82	83.65	75.73	93.41	0.0652	[76.9,84.7]

5.4 Discussion

The experimental results validate the effectiveness of the proposed two-stage Pix2Pix framework and provide insights into image fidelity, model generalisation, and clinical utility. The near-perfect SSIM and PSNR values in Stage 1 confirm that FoV standardisation preserved anatomical detail with virtually no distortion, an exceptional outcome in medical image translation, where even state-of-the-art models often produce lower similarity scores [195].

This near-lossless fidelity indicates that anatomical content was preserved almost pixel-for-pixel, largely due to the architectural and training design. The U-Net generator's skip connections carried low-level spatial detail directly from input to output, while the combined L1 and adversarial losses encouraged minimal unnecessary alteration. Functionally, this learned transformation acts as a form of data-driven calibration, aligning images from diverse acquisition protocols into a shared anatomical frame of reference. Similar harmonisation has been shown to improve reproducibility across imaging sites in radiology [196]; here, it enabled classifiers to focus on pathological versus typical articulations rather than acquisition differences, yielding a marked improvement in accuracy and generalisability. This evaluates whether the learned generative standardisation provides tangible benefits beyond deterministic geometric transformations. To address this, we compare the GAN-based approach directly with the interpolation baselines established in Chapter 4, as summarised in Table 5.6.

Table 5.6: Comparison of FoV Standardisation Approaches.

Approach	Method	Accuracy (%)	Key Characteristics
None (Chapter 4)	Mixed Fov	97.44	High variability, domain shift
Interpolation	Bicubic	94.53	Deterministic, geometric only
Interpolation	Spline	94.45	Deterministic, geometric only
GAN (real only)	Pix2Pix Stage 1	95.60	Learned texture/boundary refinement
GAN (mixed)	Pix2Pix Stage 1	98.80	Learned refinement + data augmentation

The 97.44% accuracy achieved under the mixed-FoV condition (no standardisation, Chapter 4) appears counterintuitive at first, since it exceeds both interpolation baselines and the GAN real-only result. This finding, discussed in detail in Chapter 4, reflects a specific experimental context rather than a genuine advantage of heterogeneity: the mixed-FoV model was trained and tested on data from the same combined acquisition pool, meaning that FoV variation was present in both training and test distributions. Under this matched-distribution setting, the classifier implicitly learned to handle geometric variance as part of its feature representation. In contrast, the interpolation and GAN experiments enforced strict standardisation, removing the FoV cue entirely, which requires the classifier to rely more exclusively on tongue-shape features. The improvement from GAN mixed training to 98.80%, therefore, does not simply recover the advantage of seeing varied FoVs; it demonstrates that combining learned standardisation with data augmentation provides a more generalisable representation than either heterogeneity or deterministic rescaling alone, particularly under cross-domain evaluation.

The comparison reveals GAN-based standardisation trained on real data alone marginally outperforms bicubic interpolation by approximately 1.1pp. While modest, this improvement is consistent and reflects the GAN's capacity to learn subtle textural and boundary corrections beyond pure geometric rescaling. The adversarial training objective enables the generator to adapt intensity patterns and edge characteristics in ways that better preserve articulatory salience, particularly around the tongue dorsum and root, where interpolation methods may introduce smoothing artefacts. Further, mixed training combining real and synthetic data achieved a representation of 4.3pp improvement over bicubic interpolation.

This finding demonstrates that the GAN's primary contribution extends beyond standardisation quality alone; it enables effective data augmentation by generating diverse, anatomically plausible training examples. The mixed training regime exposed the classifier to a broader distribution of tongue shapes and imaging variations, improving robustness and reducing overfitting. This aligns with findings from broader medical imaging research, where GAN-generated synthetic data has been shown to enhance model generalisation in data-limited settings [197], [198].

The performance contrast between Stage 1 and Stage 2 highlights important design trade-offs. Stage 1 produced a major leap in classification accuracy. In contrast, Stage 2 was introduced to test whether a more focused input would enhance the detection of subtle tongue-motion differences relevant to SSD. Indeed, Stage 2 improved sensitivity for positive cases: by removing background structures and concentrating exclusively on the tongue, the classifier achieved a recall approaching 98%, indicating that nearly all SSD instances were correctly identified. Clinically, this is valuable in contexts where ultrasound-based visual feedback is used during therapy sessions, as enhanced tongue visibility can support more effective biofeedback interventions [199]. High sensitivity ensures that relevant articulatory gestures are detected and represented clearly to both clinicians and patients. However, the accompanying decline in overall accuracy and precision indicates that contextual cues outside the ROI contribute to discrimination. Removing peripheral anatomy may eliminate information about tongue–palate or tongue–jaw relationships that provide implicit phonetic context [200]. Without this context, the classifier must rely solely on fine-grained tongue features, a more demanding task. Stage 1, therefore, benefits from comprehensive spatial information, while Stage 2 focuses exclusively on articulatory detail; the latter improves sensitivity but sacrifices some contextual awareness.

Stage 2's lower precision also suggests subtle style or texture inconsistencies in the synthetic ROI outputs, which may prompt false positives. The cross-domain evaluation supports this interpretation: a classifier trained on real ROI data underperformed when tested on synthetic ROI images, revealing a minor distribution shift. These differences likely stem from edge sharpness or texture uniformity at ROI boundaries. Future iterations could address this through a texture-consistency or perceptual loss, enforcing that generated ROI images match real ones not only structurally (via SSIM/PSNR) but also perceptually [40], [41]. Incorporating pretrained feature extractors for perceptual regularisation could further reduce any remaining domain gap.

The suitability of Pix2Pix as the backbone for both stages was reaffirmed by these findings. The paired-training objective was ideal for this setting, where corresponding input–target pairs were available. In Stage 1, it enabled geometric rescaling without distorting tongue shape, and in Stage 2, it supported accurate localisation and in-painting of the tongue region. The adversarial component ensured outputs remained visually plausible, while the L1 reconstruction term enforced pixel-level fidelity, together producing the extremely high SSIM observed. The U-Net architecture's encoder–decoder symmetry preserved spatial layout throughout both stages, a property crucial for maintaining clinical trust. Interestingly, classifiers trained on Pix2Pix-generated ROIs sometimes exceeded performance obtained using manually cropped ROIs, implying that the generative model acted as an adaptive denoiser, smoothing speckle noise and normalising intensity. This echoes findings from broader ultrasound research, where GAN-based enhancement improves downstream analysis by suppressing artefacts while preserving anatomy [195].

Contextual comparison with prior UTI classification work is instructive but must be interpreted carefully. Ribeiro et al. [16] achieved 59.4% accuracy on a four-class phonetic classification task using raw UTI frames from TD children, a setting that differs from ours in three important respects: the classification problem is multi-class rather than binary, the population is TD rather than a clinical versus control comparison, and no domain standardisation was applied. These differences mean the comparison is motivational rather than strictly benchmarked. Nevertheless, the contrast is meaningful: the well-documented difficulty of learning from heterogeneous, unprocessed UTI data, evident in Ribeiro et al.'s result, is precisely the challenge that the two-stage Pix2Pix framework is designed to address. The substantial performance gain observed in Stage 1 reflects both the simplification of the classification problem and, more substantively, the reduction in acquisition-induced variance achieved by generative standardisation.

More broadly, the GAN-based harmonisation approach builds on established methods in radiology and neuroimaging, where domain adaptation techniques have become essential for multi-centre studies [201]. Our contribution extends this paradigm to speech ultrasound, where standardisation tools have historically been underdeveloped. By demonstrating that learned generative mappings can surpass interpolation-only methods while preserving anatomical fidelity, this work establishes a foundation for future UTI standardisation efforts.

From a clinical perspective, several implications emerge. First, Stage 1's ability to generate a common imaging plane across different ultrasound devices addresses a major barrier to multi-centre use of UTI-based diagnostics. Variations in probe design or FoV currently hinder reliable comparison between datasets [202]. Our results show that GAN-based harmonisation can remove these discrepancies, enabling consistent interpretation. The cross-domain tests confirmed that classifiers trained on one domain performed equivalently on GAN-standardised data from another, effectively erasing inter-device bias.

The higher recall achieved in Stage 2 carries a distinct clinical implication beyond the biofeedback utility discussed above: in paediatric SSD populations, the cost of a missed positive case, failing to detect disordered articulation, is higher than the cost of a false alarm that triggers unnecessary clinical review. Stage 2's sensitivity profile is therefore well-matched to a screening or monitoring role, where capturing all genuine SSD instances is prioritised over minimising false positives. Finally, the faster inference observed on ROI-refined images implies potential for integration into real-time feedback systems: once ROI refinement is pre-computed, downstream diagnostic models could operate at interactive speeds, supporting visual biofeedback during therapy [203]. This represents a practical advantage for clinical translation, where latency constraints often limit the applicability of computationally intensive preprocessing.

The robustness of Stage 1 outputs and the focused sensitivity of Stage 2 thus illustrate complementary capabilities. Stage 1 FoV standardisation provides approximately 1% improvement over interpolation through learned texture refinement and 4% improvement through data augmentation when using mixed training, while cross-domain generalisation is nearly perfect, confirming that synthetic standardised images are functionally equivalent to real ones. Stage 2 ROI refinement prioritises sensitivity at the expense of some precision, making it particularly suitable for biofeedback applications, while also improving computational efficiency through ROI focus, achieving 50% faster inference times that support real-time deployment. Depending on the application, cross-centre benchmarking, diagnostic decision support, or therapy feedback, the pipeline can be configured accordingly. Moreover, the superior performance of mixed (real + synthetic) datasets reinforces the practicality of synthetic data augmentation in medical imaging.

The generative process expanded the training distribution without compromising realism, improving robustness and confirming that the Pix2Pix-generated frames were statistically representative of authentic anatomical variability.

Despite these strengths, several limitations should be acknowledged. First, the reliance on paired training data limits scalability to settings where precise ground-truth pairs are available. Future work could explore unpaired translation methods, such as CycleGAN [204] or diffusion-based models [49] that do not require explicit paired examples, potentially enabling broader application across diverse acquisition protocols. Second, although no anatomically implausible artefacts were observed during qualitative inspection, generative models inherently carry a risk of hallucination, generating plausible-looking but anatomically incorrect features [205]. For instance, when standardising FoV from 133° to 97°, the generator must infer the appearance of central tongue regions that may be partially occluded or poorly imaged in the original wide-angle view. While our high SSIM scores (0.96) suggest accurate reconstruction, the possibility of subtle interpolation artefacts cannot be entirely excluded without expert anatomical verification. This risk must be carefully managed in clinical contexts through rigorous validation and quality control procedures. Expert review of generated outputs and comparison with ground-truth anatomy remain essential safeguards. Third, the dataset size was moderate, which may limit the model's exposure to rarer articulatory patterns or edge cases. Expanding the training corpus with additional multi-centre data would likely improve robustness and generalisation to unseen speakers and acquisition settings. Finally, the modest domain shift observed in Stage 2 cross-domain testing (80.82% accuracy, Table 5.4) suggests that texture and boundary characteristics in ROI-refined outputs do not perfectly match those in real ROI images. Incorporating perceptual losses [206] could mitigate this limitation and improve cross-domain transferability.

5.5 Summary

This chapter presented a two-stage cGAN framework to enhance UTI data for automated SSD diagnosis. Stage 1 standardised the FoV, harmonising spatial representations across datasets while preserving anatomical integrity. Stage 2 refined the ROI, generating tongue-centred images that reduced background noise and improved classifier sensitivity. Together, these stages addressed two persistent challenges in UTI analysis, acquisition variability and inconsistent anatomical focus, while maintaining high visual fidelity and diagnostic reliability. The pipeline demonstrated tangible benefits for automated SSD classification. GAN-based FoV standardisation improved model accuracy and generalisability across domains, supporting interoperability in multi-centre clinical applications. ROI refinement enhanced sensitivity, reducing the likelihood of missed positive cases, an essential consideration for early identification of SSDs in paediatric populations. The high SSIM achieved in both stages confirms that the generative process preserved fine articulatory structure across both geometric standardisation and ROI refinement, a prerequisite for clinician confidence in AI-assisted imaging. Stage 1's exceptionally high PSNR reflects the deterministic nature of the bicubic interpolation targets used for training, confirming accurate encoding of the rescaling transformation rather than claiming general fidelity beyond this paired setting. Beyond quantitative performance, this generative preprocessing framework contributes to the broader goals of clinical speech pathology. By producing consistent and interpretable ultrasound representations supports real-time visual feedback in therapy, enhances training data quality, and enables longitudinal tracking of articulatory change. The same approach could also be used for data augmentation, generating synthetic examples of underrepresented phoneme classes or rare articulatory errors to mitigate class imbalance in paediatric datasets. Several directions for future work emerge from the findings, and it is worth distinguishing these by both tractability and expected impact. The most immediately actionable improvement is the incorporation of a perceptual or style-consistency loss into Stage 2 training. The domain gap identified in cross-domain generalisation testing has a clear mechanistic cause, texture and boundary inconsistencies at ROI edges, and perceptual regularisation using pretrained feature extractors offers a well-established solution that could be applied within the existing Pix2Pix framework without architectural redesign.

A closely related and similarly tractable direction is the development of a dual-input classifier that jointly processes the ROI-refined image alongside a down-sampled full-frame representation, combining Stage 2's sensitivity to fine articulatory detail with the broader anatomical context that Stage 1 preserves; multi-scale and dual-branch CNN designs have demonstrated this benefit in related medical imaging tasks and represent a natural extension of the current pipeline. At a longer timescale, extending the framework to unpaired translation methods, such as CycleGAN or diffusion-based models, would remove the dependency on paired training data and enable broader application across sites and devices where corresponding reference images are unavailable. Expanding validation to larger, multi-centre datasets represents the most impactful longer-term direction, as it would provide the cross-site generalisability evidence necessary for clinical adoption. In conclusion, the proposed two-stage framework offers a robust and generalisable solution for standardising and enhancing UTI data. It advances both the technical and clinical frontiers of SSD diagnosis by improving accuracy, interpretability, and consistency, while laying the groundwork for scalable, accessible AI-driven systems that support early intervention and broaden the reach of speech therapy for children with SSD. In the broader thesis structure, this chapter operates at the image-manifold level, complementing the representation learning focus of 4 and enabling the data-efficiency strategies explored in Chapter 6.

Chapter 6

6 Data Sufficiency and Annotation Optimisation in Ultrasound-Based Speech Classification

Following the generative standardisation framework developed in Chapter 5, the next critical challenge concerns data efficiency: achieving clinically meaningful performance when expert-annotated UTI datasets are limited and costly to obtain. Annotating UTIs requires specialist expertise, and DL models typically depend on large, labelled corpora, creating a major bottleneck for research scalability and clinical deployment. This chapter directly addresses Challenge **C2** data scarcity and annotation efficiency. It fulfils Objective 4, as defined in Chapter 1 (Section 1.3.1), by determining the minimum dataset size required for stable performance and optimising annotation effort under constrained labelling resources. Building on cost-optimisation principles previously proposed for ultrasound imaging, the framework is adapted and extended to the specific challenges of paediatric articulatory UTI. A two-phase strategy is introduced. First, statistical power-curve modelling quantifies how classification accuracy scales with dataset size, identifying the point of diminishing returns. Second, AL is applied to prioritise the most informative samples for expert annotation using uncertainty-based acquisition criteria. Together, these approaches establish a quantitative, architecture-aware methodology for balancing accuracy, cost, and annotation workload. The results demonstrate that targeted sample selection can reduce annotation requirements by approximately half while maintaining over 90% classification accuracy. By formalising the trade-off between dataset size and diagnostic reliability, this chapter provides a reproducible and economically grounded framework for scalable UTI dataset design, directly supporting the clinical translation of DL systems for paediatric speech assessment.

A version of this work has been published as: Al Ani, Saja, Cleland, Joanne, and Zoha, Ahmed (2025) A Framework for Assessing and Optimising Data Sufficiency in Ultrasound Tongue Imaging. In: 13th International Conference on Bioimaging (BIOIMAGING 2026), Marbella, Spain, 02-04 Mar 2026, (Accepted for Publication).

6.1 Introduction

Automating clinical assessment from UTI is constrained less by modelling capacity than by the realities of data: small, heterogeneous cohorts, variable acquisition geometry, and the scarcity and cost of expert annotation. Unlike MRI or CT, where extensive, standardised repositories exist, UTI corpora remain modest in scale and scope [207], [208]. Each annotation must be produced manually by trained SLTs with expertise in articulatory phonetics, making large-scale labelling both time-consuming and costly [209], [210]. These constraints impose a natural ceiling on dataset size and, consequently, on the generalisability of DL models for UTI. Developing diagnostic-grade DL systems, therefore, requires substantial investment not only in data acquisition but also in expert time.

This bottleneck parallels the situation across many medical-imaging domains, where annotation costs dominate research budgets [211]. Methods such as transfer learning [212], few-shot learning [213], and self-supervised approaches, including masked autoencoders [214], reduce dependence on labelled data, but do not eliminate the fundamental limits of clinician time and financial cost. In UTI, where imaging protocols differ across speakers and clinical sites, cost optimisation is essential for sustainable research and eventual clinical deployment. Recognising that cost-optimisation strategies are well-established in clinical trial design, Lawley et al. [215] recently formalised a two-phase framework for medical ultrasound that integrates power-curve analysis with AL. Their work demonstrated substantial cost reductions across three imaging modalities: breast lesion detection, COVID-19 lung pathology, and fatal plane identification. By adapting principles from clinical feasibility studies, which estimate sample sizes required for statistically reliable outcomes, they showed that ultrasound dataset requirements follow a law of diminishing returns: adding annotated data improves model performance, but each additional sample yields progressively smaller gains [216].

Their framework provided the first systematic methodology for balancing accuracy, annotation cost, and dataset size in medical ultrasound applications. However, several limitations constrain the direct application of Lawley et al.'s findings to articulatory UTI. First, their analysis relied exclusively on a single architecture without exploring whether cost-optimisation strategies generalise across different model families, a critical consideration given substantial performance differences between architectures. Second, they employed visual curve fitting rather than quantitative goodness-of-fit validation, limiting the statistical rigour of sample-size predictions for prospective study planning. Third, their validation focused on static anatomical imaging, leaving open whether the framework generalises to imaging with higher inter-speaker variability.

Clinical-trial design offers a useful analogy for addressing these challenges. Because recruiting entire populations is rarely feasible, medical researchers conduct pilot or feasibility studies to estimate the sample size required for statistically reliable outcomes. The same principle applies to ML: strategic trade-offs can balance accuracy, cost, and feasibility while maintaining clinical validity. Statistical power analysis estimates how confidently results obtained from a sample can be expected to generalise to a population, given a chosen significance level [217]. A power curve illustrates how statistical power changes with sample size under fixed significance and variance assumptions. In general, power depends on three main factors: the significance threshold, the true effect size, and the sample size available to detect that effect [218]. Previous studies have used power analysis to predict classification performance [219] and to determine optimal dataset sizes in retinal optical coherence tomography research [220], establishing precedent for empirical curve-fitting approaches in medical imaging.

In practice, cost is often the primary constraint in ML-based medical-imaging studies, where increasing sample size dramatically raises expenses. Simple heuristic rules, for example, using 10, 100, or 1000 samples per model parameter [221], are frequently adopted, but these are arbitrary and may lead to datasets that are either too small or unnecessarily large. Such rules offer little insight into the quantitative balance between cost and performance. More advanced model-based sampling methods, which estimate sample requirements from theoretical generalisation bounds [222], [223], remain difficult to link directly to monetary or time costs.

The approach proposed here instead applies empirical curve fitting to estimate sample-size sufficiency, enabling accurate prediction of both time and financial expenditure, analogous to feasibility analyses used in controlled clinical trials [224], [225].

When large volumes of unlabelled ultrasound data are available, annotation can be made more efficient through targeted selection. Early in data analysis, unsupervised clustering [161] can reduce redundancy by grouping similar samples automatically; however, in this study, AL was adopted to guide manual annotation more effectively. Although manual labelling is more resource-intensive than self-supervised or fully automated alternatives, it remains the regulatory gold standard for medical-device validation and was therefore used as the reference method. AL is a strategy for data-efficient model training, which is an iterative procedure in which a neural network analyses unlabelled data, identifies the samples about which it is most uncertain, and requests expert annotation for those specific cases [226]. Several AL strategies exist, including diversity sampling, which selects maximally varied examples [227], and uncertainty sampling, which focuses on the lowest-confidence predictions [228]. The present study employed selective uncertainty sampling, targeting ultrasound frames where the model's confidence was lowest, typically those representing ambiguous or atypical articulatory gestures. Each iteration formed part of an AL loop, incrementally refining the model's understanding by focusing annotation on the most informative samples. AL has demonstrated substantial promise across ultrasound applications, including weakly supervised breast-lesion detection [229], multimodal liver-fibrosis assessment via elastography [230], and semi-supervised COVID-19 lung-disease classification [231]. Lawley et al. reported annotation reductions of 30-40% using uncertainty-based sampling for breast and lung ultrasound. However, whether these efficiency gains transfer to articulatory ultrasound remains empirically unvalidated. Although manual labelling is more resource-intensive than self-supervised or fully automated alternatives, it remains the regulatory gold standard for medical-device validation and was therefore adopted as the reference method in both Lawley's work and the present study.

6.1.1 Contributions

The three contributions outlined below directly advance Objective 4, establishing a principled, cost-aware framework for determining dataset sufficiency and optimising annotation effort for UTI. This chapter contributes a methodological and practical framework for data-efficient DL in UTI. Its innovations lie not in proposing new architectures, but in establishing quantitative and reproducible principles for determining how much data is needed and how that data can be most efficiently annotated. The specific contributions are as follows:

1. Model selection and statistical validation. Statistical power analysis is applied to model the empirical relationship between dataset size and classification accuracy. By systematically comparing power-law learning curves and identifying the point of diminishing returns, and adding quantitative goodness-of-fit metrics, providing an evidence-based estimate of the minimum dataset size required to achieve clinically stable performance.

   Impact on C2: This contribution enables principled dataset planning and prevents unnecessary data collection and annotation beyond the point of meaningful performance gain.
2. AL pipeline for annotation reduction. A selective uncertainty-sampling strategy is implemented to prioritise the most informative or ambiguous ultrasound frames for expert labelling. This approach reduces redundant annotation and directs SLTs' efforts toward cases that yield the greatest improvement in model performance.

   Impact on C2: By concentrating annotation on high-value samples, this contribution substantially lowers expert labelling cost while preserving diagnostic accuracy.
3. Architecture-dependent analysis of data efficiency and AL. Unified framework for scalable and reproducible dataset optimisation. The two approaches, power-curve modelling and AL, are integrated into a single cost-aware workflow. Experiments demonstrate that the combined framework can sustain over 90 % classification accuracy while cutting annotation requirements by more than half, establishing a blueprint for economically viable UTI dataset development.

Impact on C2 and C3: This integrated workflow supports scalable dataset growth and improves clinical feasibility by aligning model performance, annotation effort, and deployment constraints.

Together, these contributions move the field from heuristic dataset expansion toward quantitative, data-efficient study design. They directly address Challenge **C2** data scarcity and annotation efficiency and form the methodological bridge to the deployment and scalability considerations discussed in Chapter 7.

6.2 Method

6.2.1 Proposed Framework for Cost-Efficient Sampling and Annotation

This study was conducted in two complementary phases designed to evaluate cost-efficient dataset development for UTI. The framework addresses two questions:

1. How much data is sufficient? (Phase 1 Optimised dataset capture).
2. How can annotation effort be minimised without loss of accuracy? (Phase 2 –Cost-effective annotation).

Together, these phases form a principled framework for reducing both data-collection and annotation costs in UTI-based ML systems.

AlexNet was employed as the primary experimental architecture throughout both phases. AlexNet provides continuity with prior ultrasound cost-optimisation studies while offering computational efficiency for the repeated training cycles required in power-curve analysis (20 replicate runs per dataset fraction). Hyperparameters were purposely simple and consistent across all experiments to demonstrate framework functionality. Following completion of the two-phase analysis with AlexNet, the framework was validated using EfficientNet-B0 (the architecture employed in Chapter 4) to assess whether cost-optimisation findings generalise across architectures with different parameter efficiency and baseline performance characteristics.

EfficientNet-B0 validation results are presented in Section 6.3.4, and detailed results are provided in Appendix C.

6.2.1.1 Phase 1 – Optimised Dataset Capture

Phase 1 quantifies the relationship between dataset size and classification performance through iterative training and learning curve characterisation. Figure 6.1 illustrates the workflow used to determine dataset sufficiency.

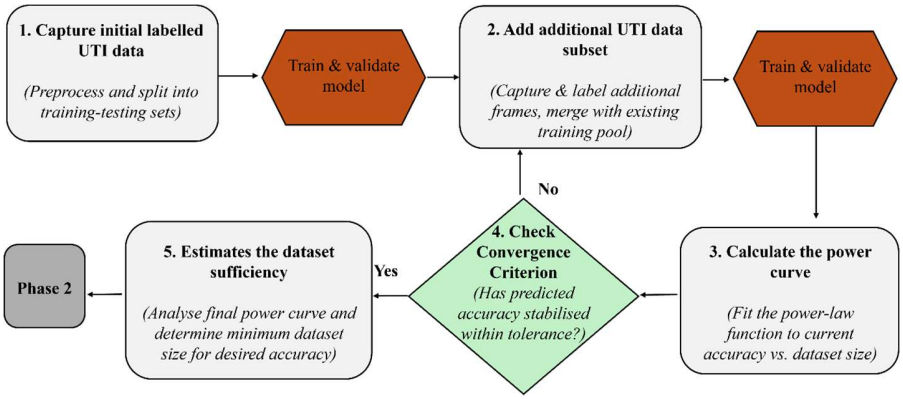

Figure 6. 1: Iterative Workflow for Estimating Dataset Sufficiency via Power-Curve Modelling.

The workflow consisted of the following steps:

1. **Initial training:** Training begins with a small, labelled subset of the available data (10% of the training pool). AlexNet is trained and evaluated, and the test accuracy is recorded. To account for stochasticity in training and subset selection, each configuration is repeated 20 times using different random seeds.
2. **Incremental dataset expansion:** The labelled training set is expanded in 10% increments up to 90% of the available training pool. At each increment, AlexNet is retrained and evaluated under identical protocols, yielding a distribution of accuracy values for each dataset size. This produces an empirical accuracy-sample size curve with associated run-to-run variability.

3. **Curve fitting:** To model the observed scaling behaviour, an exponential decay function is fitted to classification accuracy as a function of labelled training samples. Let x denote the number of training samples and A(x) the corresponding test accuracy (%). The model is defined as:

$$A(x) = A\infty - B \exp(^{-Cx}) \qquad (6.1)$$

where A∞ represents asymptotic accuracy, B > 0 controls the initial learning gap, and C > 0 determines the learning rate. This functional form ensures bounded, monotonically increasing accuracy with progressively smaller improvements, characteristic of supervised learning with diminishing returns. The exponential specification was selected following systematic comparison of five candidate functional forms: inverse power-law (as employed by Lawley et al.), exponential decay, logarithmic, and polynomial models of degrees 2 and 3. Model selection prioritised empirical fit quality (Akaike's Information Criterion, AICc) and theoretical validity (bounded asymptotic behaviour, guaranteed diminishing returns). Detailed model comparison is presented in Section 6.3.1.1. Following completion of Phase 1 and Phase 2 analysis with AlexNet, the framework was validated using EfficientNet-B0 to assess generalisability across architectures (Section 6.3.4 and Appendix C).

Model parameters are estimated using bounded non-linear least squares. To ensure physically meaningful solutions and numerical stability, parameters are constrained such that A∞ ∈ [0, 100], B > 0, and C > 0. The fit uses all replicate-level measurements, incorporating training variability into parameter estimation. Model adequacy is assessed through R^2, RMSE, and MAE. Uncertainty quantification: Uncertainty in fitted curves is quantified via replicate-level bootstrap resampling (1,000 iterations). Accuracy measurements are resampled with replacement, the exponential model refitted, and predictions evaluated on a dense sample-count grid. The resulting distribution yields 95% confidence bands around the fitted curve.

4. **Dataset sufficiency estimation:** Determined by analysing the fitted curve using: (1) a stability criterion identifying saturation regions where additional data yield negligible gains (±2% accuracy), and (2) target-accuracy thresholds (70–90%) estimated by curve inversion to determine minimum samples required for desired accuracy levels.

These steps provide quantitative, uncertainty-aware estimates of minimum dataset size, forming the basis for cost analysis and AL in Phase 2.

6.2.1.2 Phase 2 – Cost-Effective Annotation via Active Learning

Phase 2 aims to reduce annotation cost by prioritising the most informative training samples while maintaining classification performance. Building on the dataset sufficiency insights obtained in Phase 1, an AL framework is employed to iteratively select samples for annotation from an unlabelled pool. The workflow, illustrated in Figure 2, proceeds as follows:

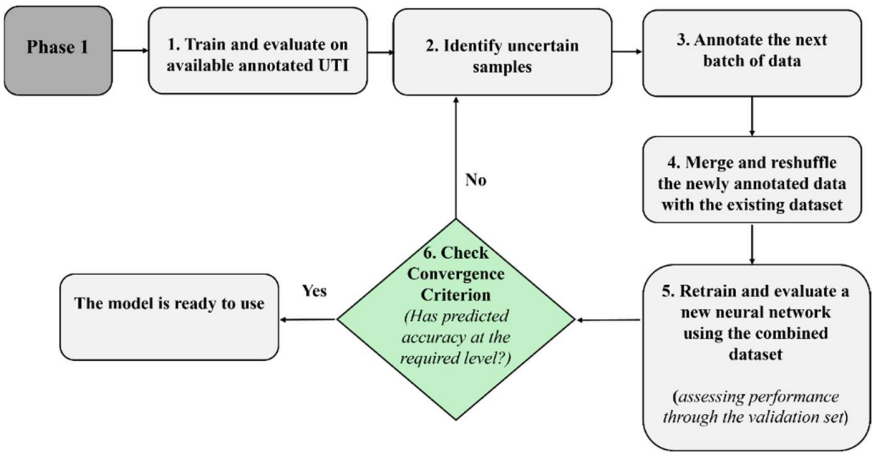

Figure 6. 2: Workflow of Phase 2 AL for Cost-Effective Annotation.

The process was implemented as follows:

1. **Initial labelled set and unlabelled pool**: The AL process begins with a small, labelled set, corresponding to 10% of the available training pool. The remaining data form an unlabelled pool from which samples are iteratively selected for annotation. A fixed, speaker-disjoint test set is retained throughout all AL experiments.
2. **Model training protocol**: At each AL iteration, the model is trained using the currently labelled dataset to avoid bias accumulation from earlier selection rounds. Training hyperparameters and evaluation protocols are kept identical across iterations and across all selection strategies to ensure a fair comparison.
3. **Uncertainty-based sample selection:** An uncertainty-based sampling strategy is used to identify informative samples. After training, the model is applied to the unlabelled pool, and an uncertainty score is computed for each sample based on the predicted class probabilities. Specifically, predictive entropy is used as the acquisition function and is calculated from the softmax output probabilities. Higher entropy values indicate greater model uncertainty. This criterion was selected because it prioritises samples for which the classifier is least confident, enabling expert annotation effort to be focused on the most informative examples. In the context of UTI, where annotation by SLTs is costly and time-consuming, entropy-based sampling improves annotation efficiency while maintaining model performance. At each iteration, a fixed batch of the most uncertain samples is selected for annotation and added to the labelled set.
4. **Baselines and comparison strategy:** AL performance is compared against a random sampling baseline, in which samples are selected uniformly at random from the unlabelled pool. Both strategies use identical initial labelled sets, batch sizes, training schedules, and evaluation protocols. To characterise sensitivity to initial conditions, both AL and random sampling were evaluated across 10 independent runs with different identical seeds for the initial 10% labelled set. This ensures that any performance difference can be attributed to the selection strategy rather than initialisation effects.
5. **Iteration schedule and stopping criterion:** The AL loop proceeds in fixed increments until the labelled dataset reaches the estimated dataset sufficiency threshold identified in Phase 1 or until no further performance improvement is observed. Performance is evaluated after each iteration on the fixed test set. AL with uncertainty sampling is

deterministic given a fixed model architecture and initial seed. Once the initial 10% labelled set is randomly selected, all subsequent samples are selected based on the model's predictive entropy scores on the unlabelled pool. Model performance is evaluated as a function of the number of labelled samples, allowing direct comparison between AL and random sampling under the same annotation budget.

6.2.2 Datasets and Preprocessing

This study utilised the same dataset introduced in Chapter 4 (TD, CP±L), ensuring continuity with the image-representation and explainability analyses presented there. The dataset comprised 2,660 ultrasound frames and differs from the dataset used in Chapter 3 (UltraSuite-UXTD corpus) in three key ways: (i) a larger sample size (28 vs. 9 children), enabling power-curve modelling and the analysis of learning behaviour across extended dataset fractions; (ii) a balanced TD/CP±L composition, ensuring clinical representativeness for binary diagnostic evaluation; and (iii) a binary classification task (TD vs. CP±L) rather than multi-class phoneme classification. While Chapter 3 established baseline modelling approaches on a smaller corpus, the present chapter addresses questions of data efficiency and annotation sufficiency, which require larger, clinically balanced cohorts and finer-grained sampling across dataset sizes.

All images were intensity-normalised, converted to grayscale, and replicated across three channels to match the input requirements of the pretrained CNN. For all experiments, the subject-disjoint split was kept fixed across runs. Variability across repetitions arose solely from random initialisation of network weights and random selection of training subsets within the fixed training pool. This design ensures that reported variability reflects training stochasticity rather than changes in participant composition.

6.2.3 Deep learning setup

A CNN based on AlexNet was used as the baseline classifier for all experiments. The model was initialised with ImageNet-pretrained weights [232] and its final fully connected layer was replaced with a two-unit softmax output corresponding to the TD and CP±L categories. AlexNet offers continuity with prior cost-efficiency studies, having been widely used in ultrasound research domains such as breast and foetal imaging as a reproducible baseline for cost-analysis tasks. Using the same architecture ensures methodological consistency and enables direct benchmarking of UTI against established imaging modalities. Although more advanced networks can achieve higher absolute accuracy on large datasets, their greater computational demand and tendency to overfit under data scarcity make them less practical for iterative experimentation. AlexNet was therefore selected not as a state-of-the-art classifier, but as a stable, interpretable, and computationally economical benchmark for dataset-efficiency evaluation. Table 6.1 shows the model hyperparameters used for the power-curve and AL experiments.

Table 6. 1: Summary of Hyperparameters for Power-Curve and AL Experiments.

Component	Setting	Justification
Model	ImageNet-pretrained AlexNet	Selected as a stable, computationally efficient baseline for repeated training across multiple data budgets and active-learning rounds.
Input	UTI frames converted to three-channel images and ImageNet-normalised	Ensures compatibility with the pretrained AlexNet input format.
Input size	227x227	Chosen to match AlexNet input requirements while preserving sufficient tongue-region detail.
Output classes	2, TD vs CLP	Matches the binary diagnostic classification task.
Optimiser	Adam	Provides stable convergence across repeated small-data experiments.
Learning rate	0.001	Standard initial learning rate for Adam and suitable for baseline fine-tuning.
Batch size	64	Provides efficient training while maintaining stable gradient estimates.
Epochs	20	Chosen to allow convergence while keeping repeated power-curve and AL experiments computationally feasible.
Loss function	Cross-entropy	Appropriate for two-class classification.

Early stopping	Patience = 10	Reduces overfitting and prevents unnecessary training when validation performance plateaus.
Scheduler	ReduceLROnPlateau; factor = 0.5 after 3 epochs	Stabilises training by reducing the learning rate when validation performance stops improving.
AL acquisition	Predictive entropy from softmax probabilities	Selects samples for which the classifier is least confident, focusing expert annotation on informative cases.

6.2.3.2 Training and Evaluation Protocol

The dataset was first divided randomly into training and test sets using an 80/20 split. Within the training set, an additional stratified random sampling was applied to ensure that each class was proportionally represented. The proportion of training data was progressively increased by 10% in each iteration to examine how sample size influences model accuracy. This strategy was designed to mimic gradual data collection over time. Each Phase 1 configuration was repeated 20 times, with each training run lasting 20 epochs, to ensure statistical robustness in learning curve estimation. Phase 2 AL experiments were repeated 10 times to characterise sensitivity to initial seed selection, as discussed in Section 6.2.1.2. A subject-wise splitting strategy ensured that frames from the same child never appeared in more than one subset. This separation prevents information leakage and ensures that results reflect genuine generalisation to unseen speakers, critical for eventual clinical translation. Model performance was assessed on the fixed held-out test set using accuracy. This metric provides complementary views of classification reliability across the two categories. Adopting the same metric and baseline architecture as prior ultrasound cost-analysis studies ensures methodological continuity and allows direct comparison of results, isolating the impact of the proposed data-efficiency framework rather than differences in model design.

6.3 Results

6.3.1 Phase 1- Optimised Dataset Capture

To determine dataset sufficiency, AlexNet was trained on progressively larger subject-disjoint subsets of the UTI training data, ranging from 10% to 90% in 10% increments. Each configuration was repeated 20 times using different random seeds to account for variability arising from data sampling and network initialisation. Figure 6.3 illustrates the relationship between training set size and classification accuracy. Mean accuracy increased rapidly as the dataset expanded from 10% to approximately 50%, reflecting substantial early performance gains once sufficient articulatory variability was captured. At smaller dataset fractions (10–30%), the 95% CIs were wide, indicating unstable performance when the model was trained on limited data. As the training set size increased, the CIs narrowed, demonstrating more stable and reliable performance estimates. Beyond approximately 60–70% of the dataset, both the mean accuracy and its confidence bands began to flatten, indicating that additional annotated samples yielded diminishing performance improvements.

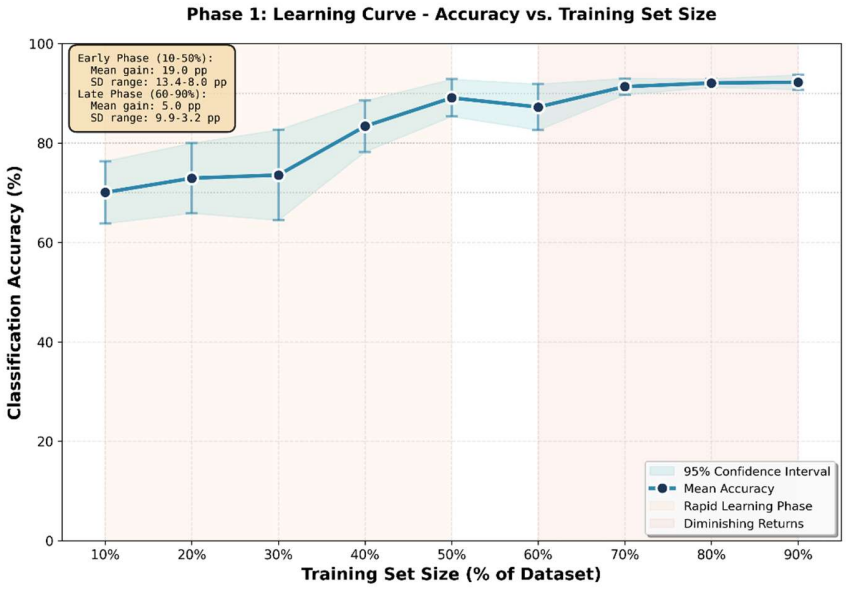

Figure 6. 3: Relationship Between Training Set Size and Classification Accuracy.

6.3.1.1 Model Selection

To ensure an appropriate functional form for learning curve modelling, five candidate models were systematically compared: inverse power-law, exponential decay, logarithmic growth, and polynomial models (degrees 2 and 3). Each model was fitted to the observed accuracy measurements across all dataset fractions and replicates using non-linear least squares optimisation. Model performance was evaluated using complementary criteria: (1) statistical fit quality via AICc, which penalises model complexity while rewarding goodness-of-fit; (2) predictive accuracy via R^2, RMSE, and MAE; and (3) theoretical validity, specifically whether the functional form exhibits bounded asymptotic behaviour and guaranteed diminishing returns, fundamental properties expected from supervised learning curves. Table 6.2 summarises the model comparison results.

Table 6.2: Comparison of Candidate Learning Curve Models.

Model	R^2	RMSE (pp)	MAE (pp)	AICc
Polynomial (deg 2)	0.947	1.95	1.46	48.3
Exponential Decay	0.939	2.07	1.61	49.5
Logarithmic	0.013	2.48	1.87	52.7
Polynomial (deg 3)	0.955	1.78	1.39	53.9
Inverse Power-Law	0.811	3.66	3.10	59.7

While polynomial models achieved the best raw statistical fit ($R^2 = 0.95$ for the second-degree polynomial), they were excluded from primary consideration due to a lack of bounded asymptotes and an inability to guarantee monotonic accuracy improvement beyond the observed data range.

Polynomial functions provide excellent interpolation but exhibit unreliable extrapolation behaviour, potentially predicting unbounded accuracy growth or non-monotonic decline, properties incompatible with learning curve theory and practical dataset planning. Among theoretically valid models with bounded asymptotes, the exponential decay model achieved substantially superior fit compared to alternatives ($R^2 = 0.939$, RMSE = 2.07 pp). The inverse power-law model, commonly employed in learning theory literature and successfully applied by Lawley et al. [215] to general clinical ultrasound imaging, showed considerably weaker fit ($R^2 = 0.811$, RMSE = 3.66 pp). This systematic deviation indicates that the power-law specification, while effective for Lawley's breast, lung, and fatal datasets, does not adequately capture the learning dynamics observed in articulatory ultrasound.

The exponential decay model was therefore selected for all subsequent analyses, balancing empirical fit quality with theoretical validity. This finding represents a methodological refinement of Lawley et al.'s framework: rather than assuming power-law scaling a priori, systematic model comparison enables domain-specific selection of the most appropriate functional form. Figure 6.5 confirms adequate model fit using the true observed means from 20 replicate runs. Residuals showed no systematic pattern when plotted against fitted values or sample size, indicating that the exponential model captures the central tendency of learning behaviour without systematic over- or under-prediction. Residuals were approximately normally distributed (mean = -0.16 pp, SD = 2.41 pp, RMSE = 2.28 pp), with 78% of data points within ±2 pp. The Q-Q plot demonstrated reasonable agreement with theoretical normal quantiles, with no extreme outliers. The largest residual (-5.12 pp at 30% fraction) reflects the high run-to-run variability in this dataset (SD = 19.43 pp across replicates) rather than systematic model inadequacy.

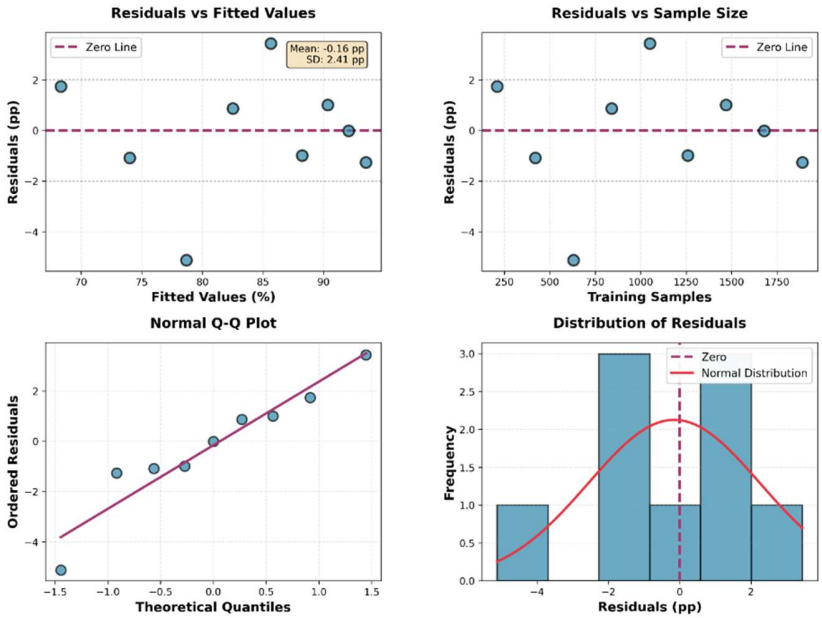

Figure 6. 4: Residual Diagnostics for the Exponential Decay Model.

The observed deviation from power-law scaling merits consideration. Lawley et al.'s successful application of inverse power-law models to general ultrasound datasets suggests that learning dynamics may vary across imaging domains. The present study's requirement for exponential modelling likely reflects domain-specific characteristics of articulatory imaging: higher inter-speaker anatomical variability, dynamic motion artefacts during speech production, and the specialised nature of phonetic annotation may produce learning curves that diverge from conventional diminishing-returns patterns observed in more standardised anatomical imaging tasks. This finding underscores a key methodological principle: learning curve functional forms should be validated empirically through systematic model comparison rather than assumed a priori based on theoretical precedent or literature reports.

6.3.1.2 Exponential Learning Curve and Dataset Sufficiency

To quantify scaling behaviour, the selected exponential decay model was fitted to classification accuracy as a function of labelled samples. Figure 6.4 presents the fitted exponential curve together with the 95% confidence band derived from bootstrap resampling. The fitted model is:

$$A(x) = 100.00 - 38.62 \times \exp^{(-0.000943x)} \qquad (6.2)$$

where $A\infty = 100.00\%$ represents the asymptotic (saturation) accuracy, $B = 38.62$ controls the initial learning gap (the difference between theoretical zero-sample accuracy and asymptotic performance), and $C = 0.000943$ determines the learning rate (the exponential rate of convergence toward saturation as training samples increase).

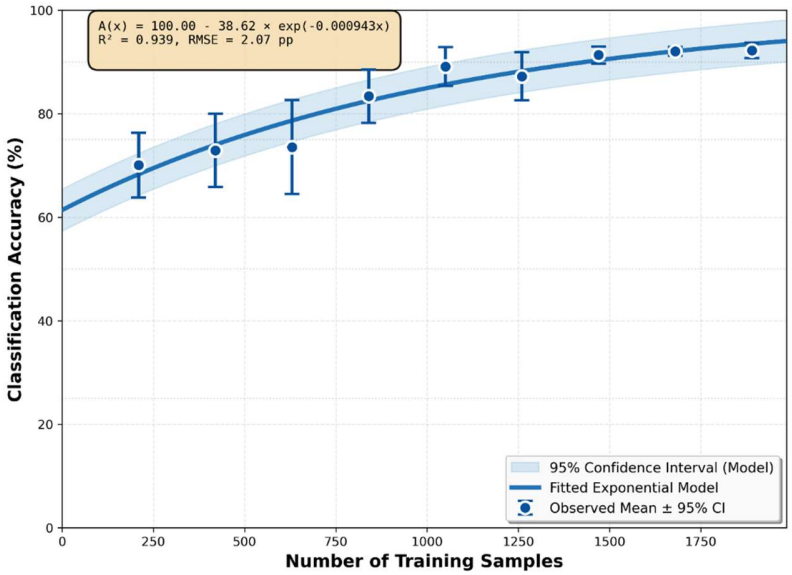

Figure 6. 5: Exponential Learning Curve for UTI Dataset Scaling.

Quantitative evaluation confirmed an excellent model fit: $R^2 = 0.939$, RMSE = 2.07 pp, and MAE = 1.6 pp. These metrics indicate that the exponential model explains approximately 94% of the variance in observed learning behaviour, with predictions typically within ±2 pp of the observed values. The superior fit quality compared to the inverse power-law alternative (R^2 =

0.811, RMSE = 3.66 pp) reflects more accurate characterisation of the dataset's specific learning dynamics, enabling more precise sample-size estimation for dataset planning.

Using the fitted exponential curve, sample size requirements were estimated for target accuracies ranging from 70% to 90%. Table 6.3 summarises the estimated number of labelled samples required to reach each target accuracy, together with 95% CIs derived from the bootstrap-refitted curves. For example, approximately 1,433 labelled samples were required to achieve 90% accuracy, while accuracies of 80–85% could be achieved using substantially fewer samples (approximately 33–48% of the dataset).

Table 6.3: Estimated UTI Dataset Requirements for AlexNet Baseline Performance using Exponential Decay Model.

Target Accuracy	Samples Required	95% CI	% of Dataset
70%	267	131-425	12.8%
75%	461	299-653	22.0%
80%	697	498-944	33.2%
85%	1002	744-1345	47.8%
90%	1433	1065-2000	68.2%

These thresholds reveal domain-specific requirements. Lawley et al. reported plateau points at 40-50% of data for breast and lung ultrasound, whereas the present UTI study requires approximately 60-70% of data before diminishing returns become pronounced (gains <2pp per 10% increment). This elevated data requirement likely reflects the increased complexity of dynamic articulatory imaging compared to static anatomical classification. However, absolute sample sizes remain modest: 1,433 samples for 90% accuracy represent a feasible annotation target for clinical speech research, particularly when combined with AL optimisation in Phase 2.

Overall, these results demonstrate that classification performance does not increase linearly with dataset size. Instead, accuracy improves rapidly at early stages before stabilising near its asymptotic value of approximately 90%. The empirically derived exponential curve analysis, therefore, provides a reproducible and uncertainty-aware basis for estimating dataset sufficiency in UTI, motivating the cost-efficient annotation strategies explored in Phase 2.

6.3.2 Phase 2- Optimised Annotation via Active Learning

Phase 2 evaluated whether uncertainty-based AL could reduce annotation costs while maintaining classification performance. To ensure statistical rigour, both AL and random sampling strategies were evaluated across 10 independent runs using identical initial seeds and training protocols. Performance was assessed on the fixed speaker-disjoint test set as the proportion of labelled training data increased from 10% to 90% in 10% increments. Figure 6.6 presents the learning trajectories for both strategies with 95% CIs. AL achieved 90% accuracy using 50% of labelled data (mean = 90.74% ± 5.45%, 95% CI: [86.84, 94.63]), whereas random sampling required 70% annotation to reach comparable performance (mean = 91.54% ± 4.05%), 95% CI: [88.64, 94.44]). This represents a 28.6% annotation reduction, corresponding to 420 fewer labelled samples (1,050 vs 1,470 samples) required for equivalent accuracy.

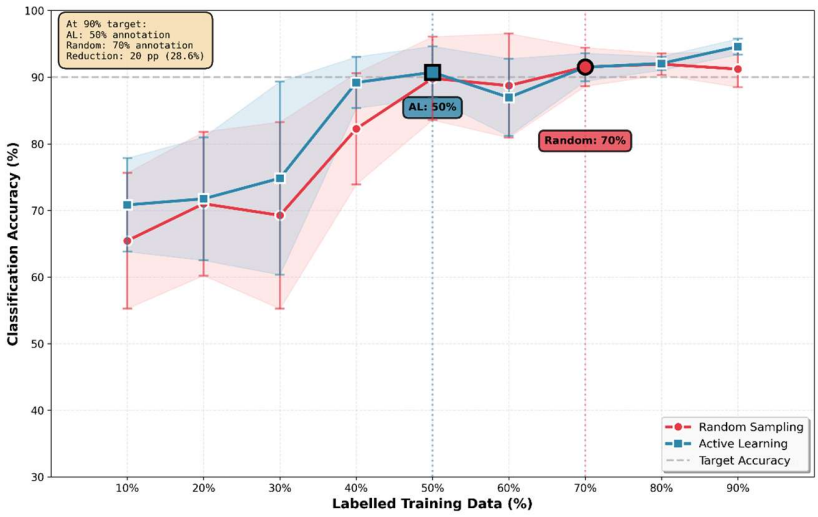

Figure 6. 6: Comparison of AL vs Random Sampling Strategies.

Paired t-tests comparing AL and random sampling at each annotation level revealed statistically significant differences at 90% annotation ($t = 2.82$, $p = 0.020$), where AL achieved 94.57% ± 3.28% compared to random sampling's 91.21% ± 4.82%. At intermediate annotation levels (40-60%), AL showed consistent numerical advantages of 5-10pp, though these did not reach statistical significance due to high run-to-run variability.

The lack of statistical significance at intermediate levels reflects the dual challenge of small sample size ($n = 10$ runs) and substantial training variance in the mid-annotation regime, rather than the absence of a practical effect. Both strategies exhibited high run-to-run variability at low annotation levels (10-30%), with SD reaching 12-20%, before stabilising above 70% annotation. This pattern mirrors the Phase 1 findings of unstable learning dynamics at small dataset fractions, reinforcing the importance of collecting sufficient data to escape the high-variance regime. AL demonstrated slightly lower average variability (SD = 7.55%) compared to random sampling (SD = 10.04%), suggesting that uncertainty-based selection produces more consistent learning trajectories by systematically targeting high-information samples.

These findings validate the effectiveness of uncertainty-based AL for cost-efficient dataset development in UTI. By directing expert annotation effort toward samples where the model was least confident, AL accelerated learning and achieved high accuracy with substantially fewer labels. The observed pattern aligns with reports from Lawley et al. [215], who demonstrated similar efficiency gains using entropy-based uncertainty sampling for breast and lung ultrasound classification. This convergence across ultrasound applications suggests that selective annotation strategies transfer effectively to articulatory imaging despite its increased complexity. The 28.6% annotation efficiency gain translates directly to expert time savings. For the present UTI dataset, AL reduces the annotation burden from 1,470 to 1,050 labelled samples to achieve 90% accuracy, a saving of 420 samples that would otherwise require manual SLT review. When combined with Phase 1 dataset optimisation (Section 6.3.1), which established that 1,433 samples suffice for 90% performance, the integrated framework achieves substantial cost reductions detailed in Section 6.3.3.

6.3.3 Combined Cost Analysis

The outcomes of Phases 1 and 2 were integrated into a unified cost model to quantify the combined savings achievable through optimised data capture and AL-based annotation. Three data-collection and annotation strategies were evaluated:

1. Full Capture + Full Annotation (FC/FA): all 2,100 training frames are collected and fully annotated.
2. Full Capture + AL (FC/AL): all 2,100 training frames are collected, with 50% (1,050 frames) selected for expert annotation via AL.
3. Optimised Capture + AL (OC/AL): only 1,433 training frames are collected, corresponding to the dataset sufficiency threshold predicted by exponential decay modelling, with 50% selected for annotation via AL.

A relative cost ratio of 1:2 between data collection and expert annotation was adopted as the base case, reflecting the substantially higher time and effort required for frame-level expert annotation compared with data acquisition and automated processing, consistent with prior medical imaging studies [29]. The total cost for each scenario was computed using Equation 6.3:

$$\text{Cost} = (N_{\text{collect}} \times C_{\text{collect}}) + (N_{\text{annotate}} \times C_{\text{annotate}}) \quad (6.3)$$

where N_{collect} and N_{annotate} represent the number of collected and annotated samples, respectively, and C_{collect} and C_{annotate} denote their corresponding unit costs. Figure 6.7 presents the total cost comparison across the three scenarios under the 1:2 cost ratio. The FC/FA baseline required a total of 6,300 cost units. Incorporating AL while maintaining full data collection (FC/AL) reduced this to 4,200 units, representing a 33.3% reduction attributable solely to annotation efficiency. Combining the exponential-model-optimised dataset capture with AL (OC/AL) yielded a further reduction to 2,865 units, equivalent to a 54.5% total saving relative to the baseline.

Notably, both optimised strategies achieved comparable ~90% classification accuracy (FC/AL: 90.74% ± 5.45%; OC/AL: estimated ~90% based on Phase 1 threshold), demonstrating that substantial savings can be realised without compromising performance.

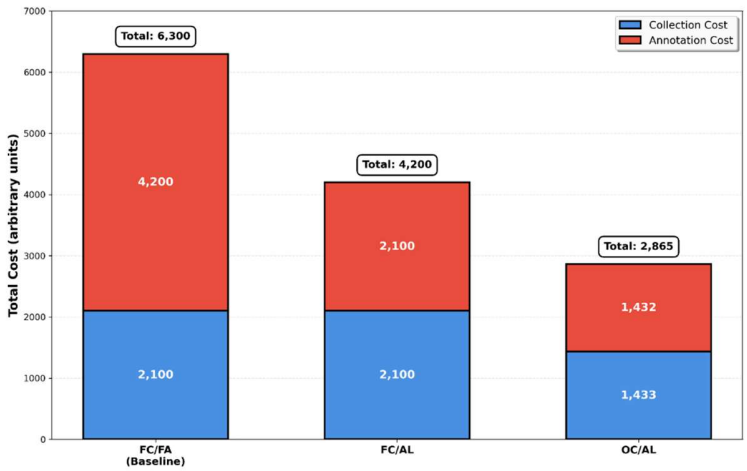

Figure 6. 7: Cost Breakdown Across Data-Collection and Annotation Strategies.

Table 6.4 summarises the detailed performance and cost outcomes for each scenario. The OC/AL configuration achieved classification accuracy equivalent to the full baseline (~90%) while requiring 32% fewer collected samples (1,433 vs 2,100) and 66% fewer annotations (716 vs 2,100). When combined with AL's 50% annotation efficiency, this produced the greatest overall cost reduction, validating the proposed framework's capacity to jointly optimise both data acquisition and expert annotation effort.

Table 6.4: Summary of Performance and Cost Across Collection and Annotation Strategies.

Strategy	Collected	Annotated	Total	Saving
FC/FA	2100	2,100	6300	0.0%
FC/AL	2100	1050 (50%)	4200	33.3%
OC/AL	1433	716 (50%)	2865	54.5%

*Note: Cost ratio 1:2 (collection: annotation).

The implications for practical deployment are significant. For exploratory or feasibility scale UTI studies, where data acquisition and annotation resources are limited, adopting this cost-aware pipeline could reduce annotation workload by two-thirds (66% reduction: from 2,100 to 716 samples) while simultaneously reducing data collection requirements by 32%. Critically, the combined framework required the collection of only 1,433 samples with annotation of just 716 samples, yet achieved performance equivalent to training on the complete 2,100-sample fully annotated corpus. The framework, therefore, provides a quantitative basis for balancing accuracy requirements against time and budget constraints in clinical research.

6.3.3.1 Sensitivity Analysis: Cost Ratio Variation

To assess the robustness of cost-optimisation conclusions across diverse institutional contexts, the combined cost analysis was repeated using collection: annotation cost ratios ranging from 1:1 to 1:4. Different research environments exhibit varying cost structures: academic institutions with established infrastructure may experience relatively lower annotation costs (approaching 1:1), whereas commercial laboratories or resource-limited settings requiring external expert consultation may face substantially higher annotation expenses (up to 1:4 or beyond). Evaluating the framework's performance across this range establishes whether the observed efficiency gains represent a robust property of the optimisation strategy or depend critically on a specific cost assumption. Table 6.5 presents the absolute costs and relative savings for each strategy across four representative cost ratios. At the 1:1 ratio, where collection and annotation costs are equivalent, the OC/AL strategy achieves 49.0% savings (3,149 vs 6,300 units), slightly lower than the base case due to the reduced relative weight of annotation costs. As the annotation-to-collection cost ratio increases to 1:3 and 1:4, the savings amplify to 57.3% and 59.1%, respectively, reflecting the increasing dominance of annotation cost in the total budget and the correspondingly greater impact of AL's annotation reduction.

Table 6.5: Cost Analysis Sensitivity to Collection: Annotation Cost Ratios.

Strategy	1:1 Ratio	1:2 Ratio	1:3 Ratio	1:4 Ratio
FC/FA	4200	6300	8400	10500
FC/AL	3150	4200	5250	6300
OC/AL	2149	2865	3581	4297

Figure 6.8 visualises these trends, illustrating how total cost and savings percentage vary with cost ratio for each strategy. Two key observations emerge: (i) a monotonic increase in absolute savings. As annotation costs rise relative to collection, the absolute cost advantage of OC/AL over FC/FA increases linearly, from 2,051 units (1:1) to 6,203 units (1:4). (ii) stable relative savings, OC/AL savings remain substantial across all ratios, ranging from 49.0% to 59.1%. The framework's efficiency gains persist even under conservative cost assumptions.

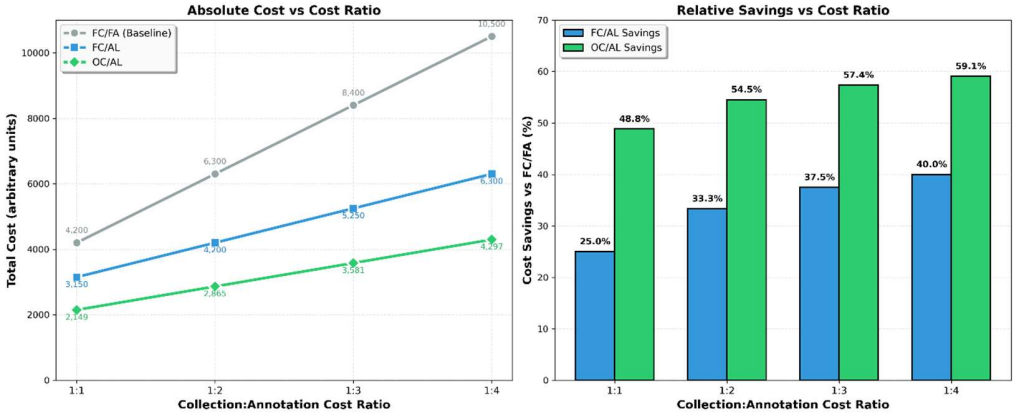

Figure 6. 8: Sensitivity Analysis.

6.3.4 Architecture Validation

To assess framework generalisability across architectures, the two-phase methodology was validated using EfficientNet-B0, the architecture employed in Chapter 4 for UTI classification. Phase 1 experiments were conducted with 20 independent runs per dataset fraction, identical to the AlexNet protocol, and Phase 2 AL employed 10 independent runs to characterise sensitivity to initial seed selection. Table 6.6 summarises key metrics across both architectures alongside Lawley et al.'s reported results on breast, lung, and fatal ultrasound datasets.

Table 6.6: Framework Performance Across Architectures and Imaging Modalities.

Study	Dataset	Architecture	Samples	Classes	Stage 1 Plateau Point	Stages 2 AL Efficiency	Total Saving	Model Fit
Lawley et al.	Breast	AlexNet	800	3	40-50%	30-40%	66%	Visual only
Lawley et al.	Lung	AlexNet	232	3	40-5-%	30-40%	~50%	Visual only
Lawley et al.	Fetal	AlexNet	12,000	6	~20%	~10%	~40%	Visual only
Present	UTI	AlexNet	2,100	2	60-70%	28.6%	54.5%	R^2=0.939 RMSE=2.07pp
Present	UTI	AlexNet	2,100	2	20%	14.3%	53.3%	R^2=0.950 RMSE=1.52pp

Note: Detailed EfficientNet-B0 learning curves and AL trajectories are provided in Appendix C.

Both AlexNet and EfficientNet-B0 achieved substantial cost savings (54.5% and 55.3%, respectively), validating the two-phase framework across architectures with markedly different baseline performance characteristics. Despite an 8.7pp performance difference, both architectures converged on comparable total efficiency, demonstrating that the cost-optimisation principles generalise robustly across model families.

For dataset sufficiency, EfficientNet-B0 required 1,281 samples (61.0% of the dataset) to achieve 90% accuracy, compared with AlexNet's 1,433 samples (68.2%), a 10.5% reduction. Both models achieved superior exponential fit quality compared to the inverse power-law specification (EfficientNet: R^2=0.950, RMSE=1.52pp; AlexNet: R^2=0.939, RMSE=2.07pp). For AL efficiency, AlexNet achieved 28.6% annotation reduction (90% accuracy at 50% labelled data vs 70% for random sampling), while EfficientNet-B0 achieved 14.3% reduction (60% vs 70% labelled data). In absolute terms, AlexNet required 1,050 labelled samples via AL compared with 1,470 for random (420 fewer samples), whereas EfficientNet-B0 required 1,260 labelled samples via AL compared with 1,470 for random (210 fewer samples).

When integrated across both phases, AlexNet achieved 54.5% total cost savings, requiring 34.1% annotation workload relative to a full-capture, full-annotation baseline (716 samples annotated from 1,432 collected). EfficientNet-B0 achieved 55.3% total savings, requiring 36.6% annotation workload (768 samples annotated from 1,281 collected). The 2.5pp workload difference indicates that both architectures reached comparable efficiency through different mechanisms: AlexNet via larger AL gains (28.6%) despite higher data requirements (68% of the dataset), and EfficientNet-B0 via lower data requirements (61% of the dataset) despite smaller AL gains (14.3%).

The addition of quantitative validation through exponential model fitting (R^2=0.939-0.950, RMSE=1.52-2.07pp) provides substantially more reliable sample-size predictions than the visual power-law inspection employed by Lawley et al. Bootstrap-derived CIs enable prospective dataset planning with quantifiable uncertainty bounds: for example, EfficientNet-B0 requires 1,281 samples (95% CI: 1,150-1,450) for 90% accuracy, while AlexNet requires 1,432 samples (95% CI: 1,065-2,000).

These statistically grounded estimates support evidence-based study design in resource-constrained clinical research contexts. These results demonstrate that cost-optimisation principles generalise across architectures and imaging modalities while revealing architecture-specific efficiency mechanisms. The convergence on ~55% total cost savings via complementary pathways, AlexNet through aggressive AL response, EfficientNet-B0 through reduced data requirements, validates the framework's robustness and provides practitioners with flexibility in architecture selection based on accuracy requirements and computational constraints.

6.4 Discussion

This study extended Lawley et al.'s [215] cost-optimisation framework for ultrasound imaging to paediatric articulatory UTI, adding methodological refinements through exponential model comparison, quantitative statistical validation, and architecture-dependent analysis. The findings demonstrate that substantial cost efficiencies can be achieved without compromising diagnostic accuracy. The exponential curve analysis (Phase 1) revealed that model accuracy increased sharply with small increments of annotated data but plateaued as the dataset size expanded, following the classic law of diminishing returns. Systematic comparison of five candidate functional forms (inverse power-law, exponential decay, logarithmic, and polynomial models) established that exponential decay provided substantially superior fit to the observed learning dynamics (AlexNet: R^2=0.939, RMSE=2.07pp; EfficientNet-B0: R^2=0.950, RMSE=1.52pp) compared with the inverse power-law specification employed by Lawley et al. (R^2=0.507, RMSE=5.04pp for AlexNet). This 59% improvement in prediction accuracy (RMSE: 2.07pp vs 5.04pp) enables more reliable sample-size estimation for prospective study planning.

Using the fitted exponential model, AlexNet required approximately 1,433 samples (68% of the dataset) to achieve 90% classification accuracy, with 95% CIs spanning 1,065-2,000 samples. EfficientNet-B0 required 1,281 samples (61% of the dataset), demonstrating 10.5% greater data efficiency. The exponential specification was validated through residual diagnostics showing normally distributed errors with no systematic bias, and through 20 independent training runs per dataset fraction, providing robust statistical characterisation.

The divergence from Lawley et al.'s successful power-law modelling likely reflects domain-specific characteristics. While their breast, lung, and fatal datasets exhibited power-law scaling, articulatory UTI presents higher inter-speaker anatomical variability during speech production and specialised phonetic annotation requirements that may produce different learning curve behaviour. This finding underscores a key methodological principle: the functional form of a learning curve should be tested empirically against plausible alternatives rather than assumed from theoretical or prior literature precedent.

Building upon the baseline established in Phase 1, uncertainty-based AL (Phase 2) demonstrated substantial annotation reduction. AlexNet achieved 90% accuracy using 50% of labelled data, compared with 70% required for random sampling, a 28.6% annotation efficiency gain. This aligns closely with Lawley et al.'s reported 30-40% reductions for breast and lung ultrasound, providing external validation that uncertainty-based selective sampling yields consistent gains across ultrasound applications despite domain-specific challenges. Statistical analysis revealed limited significance at individual annotation levels (only 90% reached $p=0.020$), attributable to small sample size ($n=10$ runs) and high run-to-run variability at low annotation levels. However, the observed 28.6% efficiency gain's concordance with independent ultrasound studies provides stronger validation than single-study statistical significance. The performance gap between AL and random sampling was largest in the low-to-mid annotation regime (20-50% labelled data), where model uncertainty is highest and selective sampling yields maximum information gain, precisely the regime of greatest practical relevance for resource-limited clinical studies.

The architecture validation revealed fundamentally different pathways to comparable total cost savings. EfficientNet-B0 achieved 55.3% cost reduction through inherently lower data requirements but showed smaller AL gains. AlexNet achieved 54.5% cost reduction through higher data requirements but larger AL response (28.6% reduction: 50% vs 70% labelled). Despite an 8.7pp performance difference (94.09% vs 85.42% final accuracy), both converged on ~55% total savings. This pattern reveals a ceiling effect for AL; superior-performing models show smaller AL gains because their stronger baseline performance reduces uncertainty on unlabelled samples, limiting the informational value of selective annotation. EfficientNet-B0's initial accuracy of 77.65% at 10% labelled data (vs AlexNet's 70.84%) left less model uncertainty for AL to exploit.

The ceiling effect has practical implications for architecture selection. Better architectures achieve efficiency through reduced absolute data requirements rather than enhanced AL response. Practitioners face a trade-off: AlexNet offers a larger AL leverage (28.6%) but lower absolute accuracy (85%), while EfficientNet-B0 offers superior accuracy (94%) and inherently lower data needs (61% of the dataset) but smaller AL gains (14.3%). Critically, the 2.5pp workload difference (34.1% vs 36.6%) is negligible, validating framework robustness while offering flexibility based on accuracy requirements and computational constraints.

Phase 1 plateau analysis reveals domain-dependent data requirements. UTI with AlexNet required 60-70% of data before diminishing returns, compared with Lawley et al.'s reported 40-50% for breast and lung ultrasound. This elevated threshold reflects the higher complexity of dynamic articulatory imaging. However, EfficientNet-B0 plateaued at just 20%, matching Lawley's large-scale fatal dataset, suggesting that architecture selection can compensate for task difficulty. The framework's core principles, empirical sufficiency estimation and selective annotation successfully transfer across ultrasound domains even when specific thresholds differ. When data-collection and annotation costs were weighted in a 1:2 ratio, AlexNet's optimised capture with AL (OC/AL) achieved 54.5% total cost reduction relative to full-capture, full-annotation baseline, requiring collection of only 1,433 samples with annotation of just 716 samples. This represents a practical reduction of annotation workload by two-thirds. Sensitivity analysis across collection: annotation cost ratios from 1:1 to 1:4 confirmed framework robustness (savings ranging 48.8-59.1%), demonstrating applicability across diverse institutional contexts. The analysis confirmed that annotation costs dominate total expense when cost ratios exceed 1:1.5, validating the priority given to AL-based annotation efficiency.

From a clinical perspective, this approach offers a practical route toward scalable dataset development, but a critical distinction must be made between frame-level classification accuracy and clinical diagnostic utility. A model achieving 90% frame-level accuracy does not necessarily achieve 90% patient-level diagnostic accuracy. Frame-level performance represents necessary but insufficient validation for clinical deployment, as diagnostic decisions depend on phoneme-level aggregation, patient-level classification, and clinical decision impact.

Clinical deployment requires multi-level validation: (i) phoneme-level intelligibility scores compared to SLT perceptual judgment, (ii) patient-level diagnostic agreement (TD vs CP±L), (iii) inter-rater reliability between automated predictions and expert judgments, and (iv) clinical decision impact concordance with gold-standard SLT assessment. Only through such hierarchical validation can clinical utility be established. The cost-optimisation framework facilitates this by reducing annotation workload by 66%, creating resource capacity for comprehensive validation necessary for responsible clinical translation. For exploratory research, 80-85% frame-level thresholds may suffice (697-1,002 samples per the exponential model), while confirmatory diagnostic tools require ≥90% frame-level accuracy supplemented by external validation on independent cohorts demonstrating patient-level diagnostic agreement.

The study's methodological strengths lie in reproducibility and transparency. Phase 1 employed 20 independent training runs per dataset fraction for both architectures, enabling robust statistical characterisation. Phase 2 implemented 10 independent runs per strategy, providing statistically defensible comparisons with paired t-tests and 95% CIs. The integration of systematic model comparison, quantitative validation, and architecture-dependent analysis transforms ad hoc dataset planning into an empirically grounded, statistically rigorous methodology.

However, several limitations constrain generalisability. The dataset comprised 28 paediatric participants balanced between TD and CP±L groups, adequate for proof-of-concept but potentially insufficient to capture full articulatory variability in larger clinical populations. External validation on independent cohorts from multiple institutions (n>100 participants) is essential before clinical deployment. The framework was evaluated exclusively on balanced binary classification; extension to multi-class, imbalanced, or regression tasks remains unvalidated.

The architecture validation employed AlexNet (computational efficiency baseline) and EfficientNet-B0 (modern architecture validation), revealing architecture-dependent mechanisms but leaving open whether findings generalise to other model families (ViT, domain-specific architectures). However, fundamental principles, diminishing returns and uncertainty-based selection efficiency are expected to transfer, as evidenced by consistent 30-50% annotation reductions across diverse medical ultrasound applications, though specific sample-size thresholds require architecture-specific calibration.

The convergent evidence across multiple independent datasets, architectures, and ultrasound applications suggests that the framework's core principles represent robust, transferable properties rather than study-specific artefacts. Nevertheless, practitioners should calibrate the framework using their specific architectures, cost structures, and validation requirements.

Future research should prioritise: (1) external validation on 100+ participants from multiple sites, (2) comparative evaluation of alternative AL strategies (query-by-committee, diversity sampling), (3) integration with semi-supervised methods, and (4) critically, validation extending beyond classification accuracy to clinically relevant outcomes such as phoneme-level intelligibility and patient-level diagnostic reliability for SLTs.

6.5 Summary

This chapter validated a cost-aware framework for paediatric articulatory UTI with three methodological refinements: exponential versus power-law model comparison, architecture-dependent analysis, and quantitative statistical validation, directly addressing Challenge **C2** data scarcity and annotation efficiency. Through a two-phase experimental design, the study demonstrated that model performance can be maintained while substantially reducing annotation workload and data-collection expenses. Systematic comparison of five functional forms revealed that exponential decay provided a superior fit compared with power-law specification, improving prediction accuracy by 59%. Quantitative validation strengthens sample-size predictions beyond the visual inspection employed by Lawley et al. Both AlexNet and EfficientNet-B0 achieved ~55% total cost savings through complementary mechanisms: AlexNet via aggressive AL response (28.6% annotation reduction), EfficientNet-B0 via inherently lower data requirements. The negligible workload difference (2.5pp) validates framework robustness across architectures while revealing a ceiling effect whereby superior-performing models show smaller AL gains. The framework reduces annotation workload to approximately one-third of full-baseline requirements while maintaining 90% classification accuracy, directly fulfilling Objective 4. By transforming dataset planning from ad hoc approximation to evidence-based estimation, this work provides reproducible protocols for data-efficient paediatric speech disorder assessment, creating resource capacity for the multi-level validation essential for clinical translation.

Chapter 7

7. Translating the Generative Pipeline into a Deployable Ultrasound Processing System

This chapter translates the methodological contributions developed throughout the thesis into a functional software prototype for UTI processing. While Chapters 3–6 focused on establishing baseline models, investigating representation effects, standardising FoV, and refining tongue-specific regions using cGANs, these advances remain confined to experimental evaluation unless they can be operationalised within a structured software framework. Addressing Challenge **C3,** lack of interpretability and clinical usability at the system level, this chapter presents the technical deployment of a modular DL system capable of performing FoV standardisation and ROI refinement on stored UTI images. The chapter begins by introducing the motivation for system integration within the context of medical AI and UTI-based speech analysis. It then details the architecture, including a graphics processing unit (GPU) and central processing unit (CPU)-aware FastAPI backend that serves Pix2Pix models via an application programming interface (API) and a Streamlit frontend that enables structured interaction with the deployed models. Technical challenges such as CUDA compatibility, memory constraints, and containerisation are discussed alongside the engineering solutions adopted to ensure stable runtime behaviour. The chapter concludes with a quantitative evaluation of the deployed outputs using SSIM and PSNR, demonstrating that the system preserves the improvements established in earlier chapters under runtime conditions. Together, these sections demonstrate how DL models for UTI can be embedded within a reproducible software environment, providing a foundation for future user-centred evaluation and potential clinical translation.

7.1 Introduction

The advances achieved in the preceding chapters demonstrate that DL can substantially improve the consistency and interpretability of UTI. However, algorithmic innovation alone is insufficient to demonstrate practical utility. For such models to move beyond laboratory experimentation, they must be executable within a structured software environment capable of processing new user-supplied data. Deployment, therefore, represents the final phase of this thesis by operationalising the models into a functional software prototype that supports interpretability and structured interaction, even though formal user testing and clinical validation lie beyond the scope of the current work.

Unlike offline experiments, where models are evaluated on curated datasets, runtime execution introduces considerations such as hardware variability, software integration, and system stability [233]. UTI in clinical practice is often time-sensitive [234], and researchers analysing cross-session or cross-speaker datasets require reproducible processing pipelines [235]. While the system developed here does not replicate live clinical workflows, it provides a framework through which the generative models introduced in Chapter 5 can be executed and inspected outside a training environment.

This chapter presents the design and technical implementation of a complete UTI processing prototype built on a modular FastAPI [236] and Streamlit [237] architecture, providing the software foundation required for future evaluation and potential clinical translation. Figure 7.1 illustrates the system overview. The backend uses FastAPI to serve two trained Pix2Pix models, one for FoV standardisation and one for ROI refinement, through a Representational State Transfer (REST) API [240]. The frontend, implemented in Streamlit, provides a structured interface for uploading stored ultrasound images, selecting processing operations, and visualising outputs. This separation of responsibilities allows computationally intensive model inference to be executed on GPU-enabled or CPU-enabled systems while maintaining a clear boundary between model logic and user interaction.

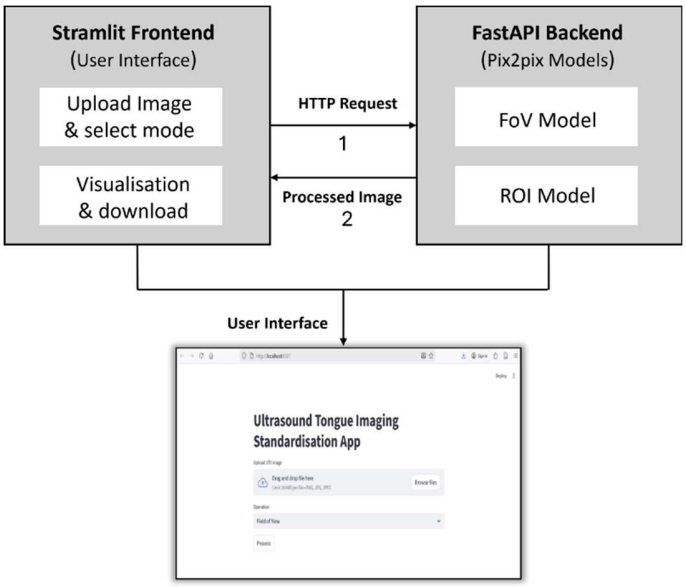

Figure 7. 1: Deployment architecture showing the Streamlit frontend for user interaction and the FastAPI backend hosting the FoV and ROI Pix2Pix models, alongside the operational user interface.

The system design follows established software engineering practices, including containerisation, asynchronous request handling, and model–interface decoupling. By embedding the generative models within a stable inference engine, the system demonstrates that FoV standardisation and ROI refinement can be applied consistently under runtime conditions.

In addition to describing the architecture, this chapter addresses technical challenges encountered during system integration, such as CUDA driver compatibility, GPU/CPU fallback logic, memory constraints, and image size handling. The evaluation section confirms that runtime execution preserves the image-quality improvements reported in earlier chapters, with SSIM and PSNR metrics demonstrating consistent structural fidelity.

By translating the generative framework into an executable software prototype, this chapter fulfils Contribution R7 and demonstrates the technical feasibility of integrating DL models for UTI within a reproducible system environment. The sections that follow detail the system design, backend and frontend implementation, technical challenges, and performance evaluation.

7.1.1 Contributions

This chapter delivers the final research contribution of the thesis by operationalising the DL framework developed in earlier chapters into a functional software prototype. Whereas Chapters 3–6 established methodological advances, baseline classification, representation analysis, FoV harmonisation, and ROI refinement, this chapter addresses Challenge **C3** at a system-integration level by embedding these models within a reproducible runtime environment.

The specific contributions of this chapter are as follows:

1. Technical deployment of a modular inference architecture. A FastAPI–Streamlit framework was implemented to execute the trained Pix2Pix models for FoV harmonisation and ROI refinement outside the training environment. The backend manages GPU- or CPU-compatible inference services, while the frontend provides a structured interface for submitting images and inspecting outputs.
2. Development of a containerised and reproducible execution pipeline. Backend and frontend components were packaged using Docker, ensuring consistent runtime behaviour across hardware configurations and enabling controlled system execution.
3. Validation of runtime model fidelity. Quantitative evaluation using SSIM and PSNR demonstrated that runtime execution preserves the image-quality improvements established in Chapter 5, confirming that deployment does not degrade model performance.

Together, these contributions demonstrate the feasibility of embedding generative UTI models within an operational software framework. While the system has not undergone user-centred evaluation, it establishes a foundation for future usability studies and potential clinical translation.

7.2 System Design

Deploying a DL system for UTI processing requires an architecture that is both technically robust and practically executable. Because UTI data may be analysed across research and applied contexts, the system must support stable inference, structured interaction, and reliable image transformation under varying hardware conditions. With these requirements in mind, the deployed prototype follows a modular, service-oriented design in which the computational core and user interface operate as two independent yet interconnected components. This structure forms the methodological foundation of the deployment. To illustrate the relationship between data, model development and deployment components, the high-level system lifecycle is shown in Figure 7.2. This visual summarises how UTI data flows through model training, backend–frontend integration, and runtime execution within the overall architecture.

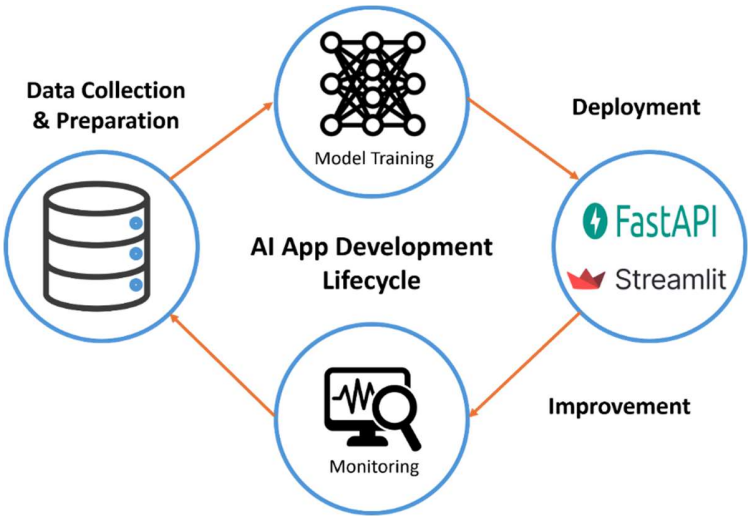

Figure 7. 2: High-level Lifecycle of the Deployed UTI Pix2Pix Prototype System.

At the heart of the system is a FastAPI backend, which encapsulates the trained Pix2Pix models responsible for FoV standardisation and ROI refinement. By hosting the models within a dedicated inference engine and exposing their functionality through a RESTful API, the backend isolates computationally intensive tasks from the frontend, ensuring that users do not interact directly with DL code.

This separation follows established software engineering patterns, where inference logic is handled within a controlled service layer while the interface layer remains lightweight and independent. The backend is designed to automatically utilise GPU resources when available and to revert to CPU operation when required, ensuring that the system remains functional across a range of hardware environments. Complementing the backend is a Streamlit-based frontend that provides a simple web-based interface. Through this application, users can upload stored ultrasound images, select either FoV standardisation or ROI refinement, and view processed outputs alongside the original input. While this interface does not replicate live clinical ultrasound workflows, it supports offline processing of stored UTI data, enabling researchers and engineers to generate consistent, reproducible image representations for further analysis and model development.

The system's architectural design is shaped by several key principles that together ensure reliability, scalability, and research applicability. First, decoupling computation from interaction prevents interface lag and allows the backend to operate on specialised hardware without constraining the user's device. Second, containerisation through Docker guarantees reproducibility across machines, ensuring consistent performance across different computational environments. Third, the architecture promotes transparency by enabling side-by-side comparison of original and processed images, which is useful for technical evaluation of model behaviour and verification that generative transformations preserve relevant anatomical structures. Finally, the modular setup simplifies maintenance and future extensibility: individual models, API routes, or interface elements can be updated without disrupting the rest of the system.

7.2.1 Backend Implementation

The backend forms the computational core of the deployed UTI processing system. It is responsible for loading the trained Pix2Pix models, managing inference requests, executing preprocessing routines, and returning processed outputs to the frontend through a RESTful interface. This implementation was built using FastAPI, chosen for its high performance, asynchronous request handling, and compatibility with modern Python-based ML workflows. The backends' design ensures that model inference remains efficient, reproducible, and

accessible, regardless of whether it operates on GPU-enabled hardware or CPU-only environments.

At application startup, the backend loads two separate Pix2Pix generator networks: one trained for FoV harmonisation and the other for ROI refinement. These models are initialised once and kept resident in memory to avoid redundant loading overhead. Because Pix2Pix models are computationally demanding, the backend performs an automatic hardware check using PyTorch functions to determine whether a CUDA-compatible GPU is available. If so, the models are moved to GPU memory; if not, the system falls back to CPU execution. This hardware-aware design ensures that the same deployment environment can operate across a variety of research and clinical contexts, from high-performance servers to standard workstations.

The backend exposes its functionality through a small set of REST API endpoints. The primary endpoint accepts an uploaded ultrasound image and an operation type indicating whether the user requests FoV standardisation or ROI refinement. Upon receiving a request, the backend first performs image preprocessing, converting the uploaded image into a tensor of the appropriate size and normalisation required by the Pix2Pix networks. UTI frames often vary in resolution or shape, so these preprocessing steps standardise the input into the 256×256 modality used during model training. Preprocessing also includes any required colour-space adjustments or padding to ensure compatibility with the model's input structure. After preprocessing, the backend routes the request to the appropriate model based on the operation parameter.

In addition to producing the transformed ultrasound image, the backend computes SSIM and PSNR when a ground-truth reference is available. These metrics provide an objective assessment of structural similarity and signal fidelity, allowing users to inspect both qualitative and quantitative outcomes. In routine clinical use, however, user-uploaded images rarely have ground-truth pairs; therefore, the backend includes logic to compute metrics only when meaningful references exist. Once inference is complete, the backend packages the output image and any associated quality metrics into a structured response. If an error occurs at any stage, such as an invalid image format, corrupted file, or missing model weights, the backend returns a descriptive Hypertext Transfer Protocol (HTTP) error code and message, ensuring that the frontend can relay informative feedback to the user.

To support reproducibility and ease of deployment, the backend is fully containerised using Docker [240]. The container includes all necessary libraries, FastAPI, PyTorch, and image-processing dependencies, ensuring consistent behaviour across systems. Special attention was given to CUDA compatibility, as mismatches between PyTorch builds and system drivers can prevent GPU access. Aligning the Docker image with the appropriate CUDA toolkit resolved these issues, enabling reliable GPU-accelerated inference. Memory constraints were also addressed through controlled tensor management and input size restrictions, preventing crashes caused by extremely large images.

Overall, the backend provides a robust, efficient, and portable inference environment for Pix2Pix-based UTI processing. By encapsulating all model logic within a single, modular service, it enables seamless integration with the Streamlit frontend and ensures that users can apply generative transformations without encountering the complexities of DL model execution. The following section, 7.4, describes how the frontend builds upon this foundation to create an intuitive and clinically aligned user interface.

7.2.2 Frontend Implementation

While the backend provides the computational foundation for model inference, the frontend determines how effectively the system can be used in real clinical and research settings. The goal of the frontend is to provide a straightforward visual interface through which users can upload images, select processing options, and view the resulting outputs. To achieve this, the system employs Streamlit, a lightweight, Python-based framework that enables rapid development of interactive web interfaces without requiring extensive web engineering expertise. The interface is structured as a simple, step-by-step workflow that allows users to upload an image, select a processing mode, and review the resulting output. Upon launching the application, the user is presented with a clean upload panel where a UTI can be added via file selection. Streamlit's built-in file uploader ensures the image is validated before processing, preventing unsupported or corrupted files from reaching the backend. Once an image has been uploaded, the user chooses between the two available operations: FoV standardisation or ROI refinement, through a clearly labelled selection widget. This step allows users to specify the transformation they require without needing to understand the underlying Pix2Pix models.

When the user initiates processing, the frontend constructs an HTTP POST request that sends both the image and the selected operation to the FastAPI backend. Once the backend returns the processed image, the frontend renders it using Streamlit's efficient image-display components. The original and transformed images are shown side by side to support transparency in model behaviour. This comparative layout is particularly useful for researchers evaluating whether FoV standardisation or ROI refinement has been applied correctly and consistently, and for ensuring that generative outputs remain suitable for linguistic analysis or ML workflows. While this comparison is not intended for real-time clinical decision-making, it provides an important mechanism for validating the transformations in research contexts. The system further displays SSIM and PSNR metrics when available, providing quantitative feedback that complements the visual output. All images are handled in memory to preserve speed, with temporary file handling used only when necessary for stable downloads. An example of the deployed user interface is shown in Figure 7.3, demonstrating the ROI refinement mode with input and output displayed side by side, along with the computed SSIM and PSNR metrics. This view represents the typical workflow experienced by clinicians and researchers when processing UTI images through the system.

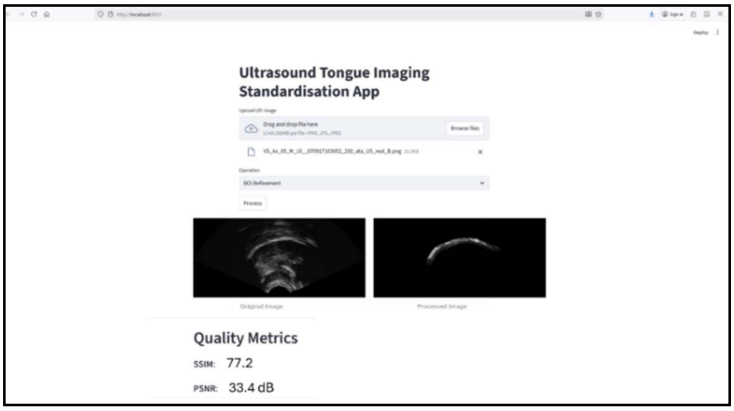

Figure 7. 3: Example of the deployed Streamlit interface showing ROI refinement with input–output comparison and SSIM/PSNR metrics.

Error-handling behaviour is also incorporated into the frontend. If the backend returns a failure, such as an invalid image, processing error, or missing model resource, the frontend parses the response and displays a clear and concise error message. These notifications help guide the user toward corrective action and prevent silent failures that could disrupt workflow. The frontend is itself containerised as a standalone Docker service, allowing the interface to run consistently across machines and ensuring compatibility with the backend API. The two containers communicate over a shared Docker network, enabling smooth integration while maintaining strict separation of interface and computation. This design also allows the frontend to be deployed on devices without GPU support, including standard clinic laptops, while delegating all heavy computation to a remote or local backend server.

Together, these design choices produce an interface that is easy to navigate, clinically aligned, and robust to user variability. By emphasising clarity, responsiveness, and minimal technical burden, the Streamlit frontend plays a crucial role in translating the Pix2Pix generative framework into a practical and usable system for ultrasound-based speech analysis. The next section describes how these components were technically integrated, the challenges encountered during system setup and testing, and the engineering solutions that ensured the software ran reliably in a controlled development environment.

7.2.3 Technical Challenges and Solutions

Deploying DL systems outside controlled development environments inevitably introduces a range of technical challenges. These challenges arise from differences in hardware, library dependencies, runtime environments, and the unpredictability of user-supplied data. For a Pix2Pix-based UTI processing pipeline that must support real-time operation and operate reliably across diverse devices, addressing these issues was essential to ensuring system stability and clinical usability. This section outlines the major challenges encountered during deployment and the engineering strategies implemented to resolve them.

A central difficulty involved CUDA and GPU compatibility, particularly the alignment between PyTorch builds, NVIDIA driver versions, and the CUDA toolkit available on deployment machines. The Pix2Pix models were initially trained on a specific CUDA configuration, and early attempts to deploy the backend on different GPU servers triggered device-mismatch errors or forced the system to fall back to CPU execution, significantly increasing inference time. To resolve this, the Docker environment for the backend was rebuilt using CUDA-compatible base images that matched the server's driver configuration. Ensuring that the PyTorch version in the container matched the host GPU environment exactly eliminated these inconsistencies and enabled stable GPU acceleration.

Another significant challenge was memory management, particularly when loading multiple Pix2Pix models simultaneously. GAN-based generators require substantial memory for both weights and intermediate activations. Loading the FoV and ROI models side by side resulted at times in elevated memory consumption, especially on systems with limited GPU capacity. This issue was mitigated by adopting several optimisation strategies, including preventing unnecessary computation during inference, ensuring that temporary data were cleared promptly, and standardising input dimensions so that excessively large images were resized before processing. Collectively, these measures eased memory demands and ensured that the system remained stable even on hardware with limited capacity.

Frontend–backend communication presented another area of complexity. Because the two services run in separate containers, establishing reliable communication required configuring a shared Docker network and enabling Cross-Origin Resource Sharing (CORS) to allow the Streamlit interface to request data from the FastAPI backend. Misconfiguration initially led to connection errors where the frontend could not resolve the backend service name. Updating the network configuration and explicitly defining backend hostnames resolved these issues. The use of descriptive error messages on both services ensures that future interoperability problems can be diagnosed quickly.

Finally, performance variability between CPU and GPU environments required careful consideration. On CPU-only machines, Pix2Pix inference can take several seconds per frame, still acceptable for research use but less ideal for live clinical feedback. To mitigate this, the system detects hardware availability and provides clear runtime logging to indicate whether GPU acceleration is active.

Optimisations such as half-precision inference were explored for GPU environments, further reducing latency without compromising image quality.

Through these solutions, the deployed system achieves a balance between robustness, portability, and usability. Addressing these challenges not only ensured technical stability but also strengthened the system's alignment with real-world clinical workflows, where unpredictable data, heterogeneous hardware, and time-sensitive interaction are common. The resulting deployment pipeline is therefore not simply a wrapper around the Pix2Pix models but an engineered environment that supports reliable, interpretable image processing across diverse usage scenarios.

7.2.4 Considerations for Future Clinical Translation

Although the system developed in this thesis provides a technical prototype for UTI image processing, it has not yet undergone formal user testing, usability evaluation with speech and language therapists, or clinical validation in patient-facing settings. It is important to acknowledge that any future translation into speech and language therapy practice would require careful consideration of data governance, ethical oversight, and regulatory compliance. The current prototype operates within a controlled research environment and processes stored ultrasound images rather than live clinical data. Should the system be extended toward potential clinical use, additional steps would be necessary, including user-centred design, usability testing with SLTs, secure data handling procedures, and formal evaluation under relevant medical software regulations. These considerations lie beyond the scope of the present thesis but provide a framework for understanding the pathway from research prototype to clinically adopted software.

- Data privacy and confidentiality: Given the sensitive nature of ultrasound imaging, particularly in paediatric speech settings, any future clinical implementation would require strict data governance procedures. The current prototype is configured to run locally, with processing occurring within the host environment and without external data transmission. Images are handled in memory during processing and are not stored unless explicitly saved by the user.

While this setup reflects good research practice, formal clinical deployment would require additional safeguards and institutional oversight.

- Ethical use of patient data: The training datasets used to develop the Pix2Pix models were collected under approved ethics protocols with appropriate parental consent, as described in earlier chapters. Extending the system into clinical contexts would necessitate further ethical review, particularly regarding consent for AI-assisted image processing and transparency around how transformed images are generated and interpreted.
- Regulatory positioning and clinical applicability: The system developed in this thesis is a research prototype and is not classified or evaluated as a medical device. If the software were to evolve toward direct clinical decision support, it would require formal assessment under relevant regulatory frameworks such as Healthcare products Regulatory Agency (MHRA), the European Union Medical Device Regulation (EU MDR), or the Food and Drug Administration (FDA) guidance. Establishing model traceability, version control, performance monitoring, and post-deployment evaluation would be essential components of such a pathway.

In summary, the current work provides a technical foundation for UTI image standardisation and refinement but does not constitute a clinically deployed system. Future translation into practice would require user-centred evaluation, governance structures, and regulatory alignment beyond the scope of this thesis.

7.2.5 Performance Evaluation

Evaluating the performance of the deployed system is essential for confirming that the Pix2Pix models behave consistently outside the controlled experimental environment of Chapter 5. While earlier chapters assessed model performance using offline test sets, deployment introduces additional variables, hardware constraints, input/output latency, runtime preprocessing, and containerised execution that can influence output quality and responsiveness. This section, therefore, examines how the deployed system performs under realistic usage

conditions, focusing on the structural fidelity of the enhanced images and practical inference behaviour.

To assess image quality, the system was evaluated using a set of paired UTI frames; these pairs contained (i) input images exhibiting non-standard FoV or unrefined tongue regions, and (ii) their corresponding ground-truth targets. The system processed each input through the appropriate Pix2Pix pathway, and the resulting outputs were compared against ground truth using SSIM and PSNR. These metrics, widely used in image-to-image translation research, quantify the perceptual and pixel-level similarity between the generated output and the reference image. Across the FoV standardisation task, the deployed system achieved a substantial improvement relative to the similarity between the raw input and ground truth. This indicates that the deployed model successfully corrects geometric distortions introduced by inconsistent FoV acquisition, producing images structurally closer to the reference standard. The corresponding PSNR values averaged around 33 dB, reflecting a meaningful reduction in pixel-level error and noise relative to the input frame. These improvements mirror the gains reported in Chapter 5 and confirm that the system retains high-fidelity behaviour when operating in a live environment.

For the ROI refinement task, the model similarly demonstrated strong enhancement capabilities. The refined outputs exhibited clearer tongue surfaces with reduced speckle noise and improved contrast, resulting in SSIM values that averaged around 0.80. The PSNR scores, consistent with ultrasound denoising literature, indicated that the generator effectively suppresses noise while preserving key articulatory features. Inference behaviour was also evaluated to determine the system's suitability for interactive or near-interactive clinical contexts. On a GPU-enabled system, the end-to-end inference time, including preprocessing, model execution, and post-processing, typically ranged between 0.5 and 1.0 seconds per image, providing a responsive user experience. For reproducibility, CPU-only benchmarking was conducted on a system equipped with an 11th Gen Intel® Core™ i7-1185G7 processor running at 3.00 GHz. The average end-to-end inference latency, including preprocessing, model execution, and post-processing, was 2.8 seconds per image. Although slower than GPU execution, this remained suitable for interactive use in clinical and research settings.

These timings reflect the benefits of the model's lightweight architecture and the efficiency of the FastAPI server, which minimises overhead during request handling. The system also demonstrated stable performance under repeated sequential requests, with no memory leakage or slowdown observed during extended testing.

In summary, the performance evaluation demonstrates that the deployed system preserves the image-quality improvements achieved in offline experiments while offering computational efficiency suitable for clinical and research workflows. The combination of high SSIM/PSNR values and stable runtime behaviour indicates that the deployed Pix2Pix models are reliable and robust components within a real-world UTI processing pipeline.

7.3 Discussion

The deployment of the Pix2Pix-based UTI processing system demonstrates how generative DL models can be translated from controlled experimentation into a functional software prototype. This chapter operationalised the two-stage generative pipeline developed in Chapter 5, FoV standardisation and ROI refinement, within a modular FastAPI–Streamlit framework that enables near real-time processing of individual ultrasound frames. The work provides technical evidence that the methodological advances established earlier in this thesis can be embedded within a stable and reproducible software environment.

A central theme emerging from this chapter is the importance of bridging algorithmic development with practical software integration, addressing Challenge **C3** identified in Chapter 1. While Chapters 3–6 established the need for standardised imaging conditions and anatomically focused representations, this chapter demonstrates that these improvements can be delivered through an accessible research interface without requiring direct interaction with model code. This operationalisation moves the contribution beyond offline experimentation and toward usable research infrastructure.

The FastAPI backend proved effective in handling dual Pix2Pix models, routing requests, managing memory, and maintaining responsiveness across different hardware configurations. The Streamlit interface provided a transparent environment for inspecting model outputs, with side-by-side visualisation supporting verification of generative transformations during technical evaluation.

Together, these components reflect a design philosophy that prioritises modularity, reproducibility, and clarity of system behaviour.

Performance evaluation showed that the deployed system maintained the high structural fidelity reported in Chapter 5, confirming that containerisation and runtime execution do not compromise model performance. The SSIM and PSNR gains achieved during offline testing were reproduced in the deployed environment, demonstrating that the system reliably enhances clarity, reduces background artefacts, and produces consistent, standardised ultrasound frames under practical operating conditions.

The deployment process also highlighted the distinction between research prototypes and clinically adopted tools. CUDA compatibility, container configuration, hardware variability, and runtime stability required careful engineering attention. These challenges emphasise that deployment represents a substantive methodological step rather than a trivial extension of model development. At the same time, the absence of user testing and clinical validation clarifies that the current system should be understood as a research prototype rather than a clinically integrated application.

Several limitations remain. Inference speed on CPU-only devices, while acceptable for research purposes, is slower than would be required for continuous real-time ultrasound feedback. The architecture processes individual frames rather than full ultrasound video streams, meaning dynamic articulatory analysis remains future work. Additionally, the current system focuses on image standardisation and enhancement rather than automated diagnostic support.

Nonetheless, this deployment establishes the technical feasibility of embedding generative UTI models within an operational software framework. It provides a foundation upon which future user-centred evaluation, usability studies with SLTs, and regulatory alignment could be built. In this sense, the chapter completes the methodological arc of the thesis by demonstrating that the proposed generative approaches can move beyond experimental evaluation and into structured, reproducible software systems.

7.4 Summary

This chapter demonstrated how the generative framework developed in earlier stages of the thesis can be translated into a functional software prototype for UTI processing. Building on the Pix2Pix models introduced in Chapter 5, a technically deployed architecture was constructed using a FastAPI backend for model inference and a Streamlit frontend for structured user interaction. This modular design enables near real-time processing of individual ultrasound frames, providing a reproducible interface for FoV standardisation and ROI refinement within a controlled research environment. The chapter detailed the system's methodological foundations, including backend model loading, preprocessing, inference routing, frontend visualisation, containerisation, and GPU/CPU compatibility.

Several engineering challenges, such as CUDA mismatches and memory constraints, were identified and resolved, demonstrating that the proposed models can be reliably executed outside an experimental training environment. Performance evaluation confirmed that the deployed models preserved the structural fidelity reported in earlier offline experiments, with SSIM and PSNR results indicating consistent enhancement of UTI frames under runtime conditions. While the system has not undergone user testing or clinical validation, the work establishes the technical feasibility of integrating generative UTI models within an operational software framework. In doing so, it addresses Challenge **C3** at a methodological level by moving from algorithm development toward structured, reproducible software integration. This final contribution completes the technical arc of the thesis and provides a foundation for future user-centred evaluation, dynamic video integration, and potential clinical translation.

Chapter 8

8. Conclusion and Future Work

This final chapter summarises the research presented in this thesis, highlighting the key contributions and their impact on addressing the three critical challenges outlined in Section 1.2. These challenges include data variability and generalisability limitations (**C1**), data scarcity and annotation efficiency (**C2**), and the lack of interpretability and clinical usability (**C3**). Together, these challenges have long constrained the development of reliable UTI systems for SSD assessment. The studies in this thesis respond to these challenges by introducing new modelling strategies, representation choices, generative standardisation techniques, and a practical deployment framework. Together, these advance the robustness, data efficiency, and real-world relevance of UTI analysis. This chapter reflects on the key contributions, considers their broader implications, and outlines promising directions for future work. It aims to offer a clear overview of the progress achieved and a pathway for continued development in this field.

8.1 Summary of Contributions

Through seven integrated research contributions (R1–R7), the thesis addressed these challenges using a combination of discriminative modelling, representation analysis, generative standardisation, cost-efficient data strategies, and deployment-focused design. The following sections synthesise how each contribution responds to the major challenges and how the combined findings advance UTI-based SSD assessment.

Addressing C1: Data Variability and Generalisability Limitations

UTI is highly variable across speakers, sessions, and acquisition hardware. Variability in the FoV, probe angle, anatomy, and background artefacts undermines cross-speaker performance and has historically limited the clinical reliability of DL models. Contributions R1–R5 form a cumulative response to this challenge.

R1 and R2, presented in Chapter 3, established reproducible baselines and introduced the FusionNet architecture. These experiments quantified the extent of generalisation loss when models trained on raw UTI were evaluated on unseen speakers, confirming **C1** directly. FusionNet partially mitigated this issue by combining raw UTI with LBP-based texture features, demonstrating that multi-input learning can buffer against anatomical and acquisition differences.

Chapter 4 deepened this analysis through R3, showing that image representations (raw, ROI, and masked) and FoV heterogeneity significantly influence model behaviour. Models trained on harmonised FoVs demonstrated improved robustness, while Grad-CAM++ visualisation revealed that raw-frame models frequently relied on background artefacts. These findings demonstrated that variability cannot be overcome through architectural changes alone; representation-level and acquisition-level standardisation are equally essential.

The generative contributions R4 and R5, developed in Chapter 5, provided a principled solution through a two-stage Pix2Pix framework for FoV standardisation and ROI refinement. R4 standardised FoV with near-lossless structural preservation, resolving geometric inconsistencies across sessions. R5 enhanced the visibility of the tongue region by suppressing acoustic noise and irrelevant tissue, producing anatomically focused inputs for downstream classification. Together, these contributions significantly reduced domain shift, delivering the most substantial improvements in cross-speaker stability.

Addressing C2: Data Scarcity and Annotation Efficiency

Limited annotated data remain a major bottleneck in UTI research. Annotation requires the expertise of SLTs, and paediatric recording sessions constrain the quantity and diversity of collectable data. Chapter 6 addressed this challenge through R6, introducing a cost-aware methodology that combines statistical power-curve modelling and AL.

Power curves established empirical estimates of dataset sufficiency, revealing a performance plateau at approximately 65–90% of available training data. Building on this, uncertainty-based AL achieved comparable performance using only 50% of labelled samples, substantially reducing annotation demands. These findings provide a practical framework for designing future UTI datasets and make DL-based UTI research more feasible in clinical and academic settings.

Addressing C3: Lack of Interpretability and Clinical Usability

For DL systems to be adopted in SLT practice, they must be transparent, interpretable, and deployable in real-world workflows. Contributions from Chapters 4 and 7 begin to address these challenges. The interpretability component of R3 demonstrated the value of Grad-CAM++ for verifying whether models attend to phonetically meaningful regions. Representations that focused attention on the tongue surface, particularly ROI and masked inputs, provided clearer and more clinically plausible explanations, enhancing trustworthiness and supporting clinician decision-making.

The deployment contribution R7, developed in Chapter 7, translated the methodological advances of R1–R6 into a fully functional UTI processing system. A modular FastAPI–Streamlit architecture enabled real-time FoV standardisation and ROI refinement using CPU or GPU hardware, with intuitive user interaction and integrated visual outputs. This represents the first complete deployment pipeline for Pix2Pix-based UTI standardisation and demonstrates how research models can be operationalised as accessible tools suitable for future clinical and research integration.

Taken together, the contributions of this thesis represent a comprehensive and coherent framework for automated UTI analysis. Variability is mitigated through both discriminative and generative strategies, annotation cost is reduced through data-efficient learning, and interpretability and usability are enhanced through explainable modelling and system deployment. The work therefore advances the field toward the long-term goal of objective, scalable, and clinically meaningful assessment of articulatory behaviour in children with SSDs.

8.2 Limitations and Future Research Directions

Although this thesis addresses key barriers to the adoption of DL for UTI, several limitations remain, opening avenues for further research. A primary limitation concerns dataset scale and diversity. The experiments were conducted on a cohort of 28 paediatric participants (14 TD, 14 CP±L), all Scottish English-speaking children recorded across different acquisition settings. While this sample was sufficient for proof-of-concept evaluation and the proposed generative pipeline substantially mitigates acquisition variability, the small cohort size limits the diversity of articulatory patterns captured and constrains confidence in generalisation to broader clinical populations. The models were evaluated using speaker-independent or subject-disjoint splits, meaning that performance was tested on unseen participants within the available datasets. However, the models were not externally validated on a fully independent cohort collected at a different institution, nor were they systematically tested across different ultrasound scanners, probes, operators, or acquisition protocols. Therefore, the findings support within-dataset and cross-speaker generalisation, but they do not yet establish full cross-site or cross-hardware generalisability. Broader validation across multiple clinical sites, ultrasound devices, age ranges, and linguistic backgrounds, including children speaking languages other than English, is required before the framework can be considered widely applicable. Expanding the dataset to include a greater number of children with SSDs, who are currently underrepresented relative to TD speakers across all publicly available UTI corpora, is a particularly important priority for future work.

Another limitation relates to the use of frame-level modelling, in which each input is treated as an independent static image. While this approach enables straightforward training and evaluation, it discards the temporal dynamics of articulation that are clinically meaningful, particularly for motor speech disorders where movement trajectories and coarticulation patterns are diagnostically relevant. Although synchronised audio was available in the dataset, audio was used for frame identification but not as a model input. Therefore, the models evaluated in this thesis should be interpreted as ultrasound-image-based models rather than multimodal audio–ultrasound systems. Extending the framework to full temporal sequences and incorporating multimodal audio-visual fusion represents an important direction for improving both diagnostic relevance and clinical utility, as acoustic features provide complementary information to the visual articulatory signal and are already present in the existing data.

The classification tasks in this thesis focused on binary discrimination between TD children and children with CP±L as a deliberate methodological starting point rather than a clinical endpoint. CP±L was chosen because its well-characterised, structurally motivated articulatory differences provide a principled test case for establishing whether the framework can detect group-level articulatory distinctions from UTI at all.

Confirming that the framework could learn articulatory distinctions, generalise across speakers, produce interpretable outputs, and operate under data scarcity constraints were necessary preconditions before extending to more complex diagnostic problems. However, in its current form, the binary TD vs. CP±L design has limited standalone clinical utility, since CP±L is already identified through structural examination at birth; the clinical need lies in characterising the nature of associated speech errors and guiding intervention, not in detecting the condition itself. Genuine clinical utility requires moving toward differential diagnosis of SSD subtypes, for example, distinguishing childhood apraxia of speech from severe phonological disorder or childhood dysarthria, conditions that share surface perceptual features but differ in underlying mechanism and therefore require different treatment approaches. Clinical assessment also requires phoneme-level characterisation of error patterns, including whether errors are consistent across productions and whether they reflect true misarticulations or covert contrasts, where a child produces a perceptually incorrect sound that nonetheless encodes a measurable articulatory distinction, with direct implications for treatment planning. Future work should therefore extend the framework to multi-class diagnostic categories and to the detection of such clinically meaningful articulatory patterns.

From a deployment perspective, although the prototype system developed in Chapter 7 demonstrates real-time functionality and was designed with a modular architecture to facilitate future extension, scalability and clinical integration remain open challenges. Real-time video streaming, interoperability with existing clinical software such as AAA, and clinician-in-the-loop evaluation are essential steps toward practical adoption. Importantly, the proposed system should be interpreted as a clinician-support tool rather than an autonomous diagnostic system. Its intended role is to assist SLTs by standardising UTI inputs, enhancing tongue-region visibility, and providing interpretable visual evidence that may support clinical reasoning. It is not designed to replace perceptual assessment, expert judgement, or formal diagnostic decision-making.

Autonomous diagnostic use would require prospective clinical validation, patient-level evaluation, regulatory assessment, and evidence that the system improves clinical decision-making in real-world speech and language therapy settings. User studies with SLTs would provide critical insight into system usability, interpretability, and workflow integration, and would help identify the practical barriers to adoption that automated systems must address in real-world clinical settings. Finally, broader governance and regulatory considerations must be addressed before clinical deployment. The modular architecture of the prototype was designed with future compliance integration in mind; however, full alignment with GDPR, NHS AI governance frameworks, and medical device regulations will be required before the system can be used clinically. Additionally, mechanisms for uncertainty estimation, model monitoring, and audit logging will be essential to ensure safety, transparency, and accountability in real-world settings, particularly given the vulnerability of the paediatric population this system is intended to support.

In summary, this thesis demonstrates that DL can be effectively adapted to the unique imaging characteristics of UTI to support the assessment of childhood SSDs. By systematically addressing variability, data scarcity, and clinical usability through contributions R1–R7, the work provides a unified methodological and practical foundation for future research and deployment. The developments presented here move automated UTI analysis closer to clinical translation and lay the groundwork for scalable, objective, and interpretable articulatory assessment in paediatric speech assessment.

Acknowledgement for the use of AI

I acknowledge the use of language assistance tools during the proofreading of this thesis. These tools were used to improve clarity, coherence, and overall academic tone. Their role was limited to refining the language and flow of the text; all ideas, analyses, and findings presented in this thesis are entirely my own.

Bibliography

[1] P. Eadie, A. Morgan, O. C. Ukoumunne, K. Ttofari Eecen, M. Wake, and S. Reilly, "Speech sound disorder at 4 years: prevalence, comorbidities, and predictors in a community cohort of children," *Dev. Med. Child Neurol.*, vol. 57, no. 6, pp. 578–584, 2015, doi: 10.1111/dmcn.12635.

[2] Y. Wren, L. L. Miller, T. J. Peters, A. Emond, and S. Roulstone, "Prevalence and Predictors of Persistent Speech Sound Disorder at Eight Years Old: Findings From a Population Cohort Study," *J. Speech Lang. Hear. Res.*, vol. 59, no. 4, pp. 647–673, Aug. 2016, doi: 10.1044/2015_JSLHR-S-14-0282.

[3] J. Broomfield and B. Dodd, "Children with speech and language disability: caseload characteristics," *Int. J. Lang. Commun. Disord.*, vol. 39, no. 3, pp. 303–324, 2004, doi: 10.1080/13682820310001625589.

[4] S. McLeod *et al.*, "Cluster-Randomized Controlled Trial Evaluating the Effectiveness of Computer-Assisted Intervention Delivered by Educators for Children With Speech Sound Disorders," *J. Speech Lang. Hear. Res.*, vol. 60, no. 7, pp. 1891–1910, Jul. 2017, doi: 10.1044/2017_JSLHR-S-16-0385.

[5] J. Preston *et al.*, "Ultrasound Images of the Tongue: A Tutorial for Assessment and Remediation of Speech Sound Errors," *J. Vis. Exp.*, vol. 2017, Jan. 2016, doi: 10.3791/55123.

[6] M. Stone, "Imaging the tongue and vocal tract," *Br. J. Disord. Commun.*, vol. 26, no. 1, pp. 11–23, Apr. 1991, doi: 10.3109/13682829109011990.

[7] J. Cleland, "Ultrasound Tongue Imaging," in *Manual of Clinical Phonetics*, 1st ed., M. J. Ball, Ed., Routledge, 2021, pp. 399–416. doi: 10.4324/9780429320903-29.

[8] J. Cleland, J. M. Scobbie, and A. A. Wrench, "Using ultrasound visual biofeedback to treat persistent primary speech sound disorders," *Clin. Linguist. Phon.*, vol. 29, no. 8–10, pp. 575–597, 2015, doi: 10.3109/02699206.2015.1016188.

[9] E. Sugden, S. Lloyd, J. Lam, and J. Cleland, "Systematic review of ultrasound visual biofeedback in intervention for speech sound disorders," *Int. J. Lang. Commun. Disord.*, vol. 54, no. 5, pp. 705–728, Sep. 2019, doi: 10.1111/1460-6984.12478.

[10] N. Zharkova, "Using ultrasound to quantify tongue shape and movement characteristics," *Cleft Palate-Craniofacial J. Off. Publ. Am. Cleft Palate-Craniofacial Assoc.*, vol. 50, no. 1, pp. 76–81, Jan. 2013, doi: 10.1597/11-196.

[11] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[12] A. Fourcade and R. H. Khonsari, "Deep learning in medical image analysis: A third eye for doctors," *J. Stomatol. Oral Maxillofac. Surg.*, vol. 120, no. 4, pp. 279–288, Sep. 2019, doi: 10.1016/j.jormas.2019.06.002.

[13] M. Malakar and R. B. Keskar, "Progress of machine learning based automatic phoneme recognition and its prospect," *Speech Commun.*, vol. 135, pp. 37–53, Dec. 2021, doi: 10.1016/j.specom.2021.09.006.

[14] O. V. Michailovich and A. Tannenbaum, "Despeckling of medical ultrasound images," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 53, no. 1, pp. 64–78, Jan. 2006, doi: 10.1109/TUFFC.2006.1588392.

[15] M. Gazda, S. Kadoury, J. Gazda, and P. Drotar, "Generative Adversarial Networks in Ultrasound Imaging: Extending Field of View Beyond Conventional Limits," Jan. 27, 2025, *arXiv*: arXiv:2405.20981. doi: 10.48550/arXiv.2405.20981.

[16] M. S. Ribeiro, A. Eshky, K. Richmond, and S. Renals, “Speaker-independent classification of phonetic segments from raw ultrasound in child speech,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 1328–1332. doi: 10.1109/ICASSP.2019.8683564.
[17] C. Cordella, M. Marte, H. Liu, and S. Kiran, “An Introduction to Machine Learning for Speech-Language Pathologists: Concepts, Terminology, and Emerging Applications,” *Perspect. ASHA Spec. Interest Groups*, vol. 10, pp. 432–450, Sep. 2024, doi: 10.1044/2024_PERSP-24-00037.
[18] D. Fabre, T. Hueber, F. Bocquelet, and P. Badin, “Tongue Tracking in Ultrasound Images using EigenTongue Decomposition and Artificial Neural Networks,” in *Interspeech 2015 - 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, Sep. 2015. Accessed: Sep. 29, 2022. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01228917
[19] K. Xu, K. You, B. Zhu, M. Feng, D. Feng, and C. Yang, “Masked Modeling-Based Ultrasound Image Classification via Self-Supervised Learning,” *IEEE Open J. Eng. Med. Biol.*, vol. 5, pp. 226–237, 2024, doi: 10.1109/OJEMB.2024.3374966.
[20] V. Berisha and J. M. Liss, “Responsible development of clinical speech AI: Bridging the gap between clinical research and technology,” *NPJ Digit. Med.*, vol. 7, p. 208, Aug. 2024, doi: 10.1038/s41746-024-01199-1.
[21] S. Matta *et al.*, “A systematic review of generalization research in medical image classification,” *Comput. Biol. Med.*, vol. 183, p. 109256, Dec. 2024, doi: 10.1016/j.compbiomed.2024.109256.
[22] K. D. Roon *et al.*, “Comparison of auto-contouring and hand-contouring of ultrasound images of the tongue surface,” *Clin. Linguist. Phon.*, vol. 36, no. 12, pp. 1112–1131, Dec. 2022, doi: 10.1080/02699206.2021.1998633.
[23] A. V. Joshi, “Introduction to AI and ML,” in *Machine Learning and Artificial Intelligence*, A. V. Joshi, Ed., Cham: Springer International Publishing, 2020, pp. 3–7. doi: 10.1007/978-3-030-26622-6_1.
[24] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth, “A review of medical image data augmentation techniques for deep learning applications,” *J. Med. Imaging Radiat. Oncol.*, vol. 65, no. 5, pp. 545–563, Aug. 2021, doi: 10.1111/1754-9485.13261.
[25] P. Kora *et al.*, “Transfer learning techniques for medical image analysis: A review,” *Biocybern. Biomed. Eng.*, vol. 42, no. 1, pp. 79–107, Jan. 2022, doi: 10.1016/j.bbe.2021.11.004.
[26] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier,” Aug. 09, 2016, *arXiv*: arXiv:1602.04938. doi: 10.48550/arXiv.1602.04938.
[27] D. Tang, J. Chen, L. Ren, X. Wang, D. Li, and H. Zhang, “Reviewing CAM-Based Deep Explainable Methods in Healthcare,” *Appl. Sci.*, vol. 14, no. 10, Art. no. 10, Jan. 2024, doi: 10.3390/app14104124.
[28] Z. Xia, R. Yuan, Y. Cao, T. Sun, Y. Xiong, and K. Xu, “A systematic review of the application of machine learning techniques to ultrasound tongue imaging analysis,” *J. Acoust. Soc. Am.*, vol. 156, no. 3, pp. 1796–1819, Sep. 2024, doi: 10.1121/10.0028610.
[29] M. S. Ribeiro, J. Cleland, A. Eshky, K. Richmond, and S. Renals, “Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors,” *Speech Commun.*, vol. 128, pp. 24–34, Apr. 2021, doi: 10.1016/j.specom.2021.02.001.

[30] J. Cleland, M. Dokovova, L. Crampin, and L. Campbell, “An Ultrasound Investigation of Tongue Dorsum Raising in Children with Cleft Palate +/- Cleft Lip,” *Cleft Palate Craniofacial J.*, vol. 61, no. 7, pp. 1104–1115, Jul. 2024, doi: 10.1177/10556656231158965.
[31] A. Anaya-Isaza, L. Mera-Jiménez, and M. Zequera-Diaz, “An overview of deep learning in medical imaging,” *Inform. Med. Unlocked*, vol. 26, p. 100723, Jan. 2021, doi: 10.1016/j.imu.2021.100723.
[32] S. Al Ani, J. Cleland, and A. Zoha, “Two-stage GAN for field-of-view standardisation and tongue region enhancement in ultrasound for cleft palate speech pattern analysis: 6th International Conference on Medical Imaging and Computer-Aided Diagnosis,” *Proc. 2025 Int. Conf. Med. Imaging Comput.-Aided Diagn. MICAD 2025*, Nov. 2025, Accessed: Dec. 10, 2025. [Online]. Available: https://www.micad.org/index.html
[33] S. Al Ani, J. Cleland, and A. Zoha, “Deep learning in ultrasound tongue imaging: a systematic review toward automated detection of speech sound disorders,” *Front. Artif. Intell.*, vol. 8, Sep. 2025, doi: 10.3389/frai.2025.1631134.
[34] P. Eadie, A. Morgan, O. C. Ukoumunne, K. Ttofari Eecen, M. Wake, and S. Reilly, “Speech sound disorder at 4 years: prevalence, comorbidities, and predictors in a community cohort of children,” *Dev. Med. Child Neurol.*, vol. 57, no. 6, pp. 578–584, 2015, doi: 10.1111/dmcn.12635.
[35] S. Harding, S. Burr, J. Cleland, H. Stringer, and Y. Wren, “Outcome measures for children with speech sound disorder: an umbrella review protocol,” *BMJ Open*, vol. 13, no. 2, p. e068945, Feb. 2023, doi: 10.1136/bmjopen-2022-068945.
[36] S. McLeod and L. J. Harrison, “Epidemiology of speech and language impairment in a nationally representative sample of 4- to 5-year-old children,” *J. Speech Lang. Hear. Res. JSLHR*, vol. 52, no. 5, pp. 1213–1229, Oct. 2009, doi: 10.1044/1092-4388(2009/08-0085).
[37] B. A. Lewis, L. A. Freebairn, and H. G. Taylor, “Follow-up of children with early expressive phonology disorders,” *J. Learn. Disabil.*, vol. 33, no. 5, pp. 433–444, 2000, doi: 10.1177/002221940003300504.
[38] Y. Wren *et al.*, “Educational outcomes associated with persistent speech disorder,” *Int. J. Lang. Commun. Disord.*, vol. 56, no. 2, pp. 299–312, Mar. 2021, doi: 10.1111/1460-6984.12599.
[39] H. McFaul, L. Mulgrew, J. Smyth, and J. Titterington, “Applying evidence to practice by increasing intensity of intervention for children with severe speech sound disorder: a quality improvement project,” *BMJ Open Qual.*, vol. 11, no. 2, May 2022, doi: 10.1136/bmjoq-2021-001761.
[40] S. Diepeveen *et al.*, “Process-Oriented Profiling of Speech Sound Disorders,” *Children*, vol. 9, no. 10, p. 1502, Sep. 2022, doi: 10.3390/children9101502.
[41] A. Parnandi *et al.*, “Development of a Remote Therapy Tool for Childhood Apraxia of Speech,” *ACM Trans. Access. Comput.*, vol. 7, no. 3, pp. 1–23, Nov. 2015, doi: 10.1145/2776895.
[42] M. Hardin-Jones and D. Jones, “Speech Production of Preschoolers With Cleft Palate,” *Cleft Palate-Craniofacial J. Off. Publ. Am. Cleft Palate-Craniofacial Assoc.*, vol. 42, pp. 7–13, Feb. 2005, doi: 10.1597/03-134.1.
[43] P. A. Mossey and E. Castilla, *Global registry and database on craniofacial anomalies: report of a WHO Registry Meeting on Craniofacial Anomalies : Baurú, Brazil, 4-6 December 2001*. Geneva, Switzerland: Human Genetics Programme, Management of Noncommunicable Diseases, World Health Organization, 2003.
[44] S. V. Martin and M. C. Swan, “An essential overview of orofacial clefting,” *Br. Dent. J.*, vol. 234, no. 12, pp. 937–942, Jun. 2023, doi: 10.1038/s41415-023-6000-9.

[45] F. W. Wong and N. M. King, “The oral health of children with clefts--a review,” *Cleft Palate-Craniofacial J. Off. Publ. Am. Cleft Palate-Craniofacial Assoc.*, vol. 35, no. 3, pp. 248–254, May 1998, doi: 10.1597/1545-1569_1998_035_0248_tohocw_2.3.co_2.
[46] O. Hunt, D. Burden, P. Hepper, and C. Johnston, “The psychosocial effects of cleft lip and palate: a systematic review,” *Eur. J. Orthod.*, vol. 27, no. 3, pp. 274–285, Jun. 2005, doi: 10.1093/ejo/cji004.
[47] W. L. Adeyemo, O. James, and A. Butali, “Cleft lip and palate: Parental experiences of stigma, discrimination, and social/structural inequalities,” *Ann. Maxillofac. Surg.*, vol. 6, no. 2, pp. 195–203, 2016, doi: 10.4103/2231-0746.200336.
[48] A. D. Sousa, S. Devare, and J. Ghanshani, “Psychological issues in cleft lip and cleft palate,” *J. Indian Assoc. Pediatr. Surg.*, vol. 14, no. 2, pp. 55–58, 2009, doi: 10.4103/0971-9261.55152.
[49] B. Wydick *et al.*, “The Impact of Cleft Lip/Palate and Surgical Intervention on Adolescent Life Outcomes,” *Ann. Glob. Health*, vol. 88, no. 1, p. 25, doi: 10.5334/aogh.3679.
[50] E. De Cuyper, F. Dochy, E. De Leenheer, and H. Van Hoecke, “The impact of cleft lip and/or palate on parental quality of life: A pilot study,” *Int. J. Pediatr. Otorhinolaryngol.*, vol. 126, p. 109598, Nov. 2019, doi: 10.1016/j.ijporl.2019.109598.
[51] R. Wyatt, D. Sell, J. Russell, A. Harding, K. Harland, and E. Albery, “Cleft palate speech dissected: a review of current knowledge and analysis,” *Br. J. Plast. Surg.*, vol. 49, no. 3, pp. 143–149, Apr. 1996, doi: 10.1016/s0007-1226(96)90216-7.
[52] J. L. Preston *et al.*, “Ultrasound Images of the Tongue: A Tutorial for Assessment and Remediation of Speech Sound Errors,” *J. Vis. Exp. JoVE*, no. 119, p. 55123, Jan. 2017, doi: 10.3791/55123.
[53] D. Sell, “Issues in perceptual speech analysis in cleft palate and related disorders: a review,” *Int. J. Lang. Commun. Disord.*, vol. 40, no. 2, pp. 103–121, 2005, doi: 10.1080/13682820400016522.
[54] P. W. Iltis, J. Frahm, D. Voit, A. A. Joseph, E. Schoonderwaldt, and E. Altenmüller, “High-speed real-time magnetic resonance imaging of fast tongue movements in elite horn players,” *Quant. Imaging Med. Surg.*, vol. 5, no. 3, pp. 374–381, Jun. 2015, doi: 10.3978/j.issn.2223-4292.2015.03.02.
[55] Y. Yang *et al.*, “An Audio-Ultrasound Synchronized Database of Tongue Movement for Mandarin speech,” *Sci. Data*, vol. 12, no. 1, p. 607, Apr. 2025, doi: 10.1038/s41597-025-04917-w.
[56] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, “An approach to real-time magnetic resonance imaging for speech production,” *J. Acoust. Soc. Am.*, vol. 115, no. 4, pp. 1771–1776, Apr. 2004, doi: 10.1121/1.1652588.
[57] H. Dent, “The application of electropalatography (EPG) to the remediation of speech disorders in school-aged children and young adults,” *Eur. J. Disord. Commun. J. Coll. Speech Lang. Ther. Lond.*, vol. 30, no. 2, pp. 264–277, 1995, doi: 10.3109/13682829509082537.
[58] S. E. Wood, C. Timmins, J. Wishart, W. J. Hardcastle, and J. Cleland, “THE USE OF ELECTROPALATOGRAPHY IN THE TREATMENT OF SPEECH DISORDERS IN CHILDREN WITH DOWN SYNDROME: A RANDOMISED CONTROLLED TRIAL.”.
[59] H. Leniston and S. Ebbels, “Investigation into the effectiveness of electropalatography in treating persisting speech sound disorders in adolescents with co-occurring developmental language disorder,” *Clin. Linguist. Phon.*, vol. 36, no. 2–3, pp. 111–126, Mar. 2022, doi: 10.1080/02699206.2021.1957022.

[60] A. S. Lee, J. Law, and F. E. Gibbon, “Electropalatography for articulation disorders associated with cleft palate,” *Cochrane Database Syst. Rev.*, vol. 2009, no. 3, p. CD006854, Jul. 2009, doi: 10.1002/14651858.CD006854.pub2.
[61] M. Stone, “A guide to analysing tongue motion from ultrasound images,” *Clin. Linguist. Phon.*, vol. 19, no. 6–7, pp. 455–501, 2005, doi: 10.1080/02699200500113558.
[62] M. S. Ribeiro *et al.*, “Tal: A Synchronised Multi-Speaker Corpus of Ultrasound Tongue Imaging, Audio, and Lip Videos,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China: IEEE, Jan. 2021, pp. 1109–1116. doi: 10.1109/SLT48900.2021.9383619.
[63] M. Moinuddin, S. Khan, A. U. Alsaggaf, M. J. Abdulaal, U. M. Al-Saggaf, and J. C. Ye, “Medical ultrasound image speckle reduction and resolution enhancement using texture compensated multi-resolution convolution neural network,” *Front. Physiol.*, vol. 13, p. 961571, Nov. 2022, doi: 10.3389/fphys.2022.961571.
[64] M. Stone, “A guide to analysing tongue motion from ultrasound images,” *Clin. Linguist. Phon.*, vol. 19, no. 6–7, pp. 455–501, Jan. 2005, doi: 10.1080/02699200500113558.
[65] J. H. Hwang, “Principles of Ultrasound,” in *Endosonography*, Elsevier, 2019, pp. 2-14.e1. doi: 10.1016/B978-0-323-54723-9.00001-4.
[66] M. Li, C. Kambhamettu, and M. Stone, “Automatic contour tracking in ultrasound images,” *Clin. Linguist. Phon.*, vol. 19, no. 6–7, pp. 545–554, Jan. 2005, doi: 10.1080/02699200500113616.
[67] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.
[68] M. H. Mozaffari, A. R. Ratul, and W.-S. Lee, “IrisNet: Deep Learning for Automatic and Real-time Tongue Contour Tracking in Ultrasound Video Data using Peripheral Vision”.
[69] K. Xu, T. G. Csapó, and M. Feng, “Deep learning-based age estimation using B-mode ultrasound tongue imaging,” *J. Acoust. Soc. Am.*, vol. 150, no. 4_Supplement, p. A190, Oct. 2021, doi: 10.1121/10.0008083.
[70] M. Feng, Y. Wang, K. Xu, H. Wang, and B. Ding, “Improving Ultrasound Tongue Contour Extraction Using U-Net and Shape Consistency-Based Regularizer,” *ICASSP 2021 - 2021 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP*, pp. 6443–6447, Jun. 2021, doi: 10.1109/ICASSP39728.2021.9414420.
[71] M. Mozaffari, M. A. R. Ratul, and W.-S. Lee, “IrisNet: Deep Learning for Automatic and Real-time Tongue Contour Tracking in Ultrasound Video Data using Peripheral Vision,” *ArXiv*, Nov. 2019, Accessed: Mar. 02, 2026. [Online]. Available: https://www.semanticscholar.org/paper/2f73c28196cd872c4ea32e4eefafca2e7f0db1f2
[72] R. Mohanasundaram, A. S. Malhotra, R. Arun, and P. S. Periasamy, “Deep Learning and Semi-Supervised and Transfer Learning Algorithms for Medical Imaging,” in *Deep Learning and Parallel Computing Environment for Bioengineering Systems*, Elsevier, 2019, pp. 139–151. doi: 10.1016/B978-0-12-816718-2.00015-4.
[73] A. R. Luca *et al.*, “Impact of quality, type and volume of data used by deep learning models in the analysis of medical images,” *Inform. Med. Unlocked*, vol. 29, p. 100911, Jan. 2022, doi: 10.1016/j.imu.2022.100911.
[74] P. K. Mall *et al.*, “A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities,” *Healthc. Anal.*, vol. 4, p. 100216, Dec. 2023, doi: 10.1016/j.health.2023.100216.

[75] L. Alzubaidi *et al.*, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *J. Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.
[76] H. Greenspan, B. van Ginneken, and R. M. Summers, “Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique,” *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1153–1159, May 2016, doi: 10.1109/TMI.2016.2553401.
[77] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs,” May 12, 2017, *arXiv*: arXiv:1606.00915. doi: 10.48550/arXiv.1606.00915.
[78] A. Paszke *et al.*, “Automatic differentiation in PyTorch,” Oct. 2017, Accessed: Nov. 12, 2025. [Online]. Available: https://openreview.net/forum?id=BJJsrmfCZ
[79] M. Abadi *et al.*, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems,” Mar. 17, 2016, *arXiv*: arXiv:1603.04467. doi: 10.48550/arXiv.1603.04467.
[80] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[81] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” May 18, 2015, *arXiv*: arXiv:1505.04597. doi: 10.48550/arXiv.1505.04597.
[82] I. Malashin, V. Tynchenko, A. Gantimurov, V. Nelyub, and A. Borodulin, “Applications of Long Short-Term Memory (LSTM) Networks in Polymeric Sciences: A Review,” *Polymers*, vol. 16, no. 18, p. 2607, Jan. 2024, doi: 10.3390/polym16182607.
[83] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, “On the Properties of Neural Machine Translation: Encoder-Decoder Approaches,” Oct. 07, 2014, *arXiv*: arXiv:1409.1259. doi: 10.48550/arXiv.1409.1259.
[84] J. Donahue *et al.*, “Long-term Recurrent Convolutional Networks for Visual Recognition and Description:,” Defense Technical Information Center, Fort Belvoir, VA, Nov. 2014. doi: 10.21236/ADA623249.
[85] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification,” *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018, doi: 10.1016/j.neucom.2018.09.013.
[86] X. Yi, E. Walia, and P. Babyn, “Generative adversarial network in medical imaging: A review,” *Med. Image Anal.*, vol. 58, p. 101552, Dec. 2019, doi: 10.1016/j.media.2019.101552.
[87] C. Ledig *et al.*, “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 105–114. doi: 10.1109/CVPR.2017.19.
[88] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT: IEEE, Jun. 2018, pp. 8789–8797. doi: 10.1109/CVPR.2018.00916.
[89] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, “Generative Adversarial Networks for Noise Reduction in Low-Dose CT,” *IEEE Trans. Med. Imaging*, vol. 36, no. 12, pp. 2536–2545, Dec. 2017, doi: 10.1109/TMI.2017.2708987.
[90] C. Szegedy *et al.*, “Intriguing properties of neural networks,” Feb. 19, 2014, *arXiv*: arXiv:1312.6199. doi: 10.48550/arXiv.1312.6199.

[91] L. Han *et al.*, “All-in-one medical image-to-image translation,” *Cell Rep. Methods*, vol. 5, no. 8, p. 101138, Aug. 2025, doi: 10.1016/j.crmeth.2025.101138.
[92] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[93] A. Joshi, “Machine Learning and Artificial Intelligence,” Jan. 2020, doi: 10.1007/978-3-030-26622-6.
[94] J. Zhu, W. Styler, and I. C. Calloway, “Automatic tongue contour extraction in ultrasound images with convolutional neural networks,” *J. Acoust. Soc. Am.*, vol. 143, no. 3_Supplement, pp. 1966–1966, Mar. 2018, doi: 10.1121/1.5036466.
[95] Z. Ghahramani, “Unsupervised Learning,” in *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*, O. Bousquet, U. von Luxburg, and G. Rätsch, Eds., Berlin, Heidelberg: Springer, 2004, pp. 72–112. doi: 10.1007/978-3-540-28650-9_5.
[96] M. S. Ribeiro, J. Cleland, A. Eshky, K. Richmond, and S. Renals, “Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors,” *Speech Commun.*, vol. 128, pp. 24–34, Apr. 2021, doi: 10.1016/j.specom.2021.02.001.
[97] P. Huang, C. Zhang, X. Zhang, X. Li, L. Dong, and L. Ying, “Self-supervised Deep Unrolled Reconstruction Using Regularization by Denoising,” *IEEE Trans. Med. Imaging*, vol. 43, no. 3, pp. 1203–1213, Mar. 2024, doi: 10.1109/TMI.2023.3332614.
[98] L. Jing and Y. Tian, “Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey,” Feb. 16, 2019, *arXiv*: arXiv:1902.06162. doi: 10.48550/arXiv.1902.06162.
[99] H. Liu and J. Zhang, “Improving Ultrasound Tongue Image Reconstruction from Lip Images Using Self-supervised Learning and Attention Mechanism,” Jun. 20, 2021, *arXiv*: arXiv:2106.11769. Accessed: Jun. 23, 2022. [Online]. Available: http://arxiv.org/abs/2106.11769
[100] K. Mendel, H. Li, D. Sheth, and M. Giger, “Transfer Learning From Convolutional Neural Networks for Computer-Aided Diagnosis: A Comparison of Digital Breast Tomosynthesis and Full-Field Digital Mammography,” *Acad. Radiol.*, vol. 26, no. 6, pp. 735–743, Jun. 2019, doi: 10.1016/j.acra.2018.06.019.
[101] M. S. Ali, M. S. Miah, J. Haque, M. M. Rahman, and M. K. Islam, “An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models,” *Mach. Learn. Appl.*, vol. 5, p. 100036, Sep. 2021, doi: 10.1016/j.mlwa.2021.100036.
[102] H. Zhang and Y. Qie, “Applying Deep Learning to Medical Imaging: A Review,” *Appl. Sci.*, vol. 13, no. 18, p. 10521, Jan. 2023, doi: 10.3390/app131810521.
[103] A. A. Taha and A. Hanbury, “Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool,” *BMC Med. Imaging*, vol. 15, p. 29, Aug. 2015, doi: 10.1186/s12880-015-0068-x.
[104] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 5967–5976. doi: 10.1109/CVPR.2017.632.
[105] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization,” Dec. 14, 2015, *arXiv*: arXiv:1512.04150. doi: 10.48550/arXiv.1512.04150.
[106] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[107] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” Nov. 25, 2017, *arXiv*: arXiv:1705.07874. doi: 10.48550/arXiv.1705.07874.
[108] J. Cleland, “Ultrasound Tongue Imaging in Research and Practice with People with Cleft Palate ± Cleft Lip,” *Cleft Palate Craniofacial J.*, p. 10556656231202448, Sep. 2023, doi: 10.1177/10556656231202448.
[109] S. Al Ani, J. Cleland, and A. Zoha, “Automated Classification of Phonetic Segments in Child Speech Using Raw Ultrasound Imaging:,” in *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies*, Rome, Italy: SCITEPRESS - Science and Technology Publications, 2024, pp. 326–331. doi: 10.5220/0012592700003657.
[110] K. You *et al.*, “Raw Ultrasound-Based Phonetic Segments Classification Via Mask Modeling,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095156.
[111] X. Dan *et al.*, “Spatio-temporal masked autoencoder-based phonetic segments classification from ultrasound,” *Speech Commun.*, vol. 169, p. 103186, Apr. 2025, doi: 10.1016/j.specom.2025.103186.
[112] A. Eshky *et al.*, “UltraSuite: A Repository of Ultrasound and Acoustic Data from Child Speech Therapy Sessions,” in *Interspeech 2018*, Sep. 2018, pp. 1888–1892. doi: 10.21437/Interspeech.2018-1736.
[113] M. H. Mozaffari, D. Sankoff, and W.-S. Lee, “BowNet: Dilated convolutional neural network for ultrasound tongue contour extraction,” *J. Acoust. Soc. Am.*, vol. 146, no. 4, pp. 2940–2941, Oct. 2019, doi: 10.1121/1.5137212.
[114] M. H. Mozaffari, N. Yamane, and W.-S. Lee, “Deep Learning for Automatic Tracking of Tongue Surface in Real-time Ultrasound Videos, Landmarks instead of Contours,” in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2020, pp. 2785–2792. doi: 10.1109/BIBM49941.2020.9313262.
[115] G. Li, J. Chen, Y. Liu, and J. Wei, “wUnet: A new network used for ultrasonic tongue contour extraction,” *Speech Commun.*, vol. 141, pp. 68–79, Jun. 2022, doi: 10.1016/j.specom.2022.05.004.
[116] N. Mukai, K. Mori, and Y. Takei, “Tongue model construction based on ultrasound images with image processing and deep learning method,” *J. Med. Ultrason.*, vol. 49, no. 2, pp. 153–161, Apr. 2022, doi: 10.1007/s10396-022-01193-8.
[117] C. Zhao, P. Zhang, J. Zhu, C. Wu, H. Wang, and K. Xu, “Predicting Tongue Motion in Unlabeled Ultrasound Videos Using Convolutional Lstm Neural Networks,” in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5926–5930. doi: 10.1109/ICASSP.2019.8683081.
[118] K. Xu, K. You, B. Zhu, M. Feng, D. Feng, and C. Yang, “Masked Modeling-Based Ultrasound Image Classification via Self-Supervised Learning,” *IEEE Open J. Eng. Med. Biol.*, vol. 5, pp. 226–237, 2024, doi: 10.1109/OJEMB.2024.3374966.
[119] M. H. Mozaffari and W.-S. Lee, “BowNet: Dilated Convolution Neural Network for Ultrasound Tongue Contour Extraction,” *J. Acoust. Soc. Am.*, vol. 146, no. 4, pp. 2940–2941, Oct. 2019, doi: 10.1121/1.5137212.
[120] B. Guo *et al.*, “The Impact of Scanner Domain Shift on Deep Learning Performance in Medical Imaging: an Experimental Study,” Oct. 02, 2024, *arXiv*: arXiv:2409.04368. doi: 10.48550/arXiv.2409.04368.
[121] M. Shahin, B. Ahmed, D. V. Smith, A. Duenser, and J. Epps, “Automatic Screening Of Children With Speech Sound Disorders Using Paralinguistic Features,” in *2019 IEEE 29th*

International Workshop on Machine Learning for Signal Processing (MLSP), Oct. 2019, pp. 1–5. doi: 10.1109/MLSP.2019.8918725.
[122] J. Cleland, C. Mccron, and J. M. Scobbie, "Tongue reading: Comparing the interpretation of visual information from inside the mouth, from electropalatographic and ultrasound displays of speech sounds," *Clin. Linguist. Phon.*, vol. 27, no. 4, pp. 299–311, Apr. 2013, doi: 10.3109/02699206.2012.759626.
[123] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clin. Linguist. Phon.*, vol. 19, no. 6–7, pp. 455–501, Jan. 2005, doi: 10.1080/02699200500113558.
[124] C. Chen, N. A. Mat Isa, and X. Liu, "A review of convolutional neural network based methods for medical image classification," *Comput. Biol. Med.*, vol. 185, p. 109507, Feb. 2025, doi: 10.1016/j.compbiomed.2024.109507.
[125] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.
[126] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Dec. 11, 2015, *arXiv*: arXiv:1512.00567. doi: 10.48550/arXiv.1512.00567.
[127] T. Hueber *et al.*, "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, Apr. 2007, p. I-1245-I–1248. doi: 10.1109/ICASSP.2007.366140.
[128] J. Cai, B. Denby, P. Roussel, G. Dreyfus, and L. Crevier-Buchman, "Recognition and real time performances of a lightweight ultrasound based silent speech interface employing a language model," in *Interspeech 2011*, ISCA, Aug. 2011, pp. 1005–1008. doi: 10.21437/Interspeech.2011-410.
[129] K. Xu, P. Roussel, T. G. Csapó, and B. Denby, "Convolutional neural network-based automatic classification of midsagittal tongue gestural targets using B-mode ultrasound images," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. EL531–EL537, Jun. 2017, doi: 10.1121/1.4984122.
[130] K. You *et al.*, "Raw Ultrasound-Based Phonetic Segments Classification Via Mask Modeling," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095156.
[131] S. Liu, N. Wang, and Z. Wang, "Self-Supervised Learning of ECG and PPG Signals for Multi-Modal Health Monitoring," in *Proceedings of 2025 2nd International Conference on Machine Learning and Intelligent Computing*, PMLR, Oct. 2025, pp. 350–358. Accessed: Dec. 15, 2025. [Online]. Available: https://proceedings.mlr.press/v278/liu25d.html
[132] Zhenhua Guo, Lei Zhang, and D. Zhang, "A Completed Modeling of Local Binary Pattern Operator for Texture Classification," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1657–1663, Jun. 2010, doi: 10.1109/TIP.2010.2044957.
[133] P. Zhang, Z. Ma, Y. Zhang, X. Chen, and G. Wang, "Improved Inception V3 method and its effect on radiologists' performance of tumor classification with automated breast ultrasound system," *Gland Surg.*, vol. 10, no. 7, pp. 2232–2245, Jul. 2021, doi: 10.21037/gs-21-328.
[134] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002, doi: 10.1109/TPAMI.2002.1017623.
[135] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[136] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, Art. no. 1, Feb. 2017, doi: 10.1609/aaai.v31i1.11231.
[137] Y. Pan *et al.*, "Fundus image classification using Inception V3 and ResNet-50 for the early diagnostics of fundus diseases," *Front. Physiol.*, vol. 14, Feb. 2023, doi: 10.3389/fphys.2023.1126780.
[138] R. Yousef, G. Gupta, N. Yousef, and M. Khari, "A holistic overview of deep learning approach in medical imaging," *Multimed. Syst.*, vol. 28, no. 3, pp. 881–914, Jun. 2022, doi: 10.1007/s00530-021-00884-5.
[139] W. Xu, Y.-L. Fu, and D. Zhu, "ResNet and its application to medical image processing: Research progress and challenges," *Comput. Methods Programs Biomed.*, vol. 240, p. 107660, Oct. 2023, doi: 10.1016/j.cmpb.2023.107660.
[140] A. Lott and J. P. Reiter, "Wilson Confidence Intervals for Binomial Proportions With Multiple Imputation for Missing Data," *Am. Stat.*, vol. 74, no. 2, pp. 109–115, Apr. 2020, doi: 10.1080/00031305.2018.1473796.
[141] A. Cadrin-Chênevert, "Navigating Clinical Variability: Transfer Learning's Impact on Imaging Model Performance," *Radiol. Artif. Intell.*, vol. 6, no. 4, p. e240263, Jun. 2024, doi: 10.1148/ryai.240263.
[142] T. Bressmann, Radovanovic ,Bojana, Kulkarni ,Gajanan V., Klaiman ,Paula, and D. and Fisher, "An ultrasonographic investigation of cleft-type compensatory articulations of voiceless velar stops," *Clin. Linguist. Phon.*, vol. 25, no. 11–12, pp. 1028–1033, Nov. 2011, doi: 10.3109/02699206.2011.599472.
[143] J. L. Preston *et al.*, "Ultrasound Images of the Tongue: A Tutorial for Assessment and Remediation of Speech Sound Errors," *J. Vis. Exp.*, no. 119, p. 55123, Jan. 2017, doi: 10.3791/55123.
[144] M. S. Ribeiro, J. Cleland, A. Eshky, K. Richmond, and S. Renals, "Exploiting ultrasound tongue imaging for the automatic detection of speech articulation errors," *Speech Commun.*, vol. 128, pp. 24–34, Apr. 2021, doi: 10.1016/j.specom.2021.02.001.
[145] S. Al Ani, "Systematic review of deep learning models in ultrasound tongue imaging for the detection of speech disorders," Apr. 28, 2023, *TechRxiv*. doi: 10.36227/techrxiv.22699291.v1.
[146] K. L. Chapman and M. A. Hardin, "Phonetic and phonologic skills of two-year-olds with cleft palate," *Cleft Palate-Craniofacial J. Off. Publ. Am. Cleft Palate-Craniofacial Assoc.*, vol. 29, no. 5, pp. 435–443, Sep. 1992, doi: 10.1597/1545-1569_1992_029_0435_papsot_2.3.co_2.
[147] A. Harding and P. Grunwell, "Active versus passive cleft-type speech characteristics," *Int. J. Lang. Commun. Disord.*, vol. 33, no. 3, pp. 329–352, 1998, doi: 10.1080/136828298247776.
[148] Y. Wang, X. Ge, H. Ma, S. Qi, G. Zhang, and Y. Yao, "Deep Learning in Medical Ultrasound Image Analysis: A Review," *IEEE Access*, vol. 9, pp. 54310–54324, 2021, doi: 10.1109/ACCESS.2021.3071301.
[149] K. Al-hammuri, F. Gebali, I. Chelvan, and A. Kanan, "Tongue Contour Tracking and Segmentation in Lingual Ultrasound for Speech Recognition: A Review," *Diagnostics*, vol. 12, p. 2811, Nov. 2022, doi: 10.3390/diagnostics12112811.
[150] sue ann Lee, A. Wrench, and S. Sancibrian, "How To Get Started With Ultrasound Technology for Treatment of Speech Sound Disorders," *Perspect. Speech Sci. Orofac. Disord.*, vol. 25, p. 66, Oct. 2015, doi: 10.1044/ssod25.2.66.
[151] N. Salimova, J. B. Hinrichs, M. Gutberlet, B. C. Meyer, F. K. Wacker, and C. von Falck, "The impact of the field of view (FOV) on image quality in MDCT angiography of the lower

extremities," *Eur. Radiol.*, vol. 32, no. 5, pp. 2875–2882, 2022, doi: 10.1007/s00330-021-08391-x.
[152] T. Miyata *et al.*, "Influence of field of view size on image quality: ultra-high-resolution CT vs. conventional high-resolution CT," *Eur. Radiol.*, vol. 30, no. 6, pp. 3324–3333, 2020, doi: 10.1007/s00330-020-06704-0.
[153] J. Wang, "Deep Cascade Learning for Optimal Medical Image Feature Representation".
[154] A. Smith, M. Dokovova, E. Lawson, A. Kuschmann, and J. Cleland, "A pilot fieldwork ultrasound study of tongue shape variability in children with and without speech sound disorder".
[155] "http://materials.articulateinstruments.com/Manuals/Archive/AAA%20Manual_2_16_12_Ultrasound%20Module.pdf." Accessed: Mar. 25, 2025. [Online]. Available: http://materials.articulateinstruments.com/Manuals/Archive/AAA%20Manual_2_16_12_Ultrasound%20Module.pdf
[156] M. Pucher, N. Klingler, J. Luttenberger, and L. Spreafico, "Accuracy, recording interference, and articulatory quality of headsets for ultrasound recordings," *Speech Commun.*, vol. 123, pp. 83–97, Oct. 2020, doi: 10.1016/j.specom.2020.07.001.
[157] A. Eshky, M. S. Ribeiro, K. Richmond, and S. Renals, "Synchronising audio and ultrasound by learning cross-modal embeddings," Nov. 27, 2019, *arXiv*: arXiv:1907.00758. Accessed: Nov. 21, 2023. [Online]. Available: http://arxiv.org/abs/1907.00758
[158] A. Wrench and J. Balch-Tomes, "Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut," *Sensors*, vol. 22, no. 3, p. 1133, Feb. 2022, doi: 10.3390/s22031133.
[159] Y. Kimori, "Morphological image processing for quantitative shape analysis of biomedical structures: effective contrast enhancement," *J. Synchrotron Radiat.*, vol. 20, no. Pt 6, pp. 848–853, Nov. 2013, doi: 10.1107/S0909049513020761.
[160] A. Zhou *et al.*, "Multi-head attention-based two-stream EfficientNet for action recognition," *Multimed. Syst.*, vol. 29, pp. 1–12, Jun. 2022, doi: 10.1007/s00530-022-00961-3.
[161] T. Liu, H. Yu, and R. H. Blair, "Stability estimation for unsupervised clustering: A review," *WIREs Comput. Stat.*, vol. 14, no. 6, p. e1575, 2022, doi: 10.1002/wics.1575.
[162] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, Oct. 2020, pp. 747–748. doi: 10.1109/DSAA49011.2020.00096.
[163] S. Miyamoto, R. Abe, Y. Endo, and J. Takeshita, "Ward method of hierarchical clustering for non-Euclidean similarity measures," in *2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, Nov. 2015, pp. 60–63. doi: 10.1109/SOCPAR.2015.7492784.
[164] R. A. Armstrong, "When to use the Bonferroni correction," *Ophthalmic Physiol. Opt. J. Br. Coll. Ophthalmic Opt.*, vol. 34, no. 5, pp. 502–508, Sep. 2014, doi: 10.1111/opo.12131.
[165] C. R. Brydges, "Effect Size Guidelines, Sample Size Calculations, and Statistical Power in Gerontology," *Innov. Aging*, vol. 3, no. 4, p. igz036, Sep. 2019, doi: 10.1093/geroni/igz036.
[166] H. Hashemi Hosseinabad and Y. Xing, "Tongue dorsum activity in children with velopharyngeal insufficiency vs. typically developing children," *Clin. Linguist. Phon.*, pp. 1–19, Oct. 2024, doi: 10.1080/02699206.2024.2411946.
[167] R. Geirhos *et al.*, "Shortcut Learning in Deep Neural Networks," *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020, doi: 10.1038/s42256-020-00257-z.
[168] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest

radiographs: A cross-sectional study," *PLOS Med.*, vol. 15, no. 11, p. e1002683, Nov. 2018, doi: 10.1371/journal.pmed.1002683.
[169] J. K. Winkler *et al.*, "Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition," *JAMA Dermatol.*, vol. 155, no. 10, pp. 1135–1141, Oct. 2019, doi: 10.1001/jamadermatol.2019.1735.
[170] P. W. Battaglia *et al.*, "Relational inductive biases, deep learning, and graph networks," Oct. 17, 2018, *arXiv*: arXiv:1806.01261. doi: 10.48550/arXiv.1806.01261.
[171] X. Wang, X. Li, R. Du, Y. Zhong, Y. Lu, and T. Song, "Anatomical Prior-Based Automatic Segmentation for Cardiac Substructures from Computed Tomography Images," *Bioengineering*, vol. 10, no. 11, p. 1267, Oct. 2023, doi: 10.3390/bioengineering10111267.
[172] Adewale Abayomi Adeniran, Amaka Peace Onebunne, and Paul William, "Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making," *World J. Adv. Res. Rev.*, vol. 23, no. 3, pp. 2447–2658, Sep. 2024, doi: 10.30574/wjarr.2024.23.3.2936.
[173] G. Abgrall, A. L. Holder, Z. Chelly Dagdia, K. Zeitouni, and X. Monnet, "Should AI models be explainable to clinicians?," *Crit. Care*, vol. 28, p. 301, Sep. 2024, doi: 10.1186/s13054-024-05005-y.
[174] S. Abbasi, H. Lan, J. Choupan, N. Sheikh-Bahaei, G. Pandey, and B. Varghese, "Deep learning for the harmonization of structural MRI scans: a survey," *Biomed. Eng. OnLine*, vol. 23, no. 1, p. 90, Aug. 2024, doi: 10.1186/s12938-024-01280-6.
[175] A. Midya, J. Chakraborty, M. Gönen, R. K. G. Do, and A. L. Simpson, "Influence of CT acquisition and reconstruction parameters on radiomic feature reproducibility," *J. Med. Imaging*, vol. 5, no. 1, p. 011020, Jan. 2018, doi: 10.1117/1.JMI.5.1.011020.
[176] H. Alqahtani, M. Kavakli-Thorne, and G. Kumar, "Applications of Generative Adversarial Networks (GANs): An Updated Review," *Arch. Comput. Methods Eng.*, vol. 28, no. 2, pp. 525–552, Mar. 2021, doi: 10.1007/s11831-019-09388-y.
[177] X. Guo *et al.*, "Damage Detection for Conveyor Belt Surface Based on Conditional Cycle Generative Adversarial Network," *Sensors*, vol. 22, no. 9, p. 3485, Jan. 2022, doi: 10.3390/s22093485.
[178] T. C. W. Mok and A. C. S. Chung, "Learning Data Augmentation for Brain Tumor Segmentation with Coarse-to-Fine Generative Adversarial Networks," vol. 11383, 2019, pp. 70–80. doi: 10.1007/978-3-030-11723-8_7.
[179] W. Dai *et al.*, "SCAN: Structure Correcting Adversarial Network for Organ Segmentation in Chest X-rays," Apr. 10, 2017, *arXiv*: arXiv:1703.08770. doi: 10.48550/arXiv.1703.08770.
[180] Y. Xue, T. Xu, H. Zhang, R. Long, and X. Huang, "SegAN: Adversarial Network with Multi-scale L_1 Loss for Medical Image Segmentation," *Neuroinformatics*, vol. 16, no. 3–4, pp. 383–392, Oct. 2018, doi: 10.1007/s12021-018-9377-x.
[181] D. Nie, R. Trullo, C. Petitjean, S. Ruan, and D. Shen, "Medical Image Synthesis with Context-Aware Generative Adversarial Networks," Dec. 16, 2016, *arXiv*: arXiv:1612.05362. doi: 10.48550/arXiv.1612.05362.
[182] S. Gandhi, H. Rana, and N. Bhatt, "Conditional GANs in Image-to-Image Translation: Improving Accuracy and Contextual Relevance in Diverse Datasets," *Procedia Comput. Sci.*, vol. 252, pp. 954–963, Jan. 2025, doi: 10.1016/j.procs.2025.01.056.
[183] K. Ko, T. Yeom, and M. Lee, "SuperstarGAN: Generative adversarial networks for image-to-image translation in large-scale domains," *Neural Netw.*, vol. 162, pp. 330–339, May 2023, doi: 10.1016/j.neunet.2023.02.042.

[184] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, “SPA-GAN: Spatial Attention GAN for Image-to-Image Translation,” *IEEE Trans. Multimed.*, vol. 23, pp. 391–401, 2021, doi: 10.1109/TMM.2020.2975961.
[185] R. Liu *et al.*, “SCCGAN: Style and Characters Inpainting Based on CGAN,” *Mob. Netw. Appl.*, vol. 26, no. 1, pp. 3–12, Feb. 2021, doi: 10.1007/s11036-020-01717-x.
[186] A. Ben-Cohen, E. Klang, S. P. Raskin, M. M. Amitai, and H. Greenspan, “Virtual PET Images from CT Data Using Deep Convolutional Networks: Initial Results,” vol. 10557, 2017, pp. 49–57. doi: 10.1007/978-3-319-68127-6_6.
[187] Z. Zhao *et al.*, “Radiomics Harmonization in Ultrasound Images for Cervical Cancer Lymph Node Metastasis Prediction Using Cycle-GAN,” *Technol. Cancer Res. Treat.*, vol. 23, p. 15330338241302237, 2024, doi: 10.1177/15330338241302237.
[188] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery,” Mar. 17, 2017, *arXiv*: arXiv:1703.05921. doi: 10.48550/arXiv.1703.05921.
[189] H. Akbarialiabad *et al.*, “The Utility of Generative AI in Advancing Global Health,” *NEJM AI*, Feb. 2025, doi: 10.1056/AIp2400875.
[190] A. Karmakar *et al.*, “The role of generative AI in medical image synthesis: A review,” *Discov. Appl. Sci.*, vol. 7, no. 10, p. 1219, Oct. 2025, doi: 10.1007/s42452-025-07714-7.
[191] D. P. Costello and P. A. Kenny, “Fat Segmentation in Magnetic Resonance Images,” in *Medical Image Processing: Techniques and Applications*, G. Dougherty, Ed., New York, NY: Springer, 2011, pp. 89–113. doi: 10.1007/978-1-4419-9779-1_5.
[192] C. Couprie, L. Najman, and H. Talbot, “Seeded Segmentation Methods for Medical Image Analysis,” in *Medical Image Processing*, G. Dougherty, Ed., in Biological and Medical Physics, Biomedical Engineering. , New York, NY: Springer New York, 2011, pp. 27–57. doi: 10.1007/978-1-4419-9779-1_3.
[193] S. S. Mohammed and H. G. Clarke, “Conditional image-to-image translation generative adversarial network (cGAN) for fabric defect data augmentation,” *Neural Comput. Appl.*, vol. 36, no. 32, pp. 20231–20244, Nov. 2024, doi: 10.1007/s00521-024-10179-1.
[194] A. Abu-Srhan, M. A. M. Abushariah, and O. S. Al-Kadi, “The effect of loss function on conditional generative adversarial networks,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 9, pp. 6977–6988, Oct. 2022, doi: 10.1016/j.jksuci.2022.02.018.
[195] S. Cammarasana, P. Nicolardi, and G. Patanè, “Real-time denoising of ultrasound images based on deep learning,” *Med. Biol. Eng. Comput.*, vol. 60, no. 8, pp. 2229–2244, Aug. 2022, doi: 10.1007/s11517-022-02573-5.
[196] J. Chen *et al.*, “Generative models improve radiomics reproducibility in low dose CTs: A simulation study,” *Phys. Med. Biol.*, vol. 66, Aug. 2021, doi: 10.1088/1361-6560/ac16c0.
[197] I. H. Rather and S. Kumar, “Generative adversarial network based synthetic data training model for lightweight convolutional neural networks,” *Multimed. Tools Appl.*, pp. 1–23, May 2023, doi: 10.1007/s11042-023-15747-6.
[198] M. Kannan, D. Umamaheswari, B. Manimekala, I. P. S. Mary, P. M. Savitha, and J. Rozario, “An enhancement of machine learning model performance in disease prediction with synthetic data generation,” *Sci. Rep.*, vol. 15, no. 1, p. 33482, Sep. 2025, doi: 10.1038/s41598-025-15019-3.
[199] J. L. Preston, P. McCabe, A. Rivera-Campos, J. L. Whittle, E. Landry, and E. Maas, “Ultrasound visual feedback treatment and practice variability for residual speech sound errors,” *J. Speech Lang. Hear. Res. JSLHR*, vol. 57, no. 6, pp. 2102–2115, Dec. 2014, doi: 10.1044/2014_JSLHR-S-14-0031.

[200] M. Eslami, C. Neuschaefer-Rube, and A. Serrurier, “Automatic vocal tract landmark localization from midsagittal MRI data,” *Sci. Rep.*, vol. 10, no. 1, p. 1468, Jan. 2020, doi: 10.1038/s41598-020-58103-6.
[201] R. R. Jha, A. Muralie, M. Daroch, A. Bhavsar, and A. Nigam, “Enhancing Autism Spectrum Disorder identification in multi-site MRI imaging: A multi-head cross-attention and multi-context approach for addressing variability in un-harmonized data,” *Artif. Intell. Med.*, vol. 157, p. 102998, Nov. 2024, doi: 10.1016/j.artmed.2024.102998.
[202] J. Chen *et al.*, “Medical image translation with deep learning: Advances, datasets and perspectives,” *Med. Image Anal.*, vol. 103, p. 103605, Jul. 2025, doi: 10.1016/j.media.2025.103605.
[203] Z. Roxburgh, “Visualising articulation: real-time ultrasound visual biofeedback and visual articulatory models and their use in treating speech sound disorders associated with submucous cleft palate,” 2018, Accessed: Jan. 15, 2026. [Online]. Available: https://eresearch.qmu.ac.uk/handle/20.500.12289/8899
[204] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2242–2251. doi: 10.1109/ICCV.2017.244.
[205] J. Granstedt, P. Kc, R. Deshpande, V. Garcia, and A. Badano, “Hallucinations in medical devices,” *Artif. Intell. Life Sci.*, vol. 8, p. 100145, Dec. 2025, doi: 10.1016/j.ailsci.2025.100145.
[206] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual Losses for Real-Time Style Transfer and Super-Resolution,” in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 694–711. doi: 10.1007/978-3-319-46475-6_43.
[207] T. Patel and O. Scharenborg, “Using Data Augmentations and VTLN to Reduce Bias in Dutch End-to-End Speech Recognition Systems,” 2023, doi: 10.48550/ARXIV.2307.02009.
[208] Y. Zhang, Z. Yue, T. Patel, and O. Scharenborg, “Improving child speech recognition with augmented child-like speech,” in *Interspeech 2024*, ISCA, Sep. 2024, pp. 5183–5187. doi: 10.21437/Interspeech.2024-485.
[209] O. Niebuhr and A. Michaud, “Speech data acquisition: the underestimated challenge,” Feb. 2015.
[210] C. Cieri, D. Miller, and K. Walker, “Research methodologies, observations and outcomes in (conversational) speech data collection,” in *Proceedings of the second international conference on Human Language Technology Research* -, San Diego, California: Association for Computational Linguistics, 2002, pp. 206–211. doi: 10.3115/1289189.1289198.
[211] F. Galbusera and A. Cina, “Image annotation and curation in radiology: an overview for machine learning practitioners,” *Eur. Radiol. Exp.*, vol. 8, p. 11, Feb. 2024, doi: 10.1186/s41747-023-00408-y.
[212] J. Y.-L. Chan, K. T. Bea, S. M. H. Leow, S. W. Phoong, and W. K. Cheng, “State of the art: a review of sentiment analysis based on sequential transfer learning,” *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 749–780, Jan. 2023, doi: 10.1007/s10462-022-10183-8.
[213] M. Karnes, S. Perera, S. Adhikari, and A. Yilmaz, “Adaptive Few-Shot Learning PoC Ultrasound COVID-19 Diagnostic System,” Sep. 08, 2021, *arXiv*: arXiv:2109.03793. doi: 10.48550/arXiv.2109.03793.
[214] V. Rani, S. T. Nabi, M. Kumar, A. Mittal, and K. Kumar, “Self-supervised Learning: A Succinct Review,” *Arch. Comput. Methods Eng.*, vol. 30, no. 4, pp. 2761–2775, May 2023, doi: 10.1007/s11831-023-09884-2.
[215] A. Lawley, R. Hampson, K. Worrall, and G. Dobie, “Prescriptive Method for Optimizing Cost of Data Collection and Annotation in Machine Learning of Clinical Ultrasound,” in *2023*

45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Jul. 2023, pp. 1–4. doi: 10.1109/EMBC40787.2023.10340858.
[216] K. Hemming, S. Eldridge, G. Forbes, C. Weijer, and M. Taljaard, "How to design efficient cluster randomised trials," *BMJ*, vol. 358, p. j3064, Jul. 2017, doi: 10.1136/bmj.j3064.
[217] C. C. Serdar, M. Cihan, D. Yücel, and M. A. Serdar, "Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies," *Biochem. Medica*, vol. 31, no. 1, p. 010502, Feb. 2021, doi: 10.11613/BM.2021.010502.
[218] J. Cohen, "Statistical Power Analysis," *Curr. Dir. Psychol. Sci.*, vol. 1, no. 3, pp. 98–101, Jun. 1992, doi: 10.1111/1467-8721.ep10768783.
[219] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance," *BMC Med. Inform. Decis. Mak.*, vol. 12, no. 1, p. 8, Feb. 2012, doi: 10.1186/1472-6947-12-8.
[220] A. Rokem, Y. Wu, and A. Lee, *Assessment of the need for separate test set and number of medical images necessary for deep learning: a sub-sampling study*. 2017. doi: 10.1101/196659.
[221] C. Voorhis and B. Morgan, "Understanding Power and Rules of Thumb for Determining Sample Size," *Tutor. Quant. Methods Psychol.*, vol. 3, Sep. 2007, doi: 10.20982/tqmp.03.2.p043.
[222] E. B. Baum and D. Haussler, "What Size Net Gives Valid Generalization?," *Neural Comput.*, vol. 1, no. 1, pp. 151–160, Mar. 1989, doi: 10.1162/neco.1989.1.1.151.
[223] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 1st ed. USA: Prentice Hall PTR, 1994.
[224] I. Balki *et al.*, "Sample-Size Determination Methodologies for Machine Learning in Medical Imaging Research: A Systematic Review," *Can. Assoc. Radiol. J.*, vol. 70, no. 4, pp. 344–353, Nov. 2019, doi: 10.1016/j.carj.2019.06.002.
[225] M. R. Goldstein, "Are Clinical Trials Cost-effective?," *JAMA*, vol. 263, no. 11, pp. 1491–1492, Mar. 1990, doi: 10.1001/jama.1990.03440110053014.
[226] S. Budd, E. C. Robinson, and B. Kainz, "A survey on active learning and human-in-the-loop deep learning for medical image analysis," *Med. Image Anal.*, vol. 71, p. 102062, Jul. 2021, doi: 10.1016/j.media.2021.102062.
[227] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-Class Active Learning by Uncertainty Sampling with Diversity Maximization," *Int J Comput Vis.*, vol. 113, no. 2, pp. 113–127, Jun. 2015, doi: 10.1007/s11263-014-0781-x.
[228] D. D. Lewis, T. B. Laboratories, and M. Hill, "A Sequential Algorithm for Training Text Classi ers".
[229] J. Yun, J. Oh, and I. Yun, "Gradually Applying Weakly Supervised and Active Learning for Mass Detection in Breast Ultrasound Images," *Appl. Sci.*, vol. 10, no. 13, p. 4519, Jan. 2020, doi: 10.3390/app10134519.
[230] L. Gao *et al.*, "Multi-Modal Active Learning for Automatic Liver Fibrosis Diagnosis based on Ultrasound Shear Wave Elastography," Nov. 02, 2020, *arXiv*: arXiv:2011.00694. doi: 10.48550/arXiv.2011.00694.
[231] L. Liu, W. Lei, Y. Luo, C. Feng, X. Wan, and L. Liu, "Semi-Supervised Active Learning for COVID-19 Lung Ultrasound Multi-symptom Classification," Feb. 28, 2021, *arXiv*: arXiv:2009.05436. doi: 10.48550/arXiv.2009.05436.
[232] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[233] J. G. de Almeida, C. Messiou, S. J. Withey, C. Matos, D.-M. Koh, and N. Papanikolaou, “Medical machine learning operations: a framework to facilitate clinical AI development and deployment in radiology,” *Eur. Radiol.*, vol. 35, no. 11, pp. 6828–6841, Nov. 2025, doi: 10.1007/s00330-025-11654-6.
[234] J. Preston *et al.*, “Ultrasound Images of the Tongue: A Tutorial for Assessment and Remediation of Speech Sound Errors,” *J. Vis. Exp.*, vol. 2017, Jan. 2016, doi: 10.3791/55123.
[235] Y. A. Fahim, I. W. Hasani, S. Kabba, and W. M. Ragab, “Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives,” *Eur. J. Med. Res.*, vol. 30, p. 848, Sep. 2025, doi: 10.1186/s40001-025-03196-w.
[236] G. D. Luca, *FastAPI cookbook: develop high-performance APIs and web applications with Python*. Place of publication not identified: Packt Publishing, 2024.
[237] T. Monks and A. Harper, “Improving the usability of open health service delivery simulation models using Python and web apps,” *NIHR Open Res.*, vol. 3, p. 48, Dec. 2023, doi: 10.3310/nihropenres.13467.2.
[238] A. Wasik, “Principles of REST API Design,” 2017.
[239] M. Garimilla, “THE ART OF API DESIGN: BEST PRACTICES FOR MODERN SOFTWARE DEVELOPMENT,” *Int. J. Eng. Tech. Res. IJETR*, vol. 9, pp. 229–239, Sep. 2024, doi: 10.5281/zenodo.13772395.
[240] M. Openja, F. Majidi, F. Khomh, B. Chembakottu, and H. Li, “Studying the Practices of Deploying Machine Learning Projects on Docker,” in *The International Conference on Evaluation and Assessment in Software Engineering 2022*, Gothenburg Sweden: ACM, Jun. 2022, pp. 190–200. doi: 10.1145/3530019.3530039.

Appendix A: UltraSuite Speaker IDs

This appendix lists speaker identifiers and demographic information for the CP±L participants included in the UltraSuite dataset used in this thesis.

Table A1: CP±L Speakers' Details.

Speaker ID	Gender
01M	M
05M	M
06M	M
07M	M
09M	M
12F	F
14F	F
16M	M
18F	F
21M	M
20F	F
24M	M
28F	F
30F	F

Appendix B: Supplementary Training Dynamics and Error Analysis

This appendix provides complete training, validation learning curves and test-set confusion matrices for all datasets and image-representation combinations evaluated in Chapter 4. These results are included to demonstrate training stability, convergence behaviour, and error distributions, and to confirm that reported performance differences are not driven by overfitting or optimisation artefacts. Training and validation losses are plotted using dual y-axes due to differences in loss scale, allowing both convergence trends to be visualised clearly on a shared epoch axis.

B.1 Original Dataset

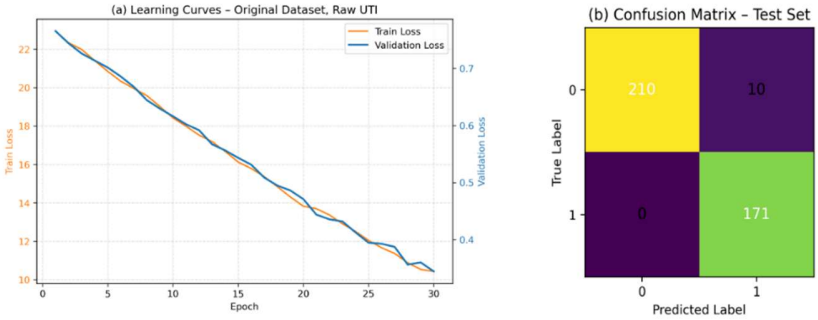

Figure B.1 (a) Training and validation learning curves for the raw UTI model trained on the original dataset. (b) Confusion matrix on the held-out test set for the same model.

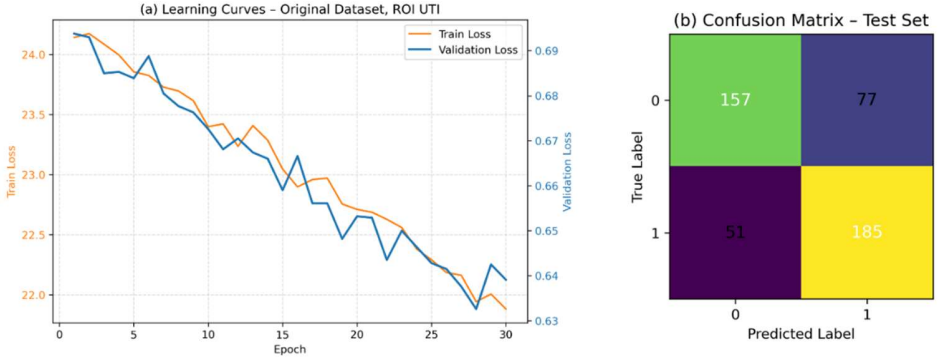

Figure B.2 (a) Training and validation learning curves for the ROI UTI model trained on the original dataset. (b) Confusion matrix on the held-out test set for the same model.

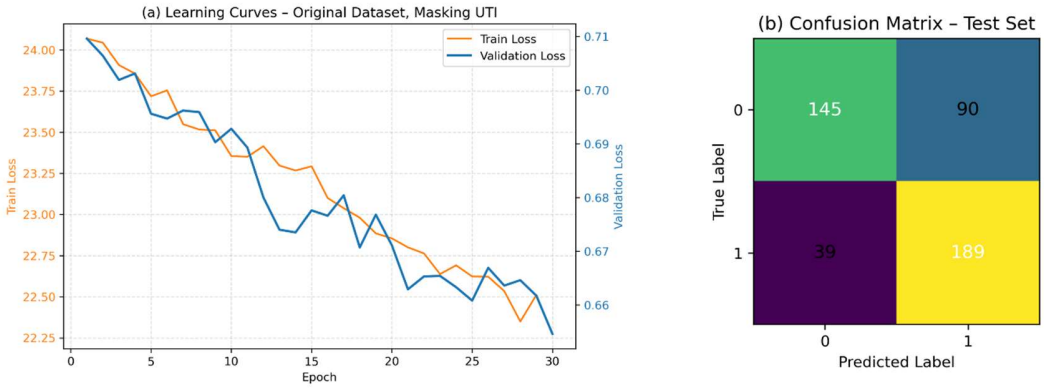

Figure B.3 (a) Training and validation learning curves for the masked UTI model trained on the original dataset. (b) Confusion matrix on the held-out test set for the same model.

B.2 Bicubic Dataset

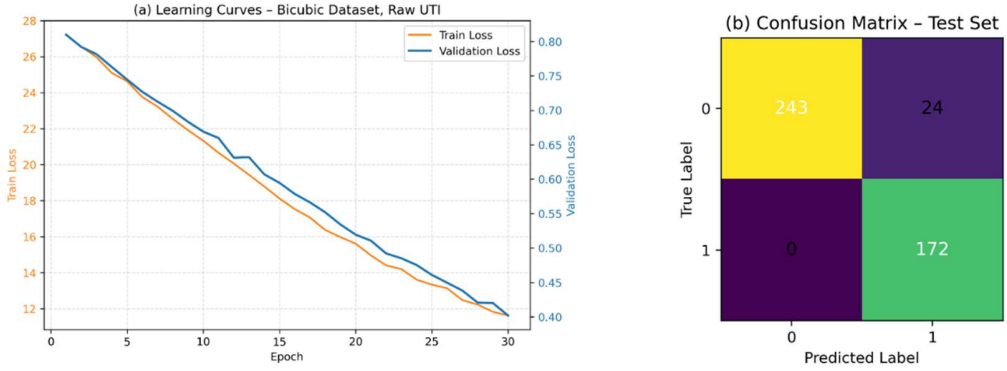

Figure B.4 (a) Training and validation learning curves for the raw UTI model trained on the bicubic dataset. (b) Confusion matrix on the held-out test set for the same model.

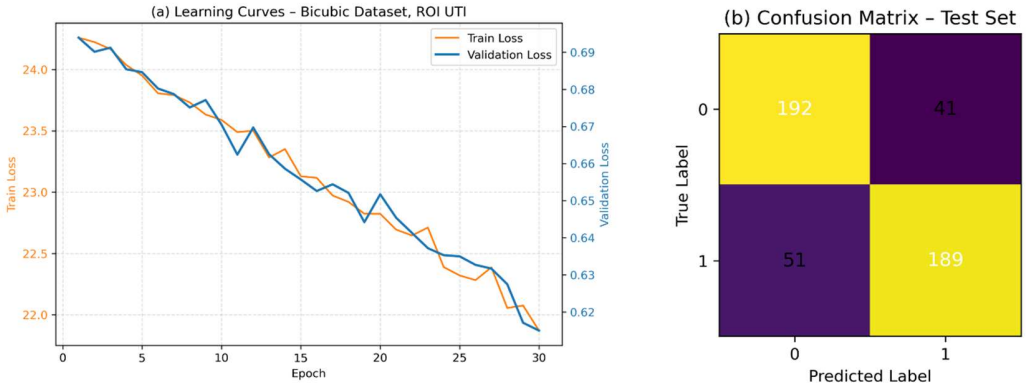

Figure B.5 (a) Training and validation learning curves for the ROI UTI model trained on the bicubic dataset. (b) Confusion matrix on the held-out test set for the same model.

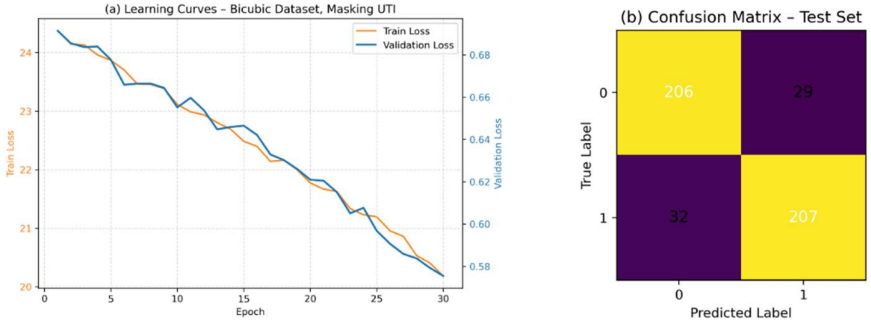

Figure B.6 (a) Training and validation learning curves for the masked UTI model trained on the bicubic dataset. (b) Confusion matrix on the held-out test set for the same model.

B.3 Spline Dataset

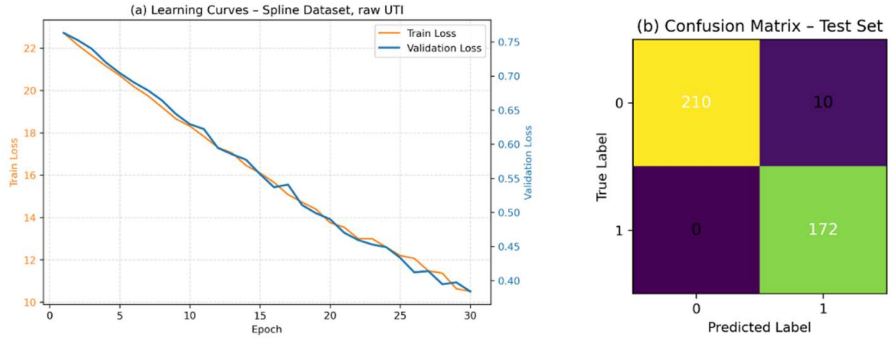

Figure B.7 (a) Training and validation learning curves for the raw UTI model trained on the spline dataset. (b) Confusion matrix on the held-out test set for the same model.

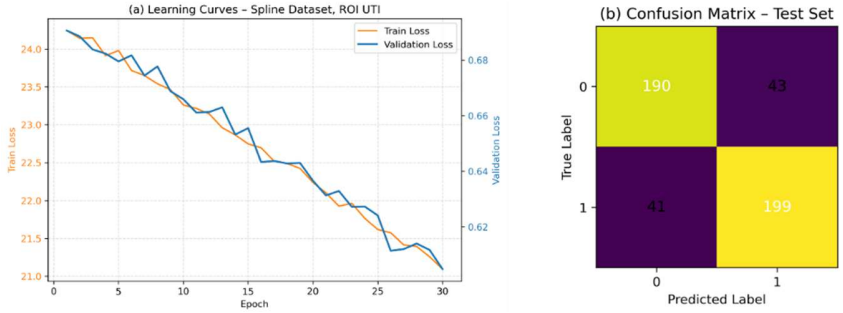

Figure B.8 (a) Training and validation learning curves for the ROI UTI model trained on the spline dataset. (b) Confusion matrix on the held-out test set for the same model.

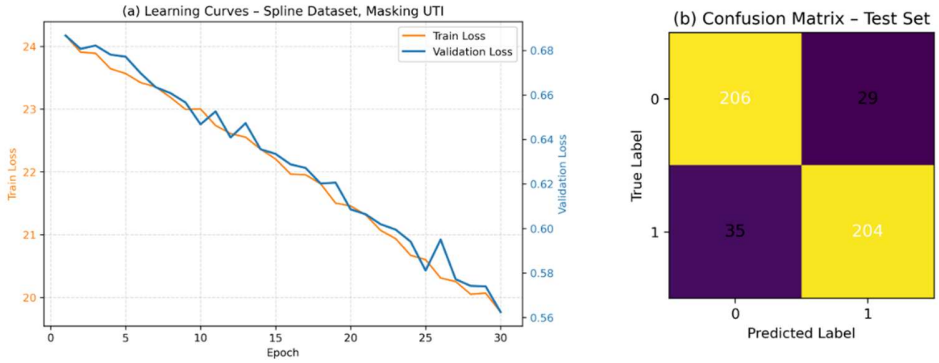

Figure B.9 (a) Training and validation learning curves for the masked UTI model trained on the spline dataset. (b) Confusion matrix on the held-out test set for the same model.

Appendix C: EfficientNet-B0 Architecture Validation

C.1.1 Learning Curve with Confidence Intervals

Figure C.1: EfficientNet-B0 Learning Curve – Accuracy vs. Dataset Size

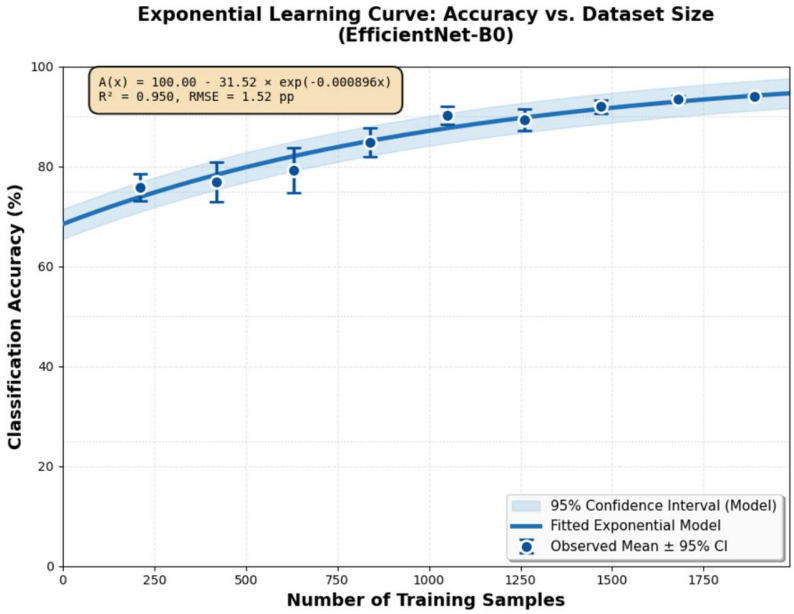

Figure C.1: EfficientNet-B0 Learning Curve.

Table C.1: EfficientNet-B0 Accuracy by Dataset Fraction

Dataset Fraction	Samples	Mean Accuracy (%)	95% CI (± pp)	Predicted (%)	Residual (pp)
10%	210	75.86	±2.61	74.73	+1.13
20%	420	76.93	±3.93	79.43	−2.50
30%	630	79.32	±4.43	83.18	−3.86
40%	840	84.89	±2.85	86.17	−1.28
50%	1050	90.25	±1.78	88.56	+1.69
60%	1260	89.36	±2.17	90.47	−1.11
70%	1470	91.99	±1.28	92.00	−0.01
80%	1680	93.48	±0.72	93.21	+0.27
90%	1890	94.10	±0.44	94.18	−0.08

Table C.2: EfficientNet-B0 Dataset Requirements by Target Accuracy

Target Accuracy	Samples Required	95% CI (samples)	% of Dataset
75%	197	125- 315	9.4%
80%	508	423- 577	24.2%
85%	830	745- 906	39.5%
90%	1,281	1,180- 1,393	61.0%

C.2 Phase 2: AL Performance

C.2.1 AL vs Random Sampling

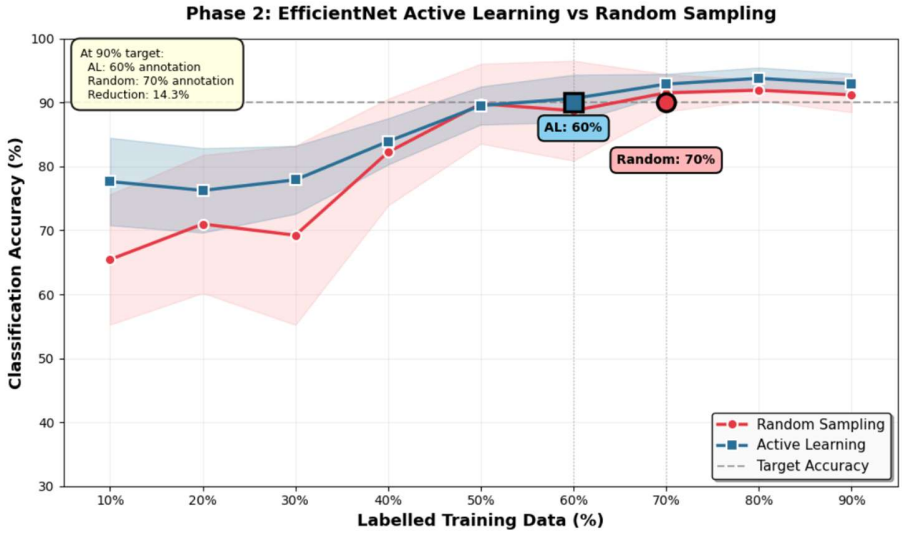

Figure C.2: EfficientNet-B0 AL versus random sampling.

C.3 Combined Cost Analysis

C.3.1 Cost Breakdown by Architecture

Table C.3: Cost Analysis Comparison – AlexNet vs EfficientNet-B0.

Strategy	Architecture	Samples Collected	% Collected	Samples Annotated	% Annotated	Cost (units)
FC/FA (baseline)	-	2,100	100	2,100	100	6300
FC/AL	AlexNet	2,100	100	1,050	50	4,200
FC/AL	EfficientNet	2,100	100	1,260	60	4,620
OC/AL	AlexNet	1,432	68	716	50	2,864
OC/AL	EfficientNet	1,281	61	768	60	2,817

Note: Cost calculated using 1:2 collection: annotation ratio. FC/FA = Full Capture/Full Annotation; FC/AL = Full Capture/Active Learning; OC/AL = Optimised Capture/Active Learning.

This appendix provides comprehensive validation data for EfficientNet-B0, demonstrating that the cost-optimisation framework generalises across architectures through:

1. Phase 1: EfficientNet-B0 required 10.5% fewer samples for 90% accuracy (1,281 vs 1,432) with superior model fit (R^2=0.950 vs 0.939).
2. Phase 2: EfficientNet-B0 showed 14.3% AL efficiency gain versus AlexNet's 28.6%, revealing architecture-dependent ceiling effects.
3. Combined: Both architectures achieved ~55% total cost savings through complementary mechanisms, validating framework robustness.