



UNIVERSITÀ DI PISA



Università di Pisa

Consiglio Nazionale delle Ricerche

National Ph.D. in Artificial Intelligence

XXXVII cycle

# Robust Computer Vision for POCUS: Achieving Reliability in Ultrasound Imaging

Candidate  
Giacomo Ignesti

Supervisors  
Davide Moroni  
Massimo Martinelli

Thesis submitted in partial fulfilment of the requirements for the degree of  
Doctor of Philosophy in Artificial Intelligence





To the realization that the future of AI extends  
far beyond easy Photoshop and Vibe Coding

To the undeniable potential it holds to reshape our world

And to the responsibility that falls upon us:  
to guide this power with wisdom and give it a smiling form



# Summary

## Robust Computer Vision for POCUS: Achieving Reliability in Ultrasound Imaging Extended Abstract

The rapid proliferation of Deep Learning (DL) has fundamentally transformed the scientific landscape, shifting the paradigm of problem solving from explicit, algorithmic logic to data driven feature learning. However, as these technologies permeate high-stakes domains such as medicine and precision agriculture, a critical tension has emerged. The very non-linearity that grants Neural Networks their immense expressive power, as described by the Universal Approximation Theorem, simultaneously causes intrinsic opacity, gradient instability, and a susceptibility to spurious correlations. Consequently, the deployment of these black-box models faces significant hurdles in terms of reliability and trust, necessitating the development of Artificial Intelligence (AI) that is not merely accurate but also fundamentally reliable and interpretable.

## The Clinical Context: Telemedicine and the TiAssisto Project

This thesis grounds its research in the critical context of the SARS-CoV-2 pandemic, which exposed systemic fragilities in healthcare and highlighted the need for resilient, decentralised care models. A significant portion of the applied research was conducted within the TiAssisto project, a server-based web platform designed to support multi-pathological patients by bridging the gap between hospitalisation and home care. To operationalise the vision of 5.0 Medicine, a human-centric approach to healthcare, TiAssisto integrates a hybrid AI framework. This comprises a Strong AI component utilising a rule-based algorithm and logic for monitoring vital signs, and a Weak AI component employing a DL Model for automatic classification of medical imaging. The platform prioritises patient security through strict local encryption and anonymisation protocols, anticipating the growing necessity for Reliable AI architectures. The Challenge of Point-of-Care Ultrasound (POCUS) within this framework, the thesis identifies Lung Ultrasound (LUS) as a pivotal technology for telemedicine, given its portability and safety. However, LUS presents unique challenges for Computer Vision; unlike MRI or CT scans, LUS images are characterised by a low signal-to-noise ratio and are artefact-rich. The diagnosis often relies on interpreting acoustic artefacts, such as "A-lines" (horizontal repetitions indicating healthy lung) or "B-lines" (vertical comet-tail artefacts indicating pathology), rather than direct anatomical visualization. To address the computational constraints of Point-of-Care (POCUS) devices and the scarcity of labelled data, the research proposes a novel Efficient Adaptive Ensembling methodology. This approach utilises the EfficientNet-b0 DL architecture as a backbone due to its optimal balance of accuracy and complexity. By training weak learners on data subsets to induce

diversity and subsequently synthesising their deep features via a trainable linear combination layer, the model achieves robustness without the computational cost of traditional ensembles. When validated on a large public LUS dataset, this ensemble model achieved 100% accuracy in distinguishing between COVID-19, Pneumonia, and Healthy classes, significantly outperforming previous state-of-the-art models. Crucially, it maintained high efficiency, requiring only 0.78 G-FLOPs, making it suitable for real-time deployment on edge devices.

#### Foundations of Reliable AI: From Diagnosis to Constraint

While achieving high accuracy is essential, the thesis argues that clinical adoption hinges on trust. Initial efforts utilised post-hoc Explainable AI (XAI) methods, specifically Grad-CAM, to visually validate that the LUS model focused on clinically relevant areas, such as the pleural line. However, the research identifies a critical limitation in post-hoc analysis: it is purely diagnostic and cannot correct a model's flawed reasoning. To transcend this limitation, the thesis introduces a paradigm shift towards Intrinsically Guided Training. The research proposes moving the generation of Class Activation Maps (CAM) into the training loop itself. By incorporating a custom Contrastive Loss term, the training process actively penalises the model if its internal attention does not align with the relevant input features. This concept evolved into the Batch-CAM paradigm. This advanced methodology replaces instance-specific reconstruction with a Class Prototype approach, where the model compares batch-averaged CAMs against a learned prototype of the target class. This enforces semantic consistency, compelling the model to learn a generalised representation of a class (e.g., the concept of a "T-shirt" or a specific digit) rather than memorising individual examples. Experimental validation on MNIST and Fashion-MNIST demonstrated that this approach yields models with competitive accuracy and significantly superior, human-intelligible internal representations.

#### Conclusion and Future Outlook

The thesis concludes that the path to Reliable AI lies in shifting from passive analysis to active constraint during the learning process. By embedding interpretability into the loss function, models can be guided to learn for the right reasons, mitigating the risk of spurious correlations. The first application of the method to LUS interpretation yields high accuracy, and the systems, now partially guided, seem to learn meaningful anatomical details to inform their decisions. Looking forward, the research outlines a trajectory towards Continual Learning to prevent catastrophic forgetting as medical protocols evolve, and Federated Learning to facilitate privacy-preserving training on decentralised data. Ultimately, integrating these Plausible AI models into hybrid Decision Support Systems offers a structure for the future of healthcare systems that are efficient, transparent, and trustworthy partners in human decision making.

# Acknowledgements

I would like to express my deepest gratitude to Davide Moroni and Massimo Martinelli for their invaluable guidance and constant help throughout this journey. Their mentorship was essential in steering this research in the right direction.

I am grateful to the University of Pisa (UNIPi) and the Institute of Information Science and Technologies "Alessandro Faedo" - ISTI of the National Research Council (CNR-ISTI) for providing me with the opportunity and the environment to conduct this work.

My sincere thanks go to Professors Alessandra Retico and Fosca Giannotti for their academic support and scientific advice, which greatly enriched this thesis.

I would also like to acknowledge Paolo Paradisi and Said Daoudagh for our collaboration and their hard work during the last year.

Finally, I dedicate this work to my father and my mother. Thank you for your resilience, your unconditional support, and for making me who I am today.



# Contents

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| Summary                                                            | i   |
| Acknowledgements                                                   | iii |
| 1 Introduction                                                     | 1   |
| 1.1 Motivations and Challenges                                     | 1   |
| 1.1.1 Gradient Problems                                            | 6   |
| 1.1.2 The Information Bottleneck (IB)                              | 7   |
| 1.1.3 Catastrophic Forgetting                                      | 7   |
| 1.1.4 Ultrasound Interpretation AI Challenge                       | 9   |
| 1.1.5 Challenges for Human and AI Interpretation                   | 10  |
| 1.2 Research Questions                                             | 11  |
| 1.3 Contributions                                                  | 13  |
| 1.3.1 TiAssisto: An Intelligent Platform for Telemedicine          | 13  |
| 1.3.2 Specialised Artificial Intelligence in Lung Ultrasound (LUS) | 14  |
| 1.3.3 Reliable Artificial Intelligence                             | 15  |
| 1.3.4 Broad Scientific Challenges and Opportunities                | 16  |
| 1.4 Publications                                                   | 17  |
| 1.4.1 Journals                                                     | 17  |
| 1.4.2 Conference Papers                                            | 18  |
| 1.4.3 Preprints                                                    | 21  |
| 1.4.4 Magazines and Newsletters                                    | 22  |
| 1.5 Outline of the Work                                            | 23  |
| I AI for medicine                                                  | 25  |
| 2 Telemedicine and the need for AI                                 | 27  |
| 2.1 The Strategic Rise of Telehealth                               | 27  |
| 2.2 TiAssisto                                                      | 28  |
| 2.3 Applying Artificial Intelligence to Temporal and Imaging Data  | 28  |
| 2.4 Ultrasound diagnostic efficiency                               | 30  |
| 2.5 The TiAssisto Project                                          | 30  |
| 2.6 Toward Medicine 5.0                                            | 31  |
| 2.7 From Weak to Strong AI                                         | 31  |
| 2.8 Towards a New Medical Interpretation Approach                  | 33  |

|       |                                                                |    |
|-------|----------------------------------------------------------------|----|
| 3     | Deep Learning for POCUS                                        | 35 |
| 3.1   | Point-of-Care Ultrasound (POCUS)                               | 35 |
| 3.1.1 | The Piezoelectric Effect                                       | 37 |
| 3.1.2 | Sound through media (Human Tissues)                            | 39 |
| 3.1.3 | From Echoes to Beamforming                                     | 40 |
| 3.1.4 | Beamforming to images                                          | 42 |
| 3.1.5 | Challenges and Artefacts                                       | 45 |
| 3.2   | Deep learning: strategies and architecture                     | 47 |
| 3.2.1 | Data preparation                                               | 47 |
| 3.2.2 | Architectures                                                  | 57 |
| 3.2.3 | Efficient Architecture                                         | 59 |
| 3.2.4 | U-net and Transformer                                          | 60 |
| 3.2.5 | The Minimal Adaptive Ensembling Methodology                    | 62 |
| 3.2.6 | Architecture of the Efficient Adaptive Ensemble                | 63 |
| 3.2.7 | Training and Validation Process                                | 65 |
| 3.3   | Case study: LUS Classification                                 | 66 |
| 3.4   | Methodological Framework: The Efficient Adaptive Ensemble      | 67 |
| 3.5   | Experiment 1: DL Model Validation for POCUS                    | 68 |
| 3.6   | Experimental Validation II: Pathology Classification           | 69 |
| 3.6.1 | Performance and State-of-the-Art Comparison                    | 70 |
| 3.6.2 | Explainability and Clinical Plausibility                       | 71 |
| 3.7   | Experimental Validation III: B-Lines                           | 72 |
| 3.7.1 | A Four-Stage Methodological Pipeline                           | 73 |
| 3.7.2 | Stage 1: Data Preparation and Privacy-Preserving Anonymisation | 73 |
| 3.7.3 | Stage 2: Image Augmentation                                    | 74 |
| 3.7.4 | Stage 3-4: Results                                             | 74 |
| 3.7.5 | Grad-CAM usage to assess model viability                       | 75 |
| 3.8   | LUS classification final pipeline                              | 77 |
| II    | Reliable method for deep learning model analysis               | 81 |
| 4     | Loss function and target classification output                 | 83 |
| 4.1   | eXplainable AI (XAI)                                           | 84 |
| 4.2   | Trustworthy AI                                                 | 84 |
| 4.3   | Reliable AI                                                    | 85 |
| 4.3.1 | Grad-CAM Post-Hoc Visual Validations: Benefits and Limits      | 85 |
| 4.3.2 | Evolution to Intrinsically Guided Training Paradigms           | 86 |
| 4.3.3 | A new Approach: Learning for the Right Reasons                 | 86 |
| 4.4   | In-Training Interpretability with Contrastive Loss             | 87 |
| 4.5   | Validation, Challenges, and Algorithmic Refinement             | 88 |
| 4.5.1 | The Initial Failure Case                                       | 88 |
| 4.5.2 | A Successful Refinement                                        | 88 |
| 4.6   | Batch-CAM and Prototypical Reasoning                           | 89 |
| 4.7   | Comprehensive Experimental Analysis of the Batch-CAM Paradigm  | 91 |
| 4.8   | Synthesis and Conclusion: A New Paradigm for Reliable AI       | 92 |



|       |                                                                       |     |
|-------|-----------------------------------------------------------------------|-----|
| 4.9   | Reliable Learning in Point-of-Care Ultrasound . . . . .               | 92  |
| 4.9.1 | The Challenge of Medical Sonography . . . . .                         | 92  |
| 4.9.2 | Methodological Adaptation: Geometry and Noise . . . . .               | 93  |
| 4.9.3 | Implicit Regularization Results . . . . .                             | 93  |
| 4.10  | Future Directions . . . . .                                           | 94  |
| 5     | Advanced AI Integration . . . . .                                     | 97  |
| 5.1   | Integration within Hybrid DSS . . . . .                               | 98  |
| 5.1.1 | Multimodal Feature Fusions . . . . .                                  | 98  |
| 5.1.2 | The Citizen Science Paradigm: Decentralised Data Governance . . . . . | 100 |
| 5.1.3 | Synthesis: Interconnecting Fusion and Crowd-Sensing . . . . .         | 100 |
| 5.2   | Lifelong Learning . . . . .                                           | 101 |
| 5.2.1 | Expanding Plausible Analysis and Alternative AI . . . . .             | 101 |
| 5.3   | The Stability-Plasticity Dilemma LUS . . . . .                        | 101 |
| 5.4   | EWC for Ensemble Maintenance . . . . .                                | 102 |
| 5.5   | Prototype Rehearsal via Batch-CAM . . . . .                           | 102 |
| 5.5.1 | Experimental Validation of CL Protocols . . . . .                     | 103 |
| 5.6   | Federated Continual Learning Architecture . . . . .                   | 104 |
| III   | Future Work and Conclusion . . . . .                                  | 107 |
| 6     | Discussion . . . . .                                                  | 109 |
| 6.1   | Beyond Classification . . . . .                                       | 110 |
| 6.2   | Beyond Reliability: Physical Reality . . . . .                        | 110 |
| 6.3   | Contextual Reasoning: Graph Neural Networks (GNNs) . . . . .          | 112 |
| 6.4   | Chat-Bot: LLM for Reasoning . . . . .                                 | 112 |
| 6.5   | Challenges . . . . .                                                  | 112 |
|       | List of Figures . . . . .                                             | 115 |
|       | List of Tables . . . . .                                              | 119 |
|       | Bibliography . . . . .                                                | 121 |
| A     | Supplementary Material for Chapter 4 . . . . .                        | 133 |
| A.1   | Proofs . . . . .                                                      | 133 |
| A.1.1 | Mathematical Preliminaries . . . . .                                  | 133 |
| A.1.2 | Standard Per-Sample Grad-CAM Analysis . . . . .                       | 133 |
| A.1.3 | Computational Steps . . . . .                                         | 133 |
| A.1.4 | Batch-CAM: Parallelisation and Aggregation Proof . . . . .            | 134 |
| A.1.5 | Algorithm Analysis . . . . .                                          | 135 |



# Chapter 1

## Introduction

"There's no such thing as a painless lesson, they just don't exist. Sacrifices are necessary. You can't gain anything without losing something first. Although if you can endure that pain and walk away from it, you'll find that you now have a heart strong enough to overcome any obstacle. Yeah... a heart made fullmetal."

---

Hiromu Arakawa,  
Fullmetal Alchemist

### 1.1 Motivations and Challenges

The proliferation of Deep Learning (DL) methodologies and their integration with Computer Vision (CV) systems has catalysed a paradigm shift across a broad spectrum of scientific and industrial applications [88]. This transformation is not merely an incremental advancement; it represents a fundamental change in how complex problems are solved, moving from explicit, human-programmed logic to data-driven, hierarchical feature learning. This revolution is powered by a joint cycle of massive datasets, accelerating computational power (primarily through GPUs and specialised hardware), and rapid algorithmic innovation.

In CV, the advent of architectures such as the Vision Transformer (ViT) [29] has revolutionised fields such as autonomous driving and logistics, enabling robust, real-time object detection and semantic segmentation. Concurrently, foundational Convolutional Neural Networks (CNNs) [74] and their specialised variants, most notably the U-Net architecture [107], have become indispensable in medicine, enabling rapid and highly accurate diagnostics through sophisticated medical image classification or segmentation. The influence of DL, however, extends far beyond visual data. Generative AI, driven primarily by Large Language Models (LLMs) [100] that adapt the same core Transformer principles, has captured public and scientific attention,

starting a new era of interactive agents, autonomous systems, and advanced applications in computational biology [65]. In the distinct domain of sequential data analysis, traditional recurrent architectures, particularly the gated structures of Long Short-Term Memory (LSTM) and their update implementation, Gated Recurrent Units (GRU), have long served as foundational architectures [18, 45]. These models provide powerful non-linear counterparts to classical statistical methods, such as the AutoRegressive Integrated Moving Average (ARIMA) model [12]. This landscape continues to evolve rapidly, with emerging methodologies such as Reservoir Computing and state-of-the-art architectures like MAMBA, which leverage Selective State Spaces (S6) to demonstrate unprecedented efficiency and efficacy in analysing complex temporal relationships and long-range dependencies [39]. Furthermore, the core principles of DL are profoundly impacting other domains of computer science, notably creating the field of Deep Reinforcement Learning (DRL) and enabling the analysis of non-Euclidean data through Graph Neural Networks (GNNs). DRL [90] leverages DNNs as powerful function approximators, overcoming the curse of dimensionality that hampered traditional reinforcement learning and effectively scaling decision-making to high-dimensional state spaces. Concurrently, DL's hierarchical feature-learning paradigm has been adapted to graph-structured data, resulting in GNNs. These models employ a mechanism of neighbourhood aggregation or message passing to iteratively update node representations based on local topology, allowing them to effectively model complex relational patterns for tasks such as link prediction and graph-level classification [112]. However, beyond superior performance, there is a critical need to develop AI systems that are not only accurate but also reliable, adaptable, and interpretable; particularly regarding model-based systems utilizing Neural Network (NN) predictors. Trust issues emerged during the early widespread adoption of NNs, as it was observed that models could produce correct predictions for the wrong reasons. A classic instance of this is the wolf/husky classification example described in the seminal work of Ribeiro et al. [103]. As shown in Figure (Fig.) 1.1, the model erroneously classified images as "wolf" but did so based solely on the presence of snow in the background, rather than the features of the animal itself.

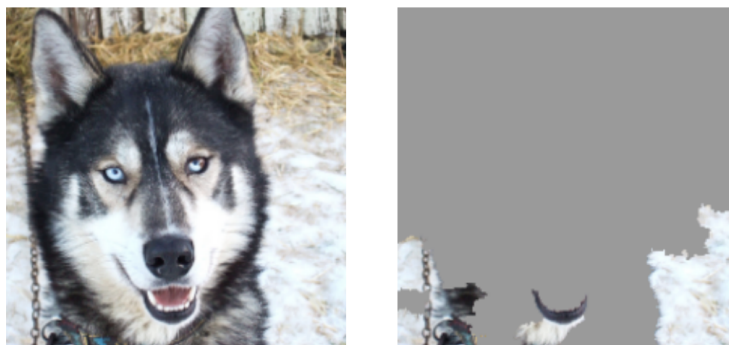


Figure 1.1: A husky dog wrongly classified as a wolf due to the snow in the image, example of the clever Hans effect in DL study, image from Ribeiro et al [103]

The difficulty in understanding the internal mechanisms of fully connected (FC) NNs stems directly from two properties: their approximation power, as described

by the Universal Approximation Theorem (UAT), and the profound compositional nonlinearity introduced by their activation functions. These two factors are, in fact, deeply intertwined. The UAT [47] provides the theoretical foundation for the power of these networks. It states, in essence, that a feed-forward network with a single hidden layer containing a sufficient number of neurons (employing a non-linear activation function selected from a wide set of possibilities) can approximate any continuous function on a compact subset of  $\mathbb{R}^n$  to an arbitrary degree of accuracy. While this theorem is a powerful guarantee of the network's expressive capacity, it is fundamentally an existence proof. It guarantees that a set of weights  $[W]$  and biases  $[b]$  exists to form the approximation. Still, it provides no insight into how this approximation is structured, how to efficiently find these parameters, or what the resulting high-dimensional parametrisation  $\theta = W, b$  means. The optimisation process, such as stochastic gradient descent (SGD) [11], finds a solution within a vast, non-convex loss landscape that may be neither parsimonious nor interpretable. The knowledge acquired by the model is not localised; instead, it is distributed and entangled across millions of parameters, rendering a direct inspection of the weight matrices largely uninformative. This challenge is compounded exponentially by the architecture's highly non-linear nature. An FC-NN computes its output through a recursive composition of functions. A single layer performs an affine transformation, as expressed in Eq 1.1.

$$z = Wx + b \quad (1.1)$$

This is a simple linear operation. The critical step is the element-wise application of a generic non-linear activation function  $a = \sigma(z)$ , such as the sigmoid Eq. 1.2 or the Rectified Linear Unit (ReLU) 1.3:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (1.2)$$

$$\sigma(z) = \max(0, z). \quad (1.3)$$

Without this non-linearity, a deep stack of layers would mathematically collapse into a single equivalent linear transformation,  $W_{eff} = W_L \cdot \dots \cdot W_2 \cdot W_1$ , thus possessing no more representative power than a simple linear model. The very non-linearity required for expressive power is the primary source of opacity. In a deep network, the final output  $y$  is a deeply nested, recursive function, Eq. 1.4 of the input  $x$ :

$$y = \sigma_L(W_L(\sigma_{L-1}(W_{L-1} \dots \sigma_1(W_1 x + b_1) \dots + b_{L-1})) + b_L) \quad (1.4)$$

This mapping from  $x$  to  $y$  is analytically intractable and exhibits extraordinarily complex behaviour. It is impossible to symbolically solve this equation to understand the global contribution of a single input feature  $x_i$ .

The activation functions create intricate, non-linear decision boundaries by effectively warping and folding the high-dimensional feature space, Fig. 1.2.

The contribution of any single input is highly context-dependent, mediated by the activations of all other neurons in computationally opaque ways. The UAT guarantees the network can represent the solution, and the deep compositional non-linearity is the mechanism it uses to build that representation. This mechanism is inherently antithetical to simple human symbolic inspection. The UAT is thus

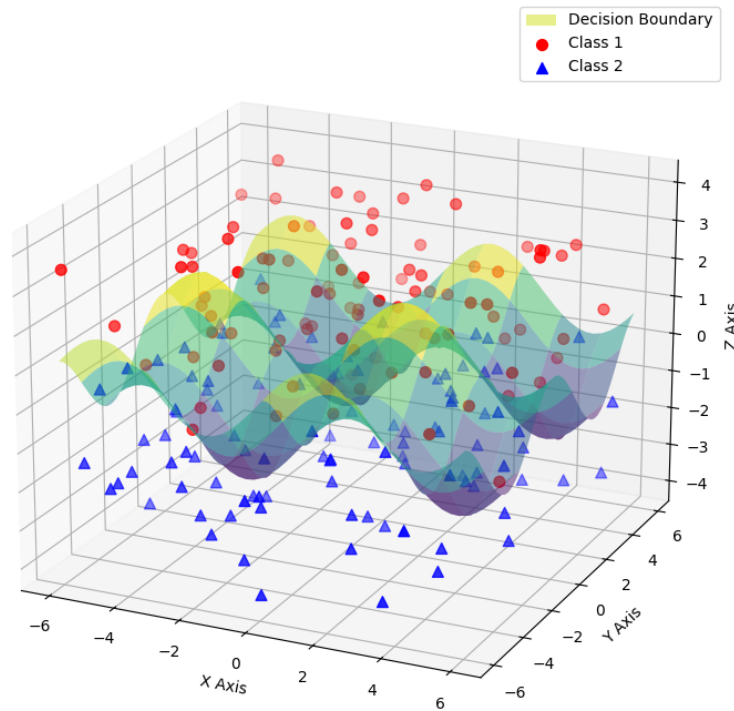


Figure 1.2: A 3-D example of a nonlinear decision boundary

both the reason for the immense power of DL models and, in a sense, the source of their most significant challenges. The deep, extensive structure built to leverage this representative power produces models that can be highly accurate but also require considerable control and oversight. This massive combination of non-linear layers and transformations leads to several key issues, as shown in the Fig. 1.3:

- **Opacity:** This is a direct consequence of the UAT and the deep network structures. The solution is distributed across millions of non-linear parameter interactions, making the model's internal decision-making process a black-box that is not human-readable.
- **Gradient problems:** The depth and complexity of the networks can lead to vanishing or exploding gradients. During backpropagation, these gradients can become exponentially small (hindering learning in early layers) or large (causing instability), which severely hinders the network's learning capability.
- **Information bottleneck:** The network must find an optimal trade-off between compression (compacting the high-dimensional input) and representation (preserving the information necessary to solve the task). The network must learn what to discard and what to retain.
- **Catastrophic forgetting:** This occurs when a network is trained sequentially. As the model's weights are updated to learn a new task, it often overwrites or forgets the knowledge patterns it had previously learned for another task.

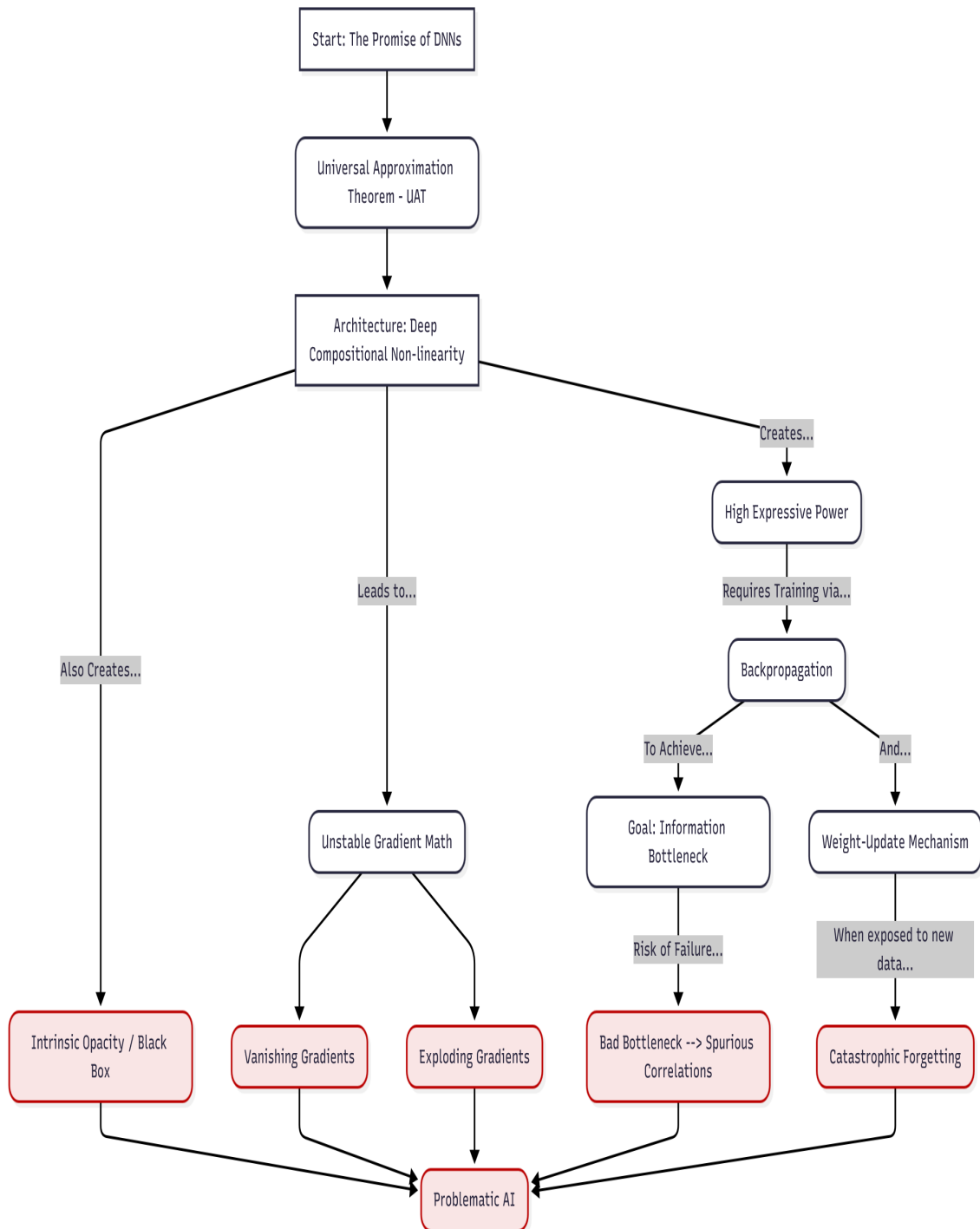


Figure 1.3: An ideal flow chart to DL network issues

### 1.1.1 Gradient Problems

The process of training a neural network relies on back-propagation, a practical application of the chain rule that computes the gradient of the loss function  $\mathcal{L}$  with respect to each weight  $\mathbf{w}$  in the network. For a deep, feed-forward network with  $L$  layers, the gradient for a weight  $\mathbf{w}_{ij}^l$  in layer  $l$  is shown in the following Eq. 1.5

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_{ij}^l} = \frac{\partial \mathcal{L}}{\partial a^L} \left( \prod_{k=l+1}^L \frac{\partial a^k}{\partial a^{k-1}} \right) \frac{\partial a^l}{\partial \mathbf{w}_{ij}^l} \quad (1.5)$$

The term  $\frac{\partial a^k}{\partial a^{k-1}}$  represents the local gradient at layer  $k$ . This term is a product of the layer's weights  $\mathbf{W}^k$  and the derivative of its activation function  $\sigma'$ , Eq. 1.6.

$$\frac{\partial a^k}{\partial a^{k-1}} = \text{diag}(\sigma'(z^k)) \cdot (\mathbf{W}^k)^T \quad (1.6)$$

where  $z^k = \mathbf{W}^k \mathbf{a}^{k-1} + \mathbf{b}^k$  is the pre-activation at layer  $k$ .

The core problem lies in the long product of these Jacobian matrices, Eq. 1.7:

$$\prod_{k=l+1}^L \frac{\partial a^k}{\partial a^{k-1}} = \prod_{k=l+1}^L \text{diag}(\sigma'(z^k)) \cdot (\mathbf{W}^k)^T \quad (1.7)$$

#### Vanishing Gradients

This problem is particularly acute with sigmoidal activation functions (e.g., logistic sigmoid or  $\tanh$ ). The derivative of the sigmoid function,  $\sigma'(x) = \sigma(x)(1 - \sigma(x))$ , has a maximum value of 0.25.

If the weights  $\mathbf{W}^k$  are initialised to be small (e.g., with a variance less than 1), the norm of each term  $\left\| \frac{\partial a^k}{\partial a^{k-1}} \right\|$  will be, on average, less than 1. As these terms are multiplied together  $L-l$  times, the gradient magnitude decreases exponentially with depth, Eq. 1.8.

$$\left\| \frac{\partial \mathcal{L}}{\partial \mathbf{a}^l} \right\| \approx \left\| \frac{\partial \mathcal{L}}{\partial \mathbf{a}^L} \right\| \cdot \prod_{k=l+1}^L \left\| \frac{\partial a^k}{\partial a^{k-1}} \right\| \xrightarrow{L-l \rightarrow \infty} 0 \quad (1.8)$$

This results in the gradients for the early layers (small  $l$ ) becoming vanishingly small, causing them to learn extremely slowly or not at all.

#### Exploding Gradients

Conversely, if the weights  $\mathbf{W}^k$  are initialised to be large, the norm of the local gradients  $\left\| \frac{\partial a^k}{\partial a^{k-1}} \right\|$  can be, on average, greater than 1. When multiplied together, this results in exponential growth of the gradient magnitude, Eq. 1.9.

$$\left\| \frac{\partial \mathcal{L}}{\partial \mathbf{a}^l} \right\| \xrightarrow{L-l \rightarrow \infty} \infty \quad (1.9)$$

This causes the weight updates to be enormous, leading to numerical instability (e.g., NaN values) and divergence. Solutions to these problems include careful



weight initialisation (e.g., Glorot [35]), non-saturating activation functions such as ReLU, batch normalisation, and architectural changes such as residual connections (ResNets).

### 1.1.2 The Information Bottleneck (IB)

The Information Bottleneck principle, introduced by Tishby, Pereira, and Bialek, [127] frames learning as a problem of optimal compression. The goal is to find a compressed representation (a bottleneck)  $Z$  of an input  $X$  that retains as much information as possible about a target variable  $Y$ . This is formalised as a constrained optimisation problem using mutual information: we want to find a stochastic mapping  $p(z|x)$  that minimises the compression cost  $I(X; Z)$  while maximising the relevance  $I(Z; Y)$ . The IB Lagrangian, Eq. 1.10, captures this trade-off:

$$\mathcal{L}_{IB}[p(z|x)] = I(X; Z) - \beta I(Z; Y) \quad (1.10)$$

where  $\beta$  is a Lagrangian multiplier that balances compression ( $\beta \rightarrow 0$ ) versus relevance ( $\beta \rightarrow \infty$ ). Minimising  $\mathcal{L}_{IB}$  for a given  $\beta$  yields a point on the optimal trade-off curve, known as the "Information Plane". A DNN can be interpreted as an IB process. The layers form a Markov chain  $X \rightarrow H_1 \rightarrow \dots \rightarrow H_L \rightarrow \hat{Y}$ , where each hidden layer  $H_l$  acts as a representation  $Z_l$ . The network, trained via SGD, implicitly optimises the IB objective:

1. Compression  $I(X; Z)$ : The network compresses the high-dimensional input  $X$  into the lower-dimensional hidden representations  $Z = H_l$ . This is  $I(X; H_l)$ ,
2. Relevance  $I(Z; Y)$ : The network must retain the information in  $H_l$  that is predictive of the target  $Y$ . This is  $I(H_l; Y)$ .

The challenge is that SGD does not explicitly control  $\beta$ . The network's architecture (layer widths) and regularisation (e.g., dropout or weight decay) implicitly define this trade-off. An over-compressed network (underfitting) loses information vital for prediction, while an under-compressed network (overfitting) retains spurious, instance-specific details from  $X$  that do not generalise.

### 1.1.3 Catastrophic Forgetting

Catastrophic forgetting, also known as catastrophic interference, is a critical failure in continual learning (CL) where a model is trained on a series of tasks  $\mathcal{T}_A, \mathcal{T}_B, \dots, \mathcal{T}_N$ . Let  $\theta$  be the parameters of the model. After training on  $\mathcal{T}_A$ , the parameters converge to  $\theta_A^*$ , which minimizes the loss  $\mathcal{L}_A$ :

$$\theta_A^* = \arg \min_{\theta} \mathcal{L}_A(\theta) \quad (1.11)$$

When the model subsequently trains on  $\mathcal{T}_B$ , it updates the parameters, Eq 1.11, to minimise  $\mathcal{L}_B$ . A standard SGD update can be represented as the following Eq. 1.12.

$$\theta_{\text{new}} = \theta_A^* - \eta \nabla_{\theta} \mathcal{L}_B(\theta_A^*) \quad (1.12)$$

The problem is that the gradient  $\nabla_{\theta}\mathcal{L}_B$  does not know  $\mathcal{L}_A$ . The update is free to move  $\theta$  into a region of the parameter space that, while good for  $\mathcal{T}_B$ , has a very high loss for  $\mathcal{T}_A$ . Parameters that were critical for performance on  $\mathcal{T}_A$  are overwritten, and the original knowledge is lost.

#### Formalising with Elastic Weight Consolidation (EWC)

A prominent solution, Elastic Weight Consolidation (EWC), formalises this problem using a Bayesian perspective. It models the new loss function for  $\mathcal{T}_B$  as Eq. 1.13:

$$\mathcal{L}'(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i (\theta_i - \theta_{A,i}^*)^2 \quad (1.13)$$

Here,  $\lambda$  is a hyperparameter controlling the elasticity,  $\theta_{A,i}^*$  is the optimal value for parameter  $i$  from Task A, and  $F_i$  is the diagonal element of the Fisher Information Matrix for Task A, evaluated at  $\theta_A^*$ . The Fisher Information  $F_i$ , Eq 1.14, is a measure of the importance of parameter  $\theta_i$  to Task A. It is defined as:

$$F_i = \mathbb{E}_{x \sim \mathcal{T}_A} \left[ \left( \frac{\partial}{\partial \theta_i} \log p(y|x, \theta_A^*) \right) \right]^2 \quad (1.14)$$

This modified loss  $\mathcal{L}'(\theta)$  creates a quadratic penalty that anchors the parameters to their old values  $\theta_A^*$ . This anchor is elastic:

- If  $F_i$  is large, parameter  $\theta_i$  is very important for  $\mathcal{T}_A$ , and the penalty for changing it is high.
- If  $F_i$  is small, parameter  $\theta_i$  is unimportant for  $\mathcal{T}_A$ , and it can be changed freely to learn  $\mathcal{T}_B$ .

This provides a principled mathematical framework to protect previously learned knowledge.

This need is particularly pronounced in high-risk contexts, such as medicine and precision agriculture, where AI-based decisions can have a direct and significant impact on people's lives or the management of vital resources. Existing architectures, such as plain-net, VGG [125], ResNet [42], and CSPNet [137], exhibit varying degrees of information loss during the feedforward process. Although methods such as reversible architectures, masked modelling, and deep supervision aim to mitigate this problem, they have drawbacks, including increased inference costs for reversible architectures and conflicts between losses and error accumulation in masked modelling and deep supervision. In particular, deep supervision is often inadequate for lightweight neural networks and can lead to under-parameterisation, thereby worsening their performance. These problems are more evident under challenging tasks and with small-sized models [50].

Implementing AI-based solutions faces significant challenges in medicine applications [129], such as telemedicine. The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) pandemic highlighted the need to improve methods and structures for managing rapidly spreading pathogens and diseases in an ageing population. The TiAssisto project [53], for example, was born precisely in response to this emergency, aiming to reduce avoidable hospitalisations and improve patients'

quality of life. However, the integration of AI in medical contexts requires not only high performance, but also that systems are safe, understandable, and used as an additional opinion, with the final decision always resting with the clinician. The scarcity of large public medical datasets is a common challenge, necessitating the development of data-generation and augmentation strategies to address it. The privacy and security of sensitive patient data are mandatory aspects, requiring robust solutions such as encryption and anonymisation [105]. In summary, the motivations and challenges emerging from the sources converge on the need for an AI that is not only effective but also understandable, robust, adaptable, and reliable, especially in critical applications where user trust and decision correctness are paramount [44].

Similarly, in other scientific applications, ensuring AI reliability remains a crucial challenge. For example, in agriculture, despite significant progress in computer vision and AI in precision agriculture, the adaptability of AI methods to new cases and domains is not always straightforward. The agricultural environment is highly dynamic [5, 67], influenced by seasonal variations, i.e. light exposure, pest outbreaks, and temporal factors. AI models must not only be able to interpolate known data but also to extrapolate to predict the impact of unprecedented climate scenarios. There is a scarcity of complete, high-quality datasets, especially for specific crops or regions. The generalisation ability of models to uncontrolled, real-world images is often unsatisfactory, with significant performance degradation exceeding 20%. This results in unreliable performance and inconsistency, which is unacceptable for users in real-world contexts and hinders the technology’s adoption [1]. Despite the availability of benchmarking datasets in agriculture, foundational models have not yet been trained in this domain, necessitating either transfer learning or training from scratch.

#### 1.1.4 Ultrasound Interpretation AI Challenge

Beyond the theoretical challenges of gradient propagation and knowledge retention, a significant frontier for DL lies in domains where data interpretation is inherently ambiguous and expert-dependent. Medical ultrasound (US), and particularly Lung Ultrasound (LUS), is a prime example of this problem. The difficulty of interpreting these scans, for both humans and AI, is precisely why robust AI solutions are critical to advancing telemedicine and the vision of Medicine 5.0.



Figure 1.4: A side-to-side comparison of chest medical imaging scans between US, CT, and MRI; the clear difference is that anatomical marks from the latter are clear, while US need further interpretation.

Unlike imaging modalities such as CT or MRI, ultrasound is fundamentally an

operator-dependent technique whose interpretation is challenging due to the way the data are displayed and the prevalence of artefacts, as shown in Fig. 1.4.

- **Operator Dependence:** The quality and diagnostic value of an ultrasound image depend entirely on the skill of the sonographers, specifically their ability to find the correct acoustic window, apply appropriate probe pressure, and optimise machine settings in real-time,
- **Artefacts as Information:** US images are filled with artefacts (e.g., acoustic shadowing, reverberation, posterior enhancement) that are not real anatomy. For a novice, these artefacts are noise; for an expert, they might represent essential diagnostic clues,
- **Low Signal-to-Noise Ratio:** Images are often "grainy" or "snowy", making fine anatomical distinctions difficult.

This reliance on expert-driven, real-time interpretation makes standardisation incredibly difficult.

### The Unique Case of Lung Ultrasound

Lung ultrasound represents an extreme case of this ambiguity because it is an exam based almost entirely on artefacts. A healthy, air-filled lung reflects nearly all ultrasound beams, making the lung parenchyma itself invisible. The diagnosis is made by interpreting the artefacts that arise from the pleural line (the lung's surface).

- **A-lines:** In a healthy lung, the pleural line is mirrored, creating horizontal, repeating artefacts known as A-lines.
- **B-lines:** In a wet lung (e.g., pulmonary oedema, interstitial syndrome), vertical, comet-tail-like artefacts called B-lines emerge from the pleural line.
- **Consolidations:** In pneumonia, parts of the lung fill with fluid and become solid-like, creating a hepatized (liver-like) appearance.

This diagnostic process is profoundly abstract. A human clinician must be trained to detect pathology in patterns of noise sources that other imaging-trained individuals would be taught to ignore.

### 1.1.5 Challenges for Human and AI Interpretation

The artefact-based nature of LUS creates a significant barrier to entry for practical interpretation by both human clinicians and AI systems. For human operators, interpretation is notoriously subjective. A pattern of B-lines deemed moderate by one clinician might be classified as severe by another. This high inter-observer variability remains a significant obstacle to its widespread adoption as a standardised tool, despite the modality's low cost and safety profile, necessitating extensive, specialised training for proficiency [82, 136]. Similarly, this domain poses unique challenges to traditional computer vision models aimed at automating interpretation. Unlike

models trained on natural image datasets, such as ImageNet [26], which are optimised to be invariant to noise, an effective LUS AI must be explicitly taught to recognise specifically the B-lines artefacts as the primary signal rather than noise to be filtered out. Furthermore, developing robust models requires vast quantities of high-quality, expertly labelled video data, as static images may miss crucial dynamic features such as lung sliding. Acquiring such datasets is both expensive and challenging [92]. Finally, these models face significant generalisation issues stemming from operator variability; an AI trained on scans acquired by an expert using a specific probe may fail to generalise to images acquired by a novice utilising different techniques, mirroring generalisation failures seen in other DL applications [36]. The very difficulty of LUS interpretation is what positions it as a prime application for AI in medicine, as an effective AI solution could address core problems of standardisation and accessibility. By providing quantitative, reproducible analysis, such as automated B-line counting or estimated LUS scores, AI models can transform previously subjective assessments into objective metrics. This objectivity is crucial [132] for accurately tracking disease progression or triaging patients in emergency settings. The most significant impact of AI lies in improving high-end diagnosis point-of-care ultrasound (POCUS) by empowering non-experts. With the increasing availability of portable, smartphone-connected ultrasound probes, AI systems could guide users – such as nurses, paramedics, or general practitioners in rural clinics – to acquire diagnostically viable images through real-time feedback, though developing an AI-guidance [24] constitutes a non-trivial DL problem in itself. This combination of POCUS and AI guidance forms the cornerstone of advanced telemedicine paradigms. For instance, a paramedic in an ambulance could acquire a LUS scan, utilise AI for preliminary analysis to indicate the probability of pathology, such as severe pulmonary oedema, and transmit this data to a hospital. In this scenario, the remote specialist receives pre-screened, high-quality scans accompanied by a quantitative report rather than raw, potentially uninterpretable video footage. This paradigm aligns perfectly with the vision of Medicine 5.0, which emphasises human-AI collaboration and decentralised, personalised care [60]. In this context, the AI does not replace the expert clinician but rather augments the non-expert, allowing specialised diagnostic power to shift from the central hospital to the patient’s bedside, home, or local clinic [128].

## 1.2 Research Questions

The challenges identified in the context of AI applied to medical diagnostics and reliable systems give rise to several critical research topics addressed in this work.

1. DL Models for Medical Imaging: The investigation of DL reveals a fundamental trade-off between model complexity, information compression, and the preservation of task-relevant features. This research area addresses the architectural challenges of propagating salient information through computationally constrained architectures to ensure stable gradient flow without the overhead of monolithic models.

- Architectural Efficiency vs. Feature Preservation: How can we design architectures that explicitly counter the computational costs of traditional ensembles while maximising feature prediction power? This involves analysing the efficacy of Efficient Adaptive Ensembling, which synthesises deep features from lightweight backbones (e.g., EfficientNet-b0 ) via a trainable linear combination layer to achieve State-of-the-Art (SOTA) accuracy with minimal G-FLOPs.
  - Methodologies for Weak Learners: What principled methodologies can be developed to integrate weak learners into a strong diagnostic tool? This requires investigating whether deep feature concatenation and fine-tuning can serve as a differentiable alternative to classical boosting or heuristic loss weighting (e.g., PCGrad), effectively filtering out spurious correlations introduced by individual learners to achieve Pareto-optimal solutions.
2. Algorithmic Transparency via Guided Training: The deployment of AI in safety-critical domains like clinical diagnostics necessitates models that are not only accurate but fundamentally interpretable. This thrust focuses on shifting Explainable AI (XAI) from a post-hoc diagnostic tool to an active component of the learning objective.
- Intrinsic Interpretability: To what extent can interpretability be embedded directly into the loss landscape to overcome the fidelity limitations of post-hoc methods (e.g., LIME, SHAP)? The objective is to develop Intrinsically Guided Training paradigms, such as Batch-CAM, which utilise a custom Contrastive Loss to penalise the model if its internal attention maps deviate from learned class prototypes.
  - Prototype-Based Reasoning: How can prototype-based reasoning be leveraged to enforce semantic consistency? This involves developing frameworks in which the model compares batch-averaged Class Activation Maps (CAMs) with a target prototype during training, compelling the network to learn generalizable features (e.g., anatomical structures) rather than memorising instance-specific noise.
3. Decentralised Edge Deployment: This research area addresses the need to optimise the trade-off between predictive performance and computational complexity (latency, memory footprint, energy) for POCUS applications.
- Edge-Compatible Performance: How can SOTA diagnostic performance be realised within the strict resource constraints of edge devices? This involves investigating the use of EfficientNet-based backbones and Anatomical Isolation using the new conceptual training method that automatically output activation maps that act as segmentation mask,
  - Scalable Decentralised Training: What strategies enable scalable training across decentralised clinical nodes? This research aims to develop Federated Learning (FL) architectures that allow local model adaptation (computing gradients on device) while aggregating updates centrally, ensuring that sensitive data remains on-site while the global model improves.

4. Enhancing Model Robustness and Maintainability through Continual Learning (CL): A critical challenge is the deployment of AI in non-stationary medical environments where models must adapt to evolving data streams (e.g., new probes, pathologies, or domain shifts) without suffering from catastrophic forgetting.
  - The Stability-Plasticity Dilemma: How can the stability/plasticity dilemma be effectively managed in privacy sensitive medical contexts? This requires a systematic evaluation of regularisation strategies, specifically Elastic Weight Consolidation (EWC), to anchor critical parameters while allowing adaptation to new tasks.
  - Privacy-Preserving Replay: Can privacy preserving replay mechanisms be developed to support lifelong learning? This research explores the utility of Prototype Rehearsal and potentially generative models for synthetic augmentation where learned class prototypes rather than raw patient data are retained to regularise future training, maintaining diagnostic performance without compromising patient data sovereignty.
5. Human-AI Collaboration within Hybrid Decision Support Systems: The ultimate goal of applied AI is to augment human intelligence within a clinical workflow. This research focuses on designing integrated Decision Support Systems (DSS) that fuse predictive outputs with medical logic and robust uncertainty quantification.
  - Multimodal Integration: How can AI systems effectively integrate multimodal data streams to enhance diagnostic reliability? This involves developing Multimodal Feature Fusion techniques that concatenate visual embeddings from ultrasound analysis with tabular clinical metadata, mimicking the holistic reasoning of a physician.
  - Data Governance in Citizen Science: What mechanisms can ensure data quality when leveraging non-expert operators (Citizen Science paradigm)? This necessitates the creation of Gatekeeper modules that utilise blind quality assessment and semantic consistency checks to filter out noise and uncertain data before it reaches the diagnostic engine, thereby ensuring the reliability of the Hybrid DSS (e.g., TiAssisto).

## 1.3 Contributions

As detailed in the publications list in Section 1.4, my work spans from the development of innovative methodologies for reliable and explainable Artificial Intelligence (AI) to their application across diverse scientific contexts, with a particular focus on telemedicine and LUS analysis.

### 1.3.1 TiAssisto: An Intelligent Platform for Telemedicine

The TiAssisto project represents a significant contribution to telemedicine, designed to support patients with multiple pathologies, particularly during the COVID-19



emergency. The paper "An intelligent platform of services based on multimedia understanding and telehealth for supporting the management of SARS-CoV-2 multi-pathological patients" describes a server-based web platform developed to assist patients and healthcare professionals, thereby reducing avoidable hospitalisations. The developed platform follows a Model-View-Controller (MVC) pattern and employs a User-Centred Design (UCD) to ensure clinical acceptance. A key focus is privacy: personal data is encrypted and anonymised with unique IDs, ensuring that only authorised clinicians can access sensitive information. TiAssisto integrates two distinct forms of AI:

- Strong AI (Rule-Based DSS): Utilises medical rules to provide diagnostic advice and assess health trends via numerical scores.
- Weak AI (ML/DL): Employs DL for the analysis of LUS images and ML (e.g., XGBoost) to predict patient health status based on vital parameters.

It is worth noting that the definition of Strong AI employed here refers specifically to Symbolic AI-systems governed by deterministic, human-defined logic rather than stochastic data patterns. However, the broader field of AI has undergone a semantic shift in how this term is understood following the rise of Generative AI. Previously, systems were characterised by a lack of massive data dependency, relying instead on expert systems and logic programming. With the advent of modern generative models, which are fundamentally data-driven yet exhibit generalised capabilities, the term Strong AI is increasingly used to denote Artificial General Intelligence (AGI). Unlike the rule-based core system of TiAssisto, modern AGI implies a system capable of broadly applying intelligence to any problem, a capability that emerged only when models shifted from rigid logical structures to massive-scale probabilistic learning. The project protocol involves a randomised controlled trial with 300 patients, aiming to validate the impact of this hybrid AI approach on the management of frail subjects and promote a shift towards Medicine 5.0. This concept encompasses all new technologies, enabling the widespread application of healthcare.

### 1.3.2 Specialised Artificial Intelligence in Lung Ultrasound (LUS)

A core component of the TiAssisto project and the main topic of this thesis research is the development of specialised AI for POCUS. The paper "Efficient Lung Ultrasound Classification" proposes a methodology that optimises the trade-off between accuracy and computational complexity, a key requirement for real-time diagnostics. The solution utilises an EfficientNet-b0 architecture, chosen for its inverted residual blocks and compound scaling. The primary contribution is an efficient adaptive ensembling method. This approach combines two pre-trained EfficientNet-b0 weak models via a trainable linear combination layer at the deep feature level. This method reduces computational costs by up to 20% compared to standard ensembles while maintaining high parameter efficiency. On a large public LUS dataset, the model achieved 100% accuracy, surpassing previous state-of-the-art benchmarks. Crucially, the research addresses explainability to ensure medical safety. Using Gradient-weighted Class Activation Mapping (Grad-CAM), the study generated



saliency maps demonstrating that the model focuses on clinically significant anatomical areas: upper lung regions for COVID-19 and diffuse lower regions for pneumonia. The practical implementation of this methodology is detailed in "Deep learning methods for point-of-care ultrasound examination." This work outlines a complete processing pipeline:

1. Data Anonymisation: An OCR-based algorithm utilising Convolutional Recurrent Neural Networks (CRNN) automatically detects and masks text in ultrasound images to preserve privacy.
2. Annotation: A custom Java tool allows for frame-by-frame video annotation.
3. Classification: The adaptive ensemble model classifies pathological signs.
4. Explainability: Grad-CAM provides a possible second opinion visual check for the clinician.

This pipeline confirms that efficient, explainable AI can be effectively deployed in POCT scenarios, with future directions pointing towards federated learning and direct video segmentation using Vision Transformers.

### 1.3.3 Reliable Artificial Intelligence

The works "Youve Got the Wrong Number: Evaluating Deep Learning Training Paradigms Using Handwritten Digit Recognition Data" and its follow-up "Youve Got the Wrong Outfit: Evaluating Deep Learning Paradigms on Digit and Fashion Recognition" lay the groundwork for more transparent and reliable AI. The research explores a new training paradigm for AI algorithms. The study focuses on the need to understand the mechanisms that guide classification systems in identifying their targets, emphasising the development of more accurate and reliable AI algorithms. Traditionally, explanation methods, such as LIME, SHAP, and Grad-CAM, are applied post hoc to interpret the decisions of a pre-trained model. This work proposes moving the reconstruction of the class activation map (CAM), particularly through a Grad-CAM-like function, into the training phase. The goal is to provide the network with a reference during training to ensure that its outputs are meaningful and interpretable. The key contribution of this approach is the introduction of a contrastive loss term to supplement the standard classification function (typically cross-entropy). This contrastive loss aims to optimise the similarity between the Grad-CAM reconstructed image and the original input, based on the given label. The underlying concept is that if the network focuses on the relevant features of the object (e.g., the shape of a digit in MNIST), its reconstruction will be more faithful to the input, providing more reliable gradients for weight updates. Initial experiments on the MNIST dataset showed that models trained with this combination of losses (termed "R models") converged faster (averaging 15.5 hours versus 20 hours) and produced a more meaningful representation of classification criteria while maintaining accuracy comparable to traditional models. The network learns to identify desired shapes more effectively, suggesting that integrating interpretability into the training process enhances model understanding. While accuracy remained

high for simple datasets, preliminary results on Fashion MNIST (FMNIST) necessitated more refined methodologies. The study "Youve Got the Wrong Outfit" extends these findings to complex architectures like ConvNeXT. This research reiterates that Trustworthy AI (TAI) requires a multidisciplinary approach, introducing the concept of Plausible AI, which defines interpretability alongside expressiveness, learning, and generalisation. Applying the contrastive loss methodology to FMNIST using a ConvNeXT architecture initially presented challenges related to early stopping and optimization. However, a dynamic training strategy utilising adaptive optimisers (AdamW) and schedulers allowed the R model to match the accuracy of standard models (approximately 89%) while offering superior sign identification. The reconstruction of target patterns appeared solid, particularly when the reconstructed sign size approached the original. The study concludes that while feature maps from pre-trained networks (ImageNet1k) are theoretically complete, they may be suboptimal for specific tasks without adequate fine-tuning, underscoring the need for this novel training methodology to develop intrinsically interpretable models.

### 1.3.4 Broad Scientific Challenges and Opportunities

Research in this domain extends the principles of reliable AI to broader scientific contexts, particularly agriculture and citizen science. The position paper "Towards the actual deployment of robust, adaptable, and maintainable AI models for sustainable agriculture" examines the hurdles in implementing AI for precision agriculture, noting that adaptability to new domains is often costly and complex. Identified challenges include the limited availability of high-quality datasets, the dynamic nature of agricultural environments (seasonality, pests), and the unpredictability introduced by climate change. Furthermore, there is a critical need to translate AI predictions into actionable knowledge for farmers. To address these, the research proposes three primary levers:

1. Integration with DSS: AI output must be accompanied by practical explanations to optimise treatments. DSS should intervene both before and after model selection to provide recommendations.
2. Improved trade-off between performance and generalisation: This is achieved via adaptive ensembling and continuous learning. Combining weak DL models at the deep feature level produces strong classifiers that facilitate domain transfer, while continuous learning allows models to evolve without catastrophic forgetting.
3. Strategic approach to crowd-sensing data: New methodologies are proposed to filter uncertainty in non-authoritative sources, utilising deep learning-based image quality assessment methods (IQAM) to verify image relevance.

The paper "Plant-traits: how citizen science and artificial intelligence can impact natural science" investigates the specific utility of citizen science (CS) datasets. Evaluating performance across repositories like PlantTraits2021 and PlantTraits2023, the study explores multimodal data analysis combining RGB images with bioclimatic and tabular data. Results indicate that while CS datasets generally possess high

reliability, domain adaptation struggles with task shift; features learned for disease classification are not directly compatible with plant trait identification. However, the research highlights that adaptive ensembling and feature fusion are promising strategies for improving robustness in these complex, heterogeneous environments.

## 1.4 Publications

In this section, the various papers I produced for my thesis work on reliable US pattern recognition, as well as other work related to the collaboration at my research lab, SI-Lab at CNR-ISTI, are listed. The papers are separated into journal, conference, pre-print and miscellaneous. The main items outside the thesis scope are the works related to the conference AGRI-AI, where I could investigate the application of a multimodal model in biological/agronomic settings. At the same time, other papers focus mainly on Dance Movement Therapy (DMT) through pose reconstruction and complexity algorithms.

### 1.4.1 Journals

- 2024

- Ignesti, G., Martinelli, M., & Moroni, D. (2024). Youve got the wrong number: evaluating deep learning training paradigms using handwritten digit recognition data. *Pattern Recognition and Image Analysis*, 34(4), 978-983.

To build more accurate and trustworthy artificial intelligence algorithms in deep learning, it is essential to understand the mechanisms driving classification systems to identify their targets. Typically, post hoc methods provide insights into this process. In this preliminary work, we shift the reconstruction of the class activation map to the training phase to evaluate how the model's performance compares with standard classification approaches. The MNIST dataset and its variants, such as Fashion MNIST, consist of well-defined images that facilitate testing this type of training process. Specifically, the classification targets are the only significant content in the images, excluding the background, enabling a direct comparison of the reconstruction with the input images. To enhance network guidance, we introduce a contrastive loss term to complement the standard classification function, which often relies on categorical cross-entropy. By comparing the accuracy and the extracted patterns of the standard approach with those of the proposed method, we can gain valuable insights into the network's learning process. This approach aims to improve the model's interpretability and effectiveness during training, ultimately leading to higher classification accuracy and reliability.

- 2023

- Bruno, A., Ignesti, G., Salvetti, O., Moroni, D., & Martinelli, M. (2023). Efficient lung ultrasound classification. *Bioengineering*, 10(5), 555.

A machine learning method for classifying lung ultrasound is proposed here to provide a point of care tool for supporting a safe, fast, and accurate diagnosis that can also be useful during a pandemic such as SARS-CoV-2. Given the advantages (e.g., safety, speed, portability, cost-effectiveness) provided by the ultrasound technology over other examinations (e.g., X-ray, computer tomography, magnetic resonance imaging), our method was validated on the largest public lung ultrasound dataset. Focusing on both accuracy and efficiency, our solution is based on an efficient adaptive ensembling of two EfficientNet-b0 models reaching 100% of accuracy, which, to our knowledge, outperforms the previous state-of-the-art models by at least 5%. The complexity is restrained by adopting specific design choices: ensembling with an adaptive combination layer, ensembling performed on the deep features, and minimal ensemble using two weak models only. In this way, the number of parameters has the same order of magnitude of a single EfficientNet-b0 and the computational cost (FLOPs) is reduced at least by 20%, doubled by parallelization. Moreover, a visual analysis of the saliency maps on sample images of all the classes of the dataset reveals where an inaccurate weak model focuses its attention versus an accurate one.

#### 1.4.2 Conference Papers

- 2024

- Ignesti, G., Martinelli, M., & Moroni, D. (2024, December). Youve Got the Wrong Outfit: Evaluating Deep Learning Paradigms on Digit and Fashion Recognition. In International Conference on Pattern Recognition (pp. 202-215). Cham: Springer Nature Switzerland.

To build more accurate and trustworthy artificial intelligence algorithms in deep learning, it is essential to understand the mechanisms driving classification systems to identify their targets. Typically, post hoc methods are used to provide insights into this process. By contrast, in this work, we investigate the possibility of using class activation maps in combination with contrastive loss to enhance the reliability of training of a deep learning model. MNIST and Fashion MNIST datasets are considered in our investigation since they have already proven a practical starting point for assessing an almost tautologic training strategy for deep learning algorithms, given that classification targets are the primary significant content of the images in these datasets. Starting from the raw comparison of accuracy and system complexity of the proposed approach, a further investigation of the techniques feasibility in a deep learning study is conducted over six random seed splits of the training data and model performance. A modern deep learning network, such as ConvNeXT, determines whether a more robust architecture trained with the proposed mechanics provides better insights than a simple convolutional neural network. This investigation also addresses the importance of skip connections, structured learning layers, and feature map dimensions in the learning process.

- Daoudagh, S., Ignesti, G., Moroni, D., Sebastiani, L., & Paradisi, P. (2024, November). Assessment of Dance Movement Therapy Outcomes: A Preliminary Proposal. In *International Conference on Computer-Human Interaction Research and Applications* (pp. 382-395). Cham: Springer Nature Switzerland.

Context: Dance Movement Therapy (DMT) is a therapeutic modality that utilizes movement to promote holistic well-being. Current DMT assessment methods rely heavily on the subjective judgment of the DMT professional. Objective: Our research aims to develop a framework composed of different components with specific functionalities that can be integrated with the DMT modality to improve the objectivity and efficiency of DMT evaluations. Method: The DMT framework consists of an experimental protocol for data collection and a reference-supporting architecture that includes components for video analysis, physiological signal management, and evaluation tools. Artificial Intelligence (AI) based human pose estimation techniques are also employed to derive the DMT participants poses during the DMT sessions for more reliable movement analysis. Results: Our preliminary results consist of demonstrating the effectiveness of the AI-based pose estimation tool, YOLO-NAS-Pose, in accurately detecting participants poses. Conclusion: The proposed framework offers a promising approach to improving DMT practices by integrating and leveraging AI-based human pose estimation to evaluate participants movement in the DMT setting objectively. Future research will focus on refining the framework and developing user-friendly tools for widespread adoption in real DMT contexts.

- Ignesti, G., Moroni, D., & Martinelli, M. (2024, September). Plant-traits: how citizen science and artificial intelligence can impact natural science. In *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)* (pp. 625-630). IEEE.

Citizen science has emerged as a valuable resource for scientific research, providing large volumes of data for training deep learning models. However, the quality and accuracy of crowd-sourced data pose significant challenges for supervised learning tasks such as plant trait detection. This study investigates the application of AI techniques to address these issues within natural science. We explore the potential of multi-modal data analysis and ensemble methods to improve the accuracy of plant trait classification using citizen science data. Additionally, we examine the effectiveness of transfer learning from authoritative datasets like PlantVillage to enhance model performance on open-access platforms such as iNaturalist. By analysing the strengths and limitations of AI-driven approaches in this context, we aim to contribute to developing robust and reliable methods for utilising citizen science data in natural science.

- Ignesti, G., Moroni, D., & Martinelli, M. (2024). Towards the actual deployment of robust, adaptable, and maintainable AI models for sustainable agriculture. *Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 33.

In the past two decades, computer vision and artificial intelligence (AI) have made significant strides in delivering practical solutions to aid farmers directly in the fields, thereby contributing to the integration of advanced technology in precision agriculture. However, extending these methods to diverse crops and broader applications, including low-resource situations, raises several concerns. Indeed, the adaptability of AI methods to new cases and domains is not always straightforward. Moreover, the dynamic global panorama requires a continuous adaptation and refinement of artificial intelligence models. In this position paper, we examine the current opportunities and challenges, and propose a new approach to address these issues, currently in the implementation phase at CNR-ISTI.

- 2023

- Ignesti, G., Deri, C., D'Angelo, G., Pratali, L., Bruno, A., Benassi, A., Salvetti, O., Moroni, D., & Martinelli, M. (2023, November). Deep learning methods for point-of-care ultrasound examination. In 2023 17th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (pp. 435-440). IEEE.

Point-of-care Test (POCT) is the delivery of medical care at or near the patients bedside. Primarily employed in emergencies, where rapid diagnosis and treatment are critical, POCT is now being used in domestic telehealth solutions, as in the TiAssisto project, thanks to technological advances such as the development of portable and affordable devices, high-speed Internet connections, video conferencing, and Artificial Intelligence (AI). Ultrasound (US) images of internal organs and structures are valuable tools in POCT medicine since this examination is portable, quick, and cost-effective. USs can help diagnose different conditions, including heart problems, abdominal pain, and pneumonia. Deep learning algorithms have proven to be highly effective in image recognition, enabling physicians to make informed decisions on-site. This article presents a pipeline approach providing remarkable and reliable results to handle point-of-care ultrasound examinations, making use of methods for: a) automating text cleaning for privacy based on an Optical Character Recognition (OCR) algorithm; b) scrolling through the video frames and annotating them using an ad hoc implemented tool; c) classifying various signs in US using a state of the art deep learning algorithm, that is an adaptive efficient method ensembling two EfficientNet-b0 weak models; d) benchmarking medical plausibility to address transparency and human in the loop setting using a post hoc explanation visual explanation method, i.e. Grad-CAM.

- 2022

- Ignesti, G., Bruno, A., Deri, C., D'Angelo, G., Bastiani, L., Pratali, L., ... & Martinelli, M. (2022, October). An intelligent platform of services based on multimedia understanding and telehealth for supporting the management of SARS-CoV-2 multi-pathological patients. In 2022 16th International Conference on Signal-Image Technology & Internet-Based Systems

(SITIS) (pp. 553-560). IEEE.

The combination of pervasive sensing and multimedia understanding with the advances in communications, makes it possible to conceive platforms of services for providing telehealth solutions responding to the current needs of society. The recent outbreak has indeed posed several concerns on the management of patients at home, urging to devise complex pathways to address the Severe Acute Respiratory Syndrome (SARS) in combination with the usual diseases of an increasingly elder population. In this paper, we present TiAssisto, a project aiming to design, develop, and validate an innovative and intelligent platform of services, having as its main objective to assist both Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) multi-pathological patients and healthcare professionals. This is achieved by researching and validating new methods to improve their lives and reduce avoidable hospitalisations. TiAssisto features telehealth and telemedicine solutions to enable high-quality standards treatments based on Information and Communication Technologies (ICT), Artificial Intelligence (AI) and Machine Learning (ML). Three hundred patients are involved in our study: one half using our telehealth platform, while the other half participates as a control group for a correct validation. The developed AI models and the Decision Support System assist General Practitioners (GPs) and other healthcare professionals in order to help them in their diagnosis, by providing suggestions and pointing out possible presence or absence of signs that can be related to pathologies. Deep learning techniques are also used to detect the absence or presence of specific signs in lung ultrasound images.

### 1.4.3 Preprints

- 2025

- Ignesti, G., Moroni, D., & Martinelli, M. (2025). Reliable and Trustworthy Learning Prototype: Insight from POCUS.

published in: TBD IEEE ISCAS 2026 publication .

Deep learning models often lack the interpretability and trustworthiness required for clinical use. This paper proposes a prototype-regularised training method to analyse 1,208 lung ultrasound images, with a focus on B-line artefacts. A ConvNeXt-Tiny architecture is used, adding a novel reconstruction loss to the standard classification loss. The model is trained to extract meaningful prototypes and to use them to classify ultrasound images. To prevent these constraints from hindering generalisation, it is used in pairs with the proposed reconstruction loss, a set of plausible data augmentation of the ideal researched prototypes, and a geometry-aware network, a spatial transformer network, to measure which solutions help the network towards outputting the most reliable outcomes. The resulting models are precise, lightweight and interpretable, indicating that the proposed solution can be embedded in an ultrasound device to assist healthcare specialists in point-of-care applications.

- Ignesti, G., Moroni, D., & Martinelli, M. (2025). Batch-CAM: Introduction to better reasoning in convolutional deep learning models. preprint arXiv:2510.00664.

Understanding the inner workings of deep learning models is crucial for advancing artificial intelligence, particularly in high-stakes fields such as healthcare, where accurate explanations are as vital as precision. This paper introduces Batch-CAM, a novel training paradigm that fuses a batch implementation of the Grad-CAM algorithm with a prototypical reconstruction loss. This combination guides the model to focus on salient image features, thereby enhancing its performance across classification tasks. Our results demonstrate that Batch-CAM achieves a simultaneous improvement in accuracy and image reconstruction quality while reducing training and inference times. By ensuring models learn from evidence-relevant information, this approach makes a relevant contribution to building more transparent, explainable, and trustworthy AI systems.

- Sebastiani, L., et al. (2025). The Impact of Dance Movement Therapy on Subjective Well-Being and Quality of Life in Healthy Older Adults. bioRxiv.

Undertaking physical and social activities such as dance and music have positive effects on the mental well-being, cognitive and physical abilities of older people. Dance Movement Therapy (DMT) is a complementary practice that uses dance and movement to stimulate creativity, enabling individuals to engage in positive interpersonal connections within a group. Therapeutic approaches involving DMT have been proven to promote positive mental health outcomes in various patient groups. This study aimed to evaluate the possible efficacy of DMT on the subjective perceptions of quality of life, satisfaction with life, and personal well-being of a sample of healthy adults aged over 65. Participants (n=25; females=17) took part in an 8-week DMT cycle on a once-weekly basis. The DMT cycle was preceded and followed by a one-hour session during which participants filled out the World Health Organisation Quality of Life assessment (WHOQoL-BRIEF), the Satisfaction With LifeScale (SWLS), and the Wellbeing Index (PWI) scale. Comparison of pre- and post-DMT scores by repeated measures ANalysis Of VAriance (rmANOVA) showed a positive effect of DMT, especially in terms of overall quality of life and social relationships. Participants in the over-70s age group who showed significantly lower scores than their younger counterparts in terms of overall quality of life, Social subscale, and PWI, exhibited analogous DMT-related changes. Our findings suggest that DMT has the potential to foster a shared positive state of mind among older individuals, thereby facilitating social interaction not only among the elderly population but also among its most senior members.

#### 1.4.4 Magazines and Newsletters

- 2023

- Bruno, A., Ignesti, G., & Martinelli, M. (2023). Explaining Ensemble Mod-



els for Lung Ultrasound Classification. ERCIM News, 134, 17.

Correct classification is the main aspect in evaluating the quality of an artificial intelligence system, but what happens when you reach top accuracy and no method explains how it works? In our study, we aim at addressing the black-box problem using an ad-hoc built classifier for lung ultrasound images.

## 1.5 Outline of the Work

This introduction delineates the primary topics addressed throughout the remainder of this dissertation. Beginning with Chapter 2, we introduce research aligning with my initial thesis studies [53] regarding the application of AI in healthcare, highlighting the potential of medical decision support systems. As a core contribution of this PhD research, the work specifically focuses on classification methods for Ultrasound imaging analysis, examining the complete pipeline from image acquisition to clinical reporting. These investigations are detailed in Chapter 3, which revisits key contributions from my academic research [14, 15, 54]. Having established a framework for medical imaging, our research interest naturally evolved toward the underlying theoretical reliability of these systems. Therefore, Chapter 4 steps away from direct clinical application to investigate the abstract algorithmic foundations necessary for building robust AI. Chapter 5 presents additional studies [23, 57, 58] utilising specific datasets to enable multimedia feature fusion experiments, alongside investigations into critical aspects required for the real-world deployment of DL models. Finally, Chapter 6 summarises the collected results and proposes future research directions directly correlated with the material presented herein.



# Part I

## AI for medicine



# Chapter 2

## Telemedicine and the need for AI

The material presented in this chapter is based on the conference paper G. Ignesti, A. Bruno, C. Deri, G. D'Angelo, L. Bastiani, L. Pratali, S. Memmini, D. Cicalini, A. Dini, G. Galesi, F. Pardini, M. Tampucci, A. Benassi, O. Salvetti, D. Moroni, and M. Martinelli. An intelligent platform of services based on multimedia understanding and telehealth for supporting the management of sars-cov-2 multi-pathological patients. In 2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pages 553–560. IEEE, 2022.

The SARS-CoV-2 pandemic served as a global stress test for healthcare infrastructures, revealing latent systemic vulnerabilities while also accelerating the adoption of novel technological paradigms [119]. The international crisis necessitated a rapid re-evaluation of patient care models, pushing distributed technology-enabled solutions from the periphery to the forefront of medical practice [34].

This chapter explores the convergence of telemedicine, point-of-care diagnostics such as LUS, and AI, as a necessary and powerful response to the challenges elucidated by the pandemic.

The initial phase of the outbreak highlighted a critical failure of non-resilient systems, exemplified by the Italian national healthcare system's lack of method and structure to address the rapid spread of contagious pathogens. The drastic measures required to manage infected patients created a secondary crisis, as multi-pathological patients were deprived of conventional care. This deferral of essential services represents a systemic breakdown, underscoring the primary motivation for research into intelligent telehealth platforms that aim to build resilience and continuity into the healthcare ecosystem. This chapter establishes the scientific and clinical rationale for investigating the potential pairing of such a platform with AI models, using the TiAssisto project as a foundational case study.

### 2.1 The Strategic Rise of Telehealth

Fundamental advancements in communication infrastructure have strategically underpinned the evolution of telehealth from a niche concept to a systemic necessity [139]. The advent and proliferation of high-speed, stable internet access via 4G, 5G, and fibre networks (FTTC/FTTH) transformed remote patient care from a

theoretical possibility into a practical, scalable solution. This technological maturation provided the bedrock for new models of healthcare delivery to be built, tested, and deployed, particularly during a crisis that demanded physical distancing. This technological shift yielded two profound and interconnected benefits. First, it empowered entirely new modes of interaction between patients and their General Practitioners (GPs), effectively overcoming the physical barriers imposed by quarantine and enabling continuity of care. Second, and of critical long-term importance, this enhanced connectivity began to generate a constant stream of newly available data. This influx of information created a big-data environment essential for modern medical analysis and for tuning complex AI models. While this new infrastructure enabled remote care, the flood of unstructured temporal and imaging data created an analytical bottleneck. Integrating advanced diagnostic and AI-driven tools thus became not just an enhancement, but an operational necessity. A telemedicine platform, by itself, is a tool for gathering and accessing more data, though the quantity gathered can vary across subjects. If we assume that a general practitioner can follow up with 1000 subjects, this could lead to a significant workload. It was clear that the tool for analysing this data was not just a commodity but a necessity [95].

## 2.2 TiAssisto

The TiAssisto project, funded by the Tuscany Region, is an intelligent telemedicine and teleconsultation platform developed collaboratively by the Italian National Research Council, CNR, and the Local Health Authority of Northwest Tuscany, ATNO, to address the fragilities in the healthcare system exposed by the COVID-19 pandemic, focusing on multi-pathological subjects. The system provides continuous remote monitoring through a secure, self-hosted Web platform featuring personalised interfaces designed for patients and for physicians [53]. Its core innovation is an AI-driven decision support system designed to enable proactive clinical interventions. By integrating DL for automated LUS analysis with a novel hybrid health score, a composite risk indicator that fuses the reliability of rule-based classical analysis with the predictive power of data-driven AI, the system empowers clinicians to anticipate patient needs rather than merely react to them. The platform's effectiveness in improving quality of life and reducing hospitalisations is being scientifically validated through a 300-patient Randomised Controlled Trial (RCT), with the ultimate goal of building trust in AI and advancing a more resilient, personalised model of 5.0 medicine [142].

## 2.3 Applying Artificial Intelligence to Temporal and Imaging Data

A central challenge in modern medicine has been the difficulty of synthesising disparate and continuous streams of patient data into a cohesive, longitudinal clinical picture. In the traditional approach, healthcare for generally healthy individuals (or 'non-pathological subjects') is fundamentally episodic. This means their health management or preventive care is not continuously monitored. In contrast, telehealth

platforms can generate a continuous flow of temporal data (vital signs) and complex imaging data (i.e. ultrasound scans). Before the development of AI and other advanced analytics, it would have been overwhelming and practically impossible for a single physician to monitor and interpret such continuous self-monitoring data streams. Lacking the necessary tools, a physician couldn't sift through that sheer volume of data to identify meaningful patterns or assess a patient's health status in real-time. A widespread adoption of telemedicine solutions, however, generates its own challenge: a high-velocity, high-volume stream of complex imaging data. The very success of LUS in data acquisition created an urgent analytical bottleneck, highlighting the need for computational systems capable of interpreting this data at scale and with a level of consistency beyond human capacity, Fig. 2.1. AI and ML

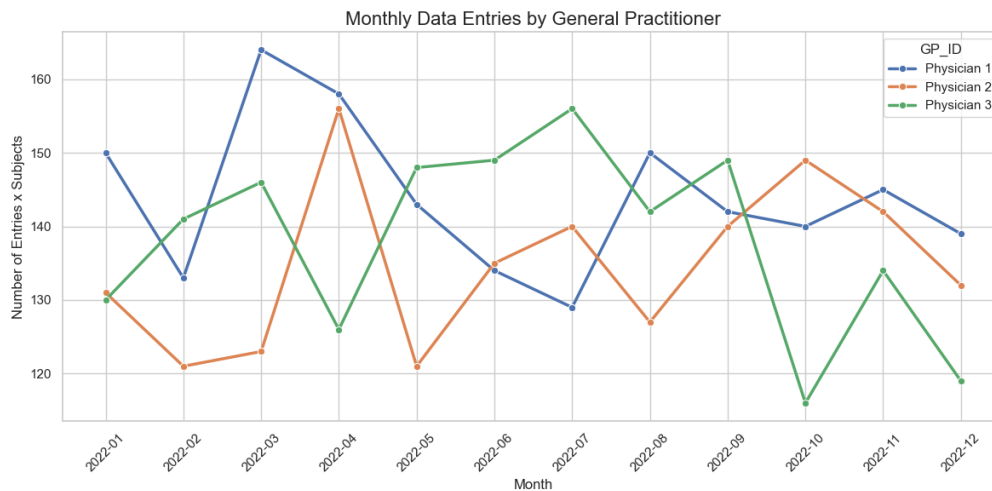


Figure 2.1: This infographic illustrates the amount of data gathered per subject during the TiAssisto trial period, specifically detailing the reports for three of the main GPs enrolled in the project.

provide the critical capabilities to overcome this challenge. As implemented within the TiAssisto project, these technologies offer a robust mechanism for synthesising multimodal data into clinically relevant insights. AI is applied to:

- **Synthesise Temporal Data:** Employ a rule-based AI system to transform the continuous stream of vital measurements into an actionable patient health trend score, providing physicians with a rapid assessment of a patient's trajectory;
- **Analyse Imaging Data:** Automatically process LUS (imaging) to detect and identify typical patterns associated with various pathologies, providing powerful decision support to physicians during diagnosis.

This integration of temporal and imaging data using AI represents more than an incremental advance; it catalyses a fundamental restructuring of medical practice. The ultimate objective is to harness this information for personalised healthcare, underpinning the evolution towards Medicine 5.0. By transcending episodic data collection in favour of continuous algorithmic synthesis, this approach enables a definitive clinical shift from reactive intervention to proactive, personalised health

monitoring and prevention. Such a system facilitates the construction of dynamic digital phenotypes, where patient health is understood as a continuous trajectory rather than a series of isolated snapshots. Through the longitudinal analysis of temporal and imaging biomarkers, the platform can identify subtle physiological deviations that precede symptomatic manifestation. This capability shifts the locus of care from the hospital to the home, empowering patients to participate actively in their health management while providing clinicians with a granular, real-time understanding of disease progression.

## 2.4 Ultrasound diagnostic efficiency

The shift toward decentralised healthcare models places immense strategic value on portable, powerful diagnostic tools. LUS exemplifies this, emerging as a pivotal technology during the SARS-CoV-2 crisis [135]. It offered a potent diagnostic capability that could be deployed far beyond traditional hospital confines, proving indispensable for managing patients in challenging new environments, such as home quarantine. The key attributes that established ultrasound as an essential point-of-care tool during the pandemic are presented in the following Table 2.1, along with the limitations of the exams.

Table 2.1: Pros and Cons of Lung Ultrasound for SARS-CoV-2 Assessment

| Feature            | Pros (Advantages)                                                                       | Cons (Disadvantages)                                                                               |
|--------------------|-----------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------|
| Versatility & Cost | Cost-effective and versatile; widely deployable across various settings.                | Highly operator-dependent; requires specialised training for acquisition and interpretation.       |
| Portability        | Portable technology enables assessment of home-quarantined patients.                    | Close proximity required, increasing operator infection risk compared to remote CT.                |
| Diagnostic Power   | Effective in identifying structural signs of pneumonia (e.g., B-lines, consolidations). | Limited to peripheral lesions; cannot visualise deep or central pathology due to air interference. |

## 2.5 The TiAssisto Project

The TiAssisto project is now a validated research framework. It serves as a comprehensive case study that embodies the convergence of the critical themes discussed throughout this chapter. It was conceived in direct response to a systemic healthcare crisis, leveraging telehealth infrastructure and POCUS to address the acute needs of a vulnerable patient population. Crucially, it integrates artificial intelligence as the core engine for data synthesis, transforming raw data streams into decision-support



tools for healthcare professionals. This study, therefore, investigates how the TiAssisto platform serves as a functional model for validating the clinical and operational impact of intelligent telemedicine. The central thesis is that such platforms can reduce avoidable hospitalisations and improve systemic resilience through AI-driven, proactive management of multi-pathological patients. The next chapters will delve into the specific research protocol and methods, the technical architecture of the TiAssisto web platform, and the design and experimental results of its integrated artificial intelligence modules.

## 2.6 Toward Medicine 5.0

The TiAssisto platform exemplifies the transition from Telemedicine 4.0, which focuses primarily on digitisation and remote connectivity, to Medicine 5.0, Fig. 2.2). This paradigm is defined by the harmonious collaboration between human intelligence and artificial cognitive systems. While Industry 4.0 prioritised automation and data exchange, the 5.0 framework re-centres the human element, prioritising patient well-being, personalisation, and the trustworthiness of the technological interaction. In this context, the TiAssisto project works as a critical bridge. By integrating UCD principles, the platform ensures that the technological layer does not obfuscate the clinical relationship but rather enhances it. This aligns with recent discoveries in AI-driven healthcare, where the focus has shifted from black-box predictive accuracy to explainable decision support that fosters trust among GPs and patients. The project's emphasis on strictly regulated privacy, using local server encryption rather than third-party cloud solutions, anticipates the growing demand for sovereign AI architectures that guarantee data security amid rising cyber vulnerability. Access to national cloud services for pharmacology management was successfully established for data acquisition and updates. Furthermore, while data integration with physicians' existing clinical systems was achieved, the research team was restricted from accessing these resources.

## 2.7 From Weak to Strong AI

The core innovation of the research lies in its hybrid algorithmic approach, which mirrors the current trajectory of medical AI discoveries. The project distinguishes between two necessary layers of intelligence. For longitudinal monitoring, the platform utilises standardised physiological scores and classical machine learning techniques, such as Gradient Boosting. Initial experiments on synthetic datasets yielded a prediction accuracy of 100%. While this figure is anomalously high due to the nature of the generated data, it demonstrates the potential of classical algorithms to effectively categorise patient risk (Worsening, Stable, Improving) when variables are clearly defined. This layer acts as a rapid triage mechanism, filtering the noise of daily monitoring into actionable alerts, Fig 2.3.

The discovery in this context is the successful application of DL to point-of-care ultrasounds POCUS. The platform employs an ensemble of EfficientNet architectures to automatically classify lung pathologies. This moves beyond simple param-

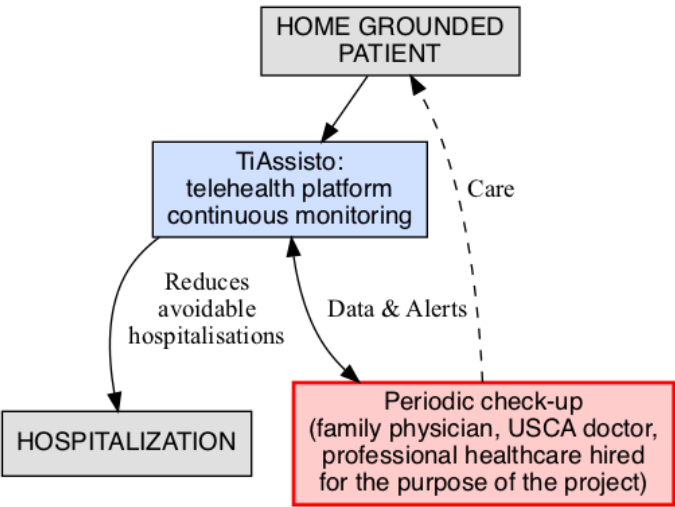


Figure 2.2: In the figure, the new 5.0 telemedicine approach is represented, in which the platform works as an intermediate step between hospital care and care-home assistance.



Figure 2.3: The figure portrays the Italian TiAssisto website, in which the involved physicians can monitor vitals, images, and notes from the related subjects; the system automatically detects anomalies and visually alerts the physicians.

eter tracking to complex pattern recognition, enabling the system to identify signs of pneumonia, pulmonary oedema, or pneumothorax in real time. This integration validates the hypothesis that portable, AI-enhanced diagnostics can decentralise complex care, effectively moving the "hospital" to the patient's home.

## 2.8 Towards a New Medical Interpretation Approach

The SARS-CoV-2 pandemic served as a catalyst, forcing the healthcare system to confront its lack of resilience and structural weaknesses in managing patients with diverse comorbidities. The TiAssisto project responds to this not merely as a temporary emergency measure, but as a foundational blueprint for the future of public health. By combining pervasive sensing, secure teleconsultation, and a dual-layer AI architecture (blending rule-based stability with DL's diagnostic power), the platform successfully operationalises the concept of Medicine 5.0. It demonstrates that the future of healthcare lies not in replacing the clinician, but in augmenting their capacity to process complex, continuous data streams. As validated by the ongoing RCT, such intelligent platforms are essential for reducing avoidable hospitalisations, ensuring that high-quality care is resilient, accessible, and deeply human-centric.

These principles will guide the development of the topics explained in the following sections and chapters, starting from the core principles and DL analysis techniques necessary to develop a fast and accurate lung ultrasound examination protocol to assist physicians. The following section examines the core mathematical ideas underlying the development of reliable AI systems and the tests necessary to develop them



# Chapter 3

## Deep Learning for POCUS

The material presented in this chapter is based on the following publications:

- A. Bruno, G. Ignesti, O. Salvetti, D. Moroni, and M. Martinelli. Efficient lung ultrasound classification. *Bioengineering*, 10(5):555, 2023,
- G. Ignesti, C. Deri, G. D'Angelo, L. Pratali, A. Bruno, A. Benassi, O. Salvetti, D. Moroni, and M. Martinelli. Deep learning methods for point-of-care ultrasound examination. In *2023 17th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 435–440. IEEE, 2023,
- A. Bruno, G. Ignesti, and M. Martinelli. Explaining ensemble models for lung ultrasound classification. *ERCIM News*, 2023(134):16–18, 2023.

The heart of the analysis presented in this and the following chapters is the criteria for developing a plausible DL model specifically for medical imaging. There are four key aspects in assessing and choosing a DL model for medical applications: data preparation, model selection, training phase, and model validation. Each step will introduce a key fundamental strategy in AI development, tailored and in-depth analysed for ultrasound image analysis. Before delving into these topics, it is better to contextualise the nature of ultrasound images, especially those related to lung diagnosis, and to define the term 'Point-of-Care'. This contextualization is necessary because every DL model is inextricably linked to the quality and nature of the data it consumes, as well as the physical principles underlying the tools used to generate that data. Outside the key fundamental model introduced in the Introduction, any specialised model is deeply influenced by the analysis, especially if the aim is to align the AI decision with that of the physicians who should analyse the images. The following section will define ultrasound and then explore the method developed for the Reliable deployment of the model to analyse medical images precisely and safely.

### 3.1 Point-of-Care Ultrasound (POCUS)

Ultrasound is a cornerstone of modern medical imaging, providing a safe, non-invasive, and real-time view of the internal human anatomy. Its transition from

bulky, stationary machines in dedicated radiology suites to portable, point-of-care devices has advanced clinical practice, particularly in emergency medicine, critical care, and resource-limited settings [91]. Understanding the basic physics of how these devices create images is fundamental to appreciating the challenges and opportunities in applying AI for their interpretation.

Ultrasound devices are simpler in terms of components than much larger imaging machines, such as Computed Tomography (CT) scan and Magnetic Resonance Imaging (MRI), but their operation is equally complex. The heart of the machine is the probe, the part that requires contact with the analysed subject. The probe geometry and implementation have evolved from a linear shape to modern designs featuring also curvilinear or spherical shapes, Fig.3.1. The shapes are directly related to the machine's utilisation and facilitate contact with the subject, but it is a component within the probe that actually enables the imaging technique, namely, a transducer inside the probe generates and then receives sound waves. The probe's

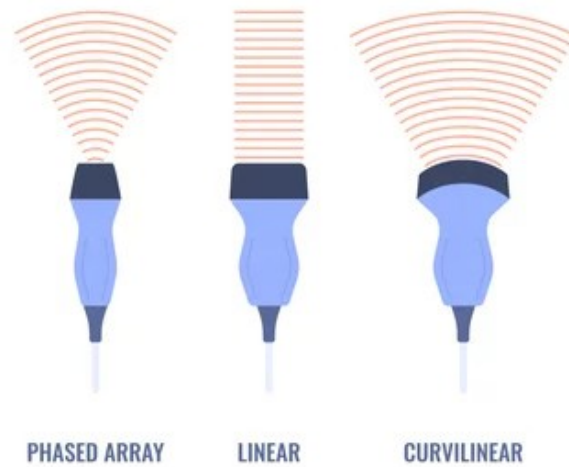


Figure 3.1: Different types of probe shapes for ultrasound devices, each probe corresponds to a type of exam with a designed anatomical target

geometry directly corresponds to the transducer configuration, since it's the relevant piece. In contrast, the external probe case protects the transducer and enables a better match with the human body. Changes in the transducer implementation over time are more complex than just the geometric shape; material, power profiles, and finishing trim have also undergone significant changes: from a single quartz piece with a direct power source to a complex ceramic-based 2D-array element that powers singularly. With the advent of digital equipment, a component known as a beamformer became a staple of ultrasound devices. This part of the equipment powers the array, creating a specific sound wave to achieve superior control over image creation. Beyond that, the ability to create such a power profile enables the beamformer to separate echoes more precisely, not only as a better transmission medium but also as a more powerful tool against noise (which is inherently high in this type of diagnostic procedure). Naturally, modern devices enable deep interaction with image generation through a series of control mechanisms (including implementation in power supplies, transducer shape and materials, and beamformer

technologies). As discussed, the probe head and transducer can have different profiles, each connected to a different emission frequency. The penetration depth of the probe is intrinsically related to its emission frequency; therefore, depending on the anatomical region being explored, different probe heads/frequencies are available. Depth, gain and focus are other key parameters. Each of them controls a key setting of the exam, which analytically determines the field-of-view dimension and the device's frame rate, as shown in Fig. 3.2. More complex parameters, such as Time Gain Compensation (TGC) and Dynamic Range, are essential for image optimisation. While overall gain provides a general increase in signal brightness, TGC enables differential amplification. This enables enhancing specific pixels based on their depth, compensating for the natural attenuation of sound as it travels deeper into tissue. Dynamic Range, on the other hand, is a control tool that defines how echo strengths are mapped to shades of grey; rather than modifying the transmitted soundwave, it adjusts the contrast resolution. It determines the ratio between the strongest and weakest displayed signals, effectively dictating whether the image appears high-contrast (more black and white) or smoother with many shades of grey.

### 3.1.1 The Piezoelectric Effect

The core technologies behind the transducer originate from a remarkable physical phenomenon known as the piezoelectric effect [21], first discovered by physicists Jacques and Pierre Curie in 1880. At the heart of a traditional ultrasound probe, or transducer, is an active element that converts electrical energy to acoustic energy and vice versa. While early transducers used natural piezoelectric crystals, such as quartz ( $\text{SiO}_2$ ), these materials were largely replaced by the 1950s because their electrical output was too low for sophisticated technology. Today, the active element is almost universally an engineered, artificial piezoelectric ceramic, with Lead Zirconate Titanate (PZT) being the most common material of choice, even though the new lead free ceramics are gaining trust [94]. More advanced probes may use specialised piezoelectric single crystals (such as PMN-PT), piezocomposites, or even entirely different non-piezoelectric technologies, such as Capacitive Micromachined Ultrasonic Transducers (CMUTs) [75, 117].

The piezoelectric effect is a dual-turn phenomenon. The converse piezoelectric effect (or electrostrictive effect) occurs when an electric voltage is applied across the material; its molecular structure changes, causing it to deform or vibrate. The direct piezoelectric effect occurs when a material is mechanically strained or vibrated by an external force (such as a returning sound wave), generating an electric charge or voltage. This dual capability enables a single transducer to act as both an ultrasonic emitter and receiver, thereby transmitting and receiving sound waves, Fig. 3.3.

To generate an image, the ultrasound machine sends an electrical pulse to the piezoelectric element. This triggers the converse piezoelectric effect, causing the element to vibrate and emit a beam of high-frequency sound waves (typically in the 2-15 Megahertz range, far above human hearing) into the body's tissues. These sound waves travel through the tissue, encounter boundaries between different structures (e.g., muscle and bone, liver and kidney), and are reflected as echoes. When these returning echoes strike the piezoelectric element, they cause it to vibrate, triggering



Figure 3.2: A modern example of an ultrasound device, its structure resembles that of a modern notebook, each key in the keyboard represents a setting, like Gain, Depth, etc.



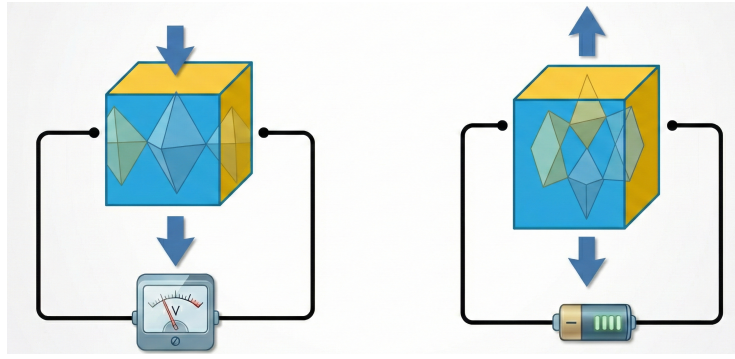


Figure 3.3: Example of direct and indirect piezoelectric effect. The stress material produces a voltage in the indirect effect, while the powered material generates stress in the direct effect, as shown in the figure: left (indirect), right (direct).

the direct piezoelectric effect. This, in turn, generates an electrical signal that is sent back to the machine that converts these echoes into an image.

### 3.1.2 Sound through media (Human Tissues)

Sound is a mechanical disturbance propagating through an elastic medium; in US settings, the human tissues serve as that medium. As stated in chapter 3.1.1, sound is generated via pressure fluctuations to its eventual dissipation. Sound can be described as a wave function dependent on the initial amplitude derived from transducer medium coupling, and the analysis focuses on signal attenuation caused by geometric divergence and the inverse-square law as the wave travels through media. Sound is inextricably linked to the mechanics of the medium through which it travels. At its core, sound is created by modulating air pressure, resulting in longitudinal waves composed of compressions (high pressure) and rarefactions (low pressure). While the perceptual experience of sound is subjective, its physical propagation is governed by rigorous mathematical and physical principles. The propagation of sound is characterised by a wave function, which mathematically describes the displacement of particles or the pressure variations over time and space. In a simplified one-dimensional model, a harmonic sound wave (Eq 3.1) travelling along the  $x$ -axis can be expressed as:

$$p(x, t) = A_0 \sin(kx - \omega t + \phi) \quad (3.1)$$

Where:

- \*  $p(x, t)$  represents the acoustic pressure at position  $x$  and time  $t$ ,
- \*  $k$  is the wave number ( $2\pi/\lambda$ ),
- \*  $\omega$  is the angular frequency,
- \*  $A_0$  is the initial amplitude,
- \*  $\phi$  is the initial phase (the full-turn variable angle).

When sound transitions from air to the physiological environment of human tissue, the mechanics of propagation shift significantly. Unlike air, soft tissues are modelled primarily as fluids where particles oscillate parallel to the direction of wave propagation. In this context, the interaction between the acoustic wave and the tissue is governed by the medium's specific acoustic impedance ( $Z$ ). Acoustic impedance determines how much of a sound wave is transmitted versus reflected at the boundary between two different tissues (after accounting for absorption). It is defined as the product of the tissue's density and the speed of sound within it. This relationship is critical for ultrasound imaging, as image generation relies on the echoes caused by impedance mismatches at tissue interfaces [28]. The acoustic impedance is defined mathematically in the following Eq. 3.2.

$$Z = \rho \cdot c \quad (3.2)$$

Where:

- \*  $Z$  is the acoustic impedance (measured in Rayls, or  $kg/(m^2 \cdot s)$ ),
- \*  $\rho$  is the density of the medium ( $kg/m^3$ ),
- \*  $c$  is the speed of sound in the medium ( $m/s$ ).

In soft human tissue, the average speed of sound is approximated at 1540 m/s. However, this varies with tissue stiffness and density; for instance, sound travels faster through bone due to its high bulk modulus, whereas it travels more slowly through adipose (fat) tissue [32, 110]. The returning echoes sum the reflected sound waves through the media, impact the crystal back, and generate electrical signals as input due to the inverse piezoelectric effect.

### 3.1.3 From Echoes to Beamforming

The ultrasound machine processes these returning electrical signals to construct a two-dimensional image. It does this by interpreting two key pieces of information for each echo:

- **Time of Flight:** The machine measures the time it takes for the sound wave to travel from the transducer, reflect off a tissue boundary, and return. Knowing the speed of sound in human tissue (approximately 1540 m/s), it can calculate the distance to that boundary. Echoes that return quickly correspond to shallow structures, while those that take longer correspond to deeper ones. This determines the vertical position of a pixel on the screen.
- **Signal Strength (Amplitude):** The strength of the returning echo determines the brightness of the corresponding pixel. Tissues that are very dense or highly reflective, like bone or air, reflect a lot of sound and appear bright white (hyperechoic). Less-dense tissues, such as fluid-filled cysts, reflect little sound and appear black (anechoic). Most soft tissues fall somewhere in between, creating various shades of grey.

Using a technique called beamforming, these two parameters are combined to create a new signal, which is then processed to generate an image of a cross-section of the probed anatomy. Beamforming functions as the requisite spatial filtering architecture, acting upon the raw acoustic data to synthesise a coherent image by rigorously manipulating the interdependence of time of flight and signal amplitude. Because acoustic wavefronts backscattered from a specific tissue location arrive at individual transducer elements at varying time instances due to geometric path differences, the beamformer applies calculated time delays to each receive channel to compensate for these variances. This process, known as dynamic focusing, realigns the signals so that the time of flight parameter accurately correlates to a precise focal point rather than a generalized depth. Once these temporal offsets are corrected, the beamformer sums the phase-aligned signals, Fig. 3.4, inducing constructive interference to maximise the amplitude of echoes originating from the focal region while simultaneously suppressing off-axis noise through incoherent summation. Consequently, the final pixel brightness a direct mapping of this processed amplitude reflects a spatially isolated interrogation of tissue reflectivity at the exact coordinates defined by the corrected time of flight, thereby determining the system's lateral resolution and contrast fidelity.

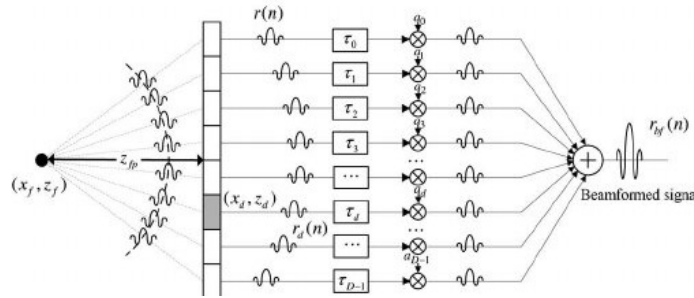


Figure 3.4: Representation of beamforming, the reflected soundwave from a generic focal point impacts the transducer back, transducers that, in function of the position, apply a series of corrections to correctly sum the multiple parts of the soundwave into a cleaned signal.

In a generalised framework encompassing the evolution of transducer technologies, modern beamforming algorithms improve the simple delay-and-sum architectures by integrating specific array geometries and adaptive control settings to fundamentally alter the mapping of time-of-flight and amplitude to pixel luminosity. Whether operating on linear, curvilinear, or multidimensional matrix arrays, the beamformer algorithmically compensates for the physical arrangement of elements, utilising the specific transducer geometry to calculate complex and delay profiles that ensure the time of flight corresponds to a precise spatial location within a dynamically steering field of view.

Concurrently, control parameters such as dynamic aperture and apodization<sup>1</sup> directly manipulate the amplitude component by applying variable weighting functions

<sup>1</sup>Apodization is a spatially variant weighting filter applied across the active probe aperture, wherein the amplitude of signals from peripheral elements is mathematically attenuated relative to the centre to suppress off-axis side lobes.

across the active elements; this process selectively suppresses off-axis energy known as side lobes and enhances the primary lobe signal, effectively filtering the raw amplitude data before summation. Consequently, the final pixel brightness is not merely a measure of echo intensity, but a synthesised value where the time of flight ensures rigorous spatial registration across evolving geometries, and the amplitude is a contrast-optimised representation of tissue reflectivity resulting from sophisticated wavefront manipulation.

### 3.1.4 Beamforming to images

At this stage, the ultrasound machine has combined signals from multiple transducer elements to create a focused line of data. This beamformed signal contains raw echo information; it looks nothing like an image, yet it is a high-frequency oscillating wave. The beamformer outputs a series of scan lines from a focal point or zone. A single scan line,  $s(t)$ , represents the echoes received from a specific direction over time (which correlates to depth). That signal is composed of a high frequency  $f_c$  that describes the oscillation of the transducer and an amplitude  $A(t)$  that is, by design, the cleaned strength of the reflection over that line, in function of phase shift over time  $\phi(t)$  represented in the following Eq. 3.3.

$$s(t) = A(t) \cdot \cos(2\pi f_c t + \phi(t)) \quad (3.3)$$

Due to the high frequency of the oscillation (3-12 MHz), the signal would be too fast for the eye to see and contain no anatomical information. The next step is to extract the signal's envelope using the Hilbert transform, which produces the signal's integral over frequency, yielding a real-valued representation. The transform shifts the signal 3.4 phase by  $(\frac{\pi}{2})$ .

$$\hat{s}(t) = \mathcal{H}[s(t)] = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{s(\tau)}{t - \tau} d\tau \quad (3.4)$$

We create a complex signal whose real part is the original signal and whose imaginary part is the Hilbert transform.

$$s_a(t) = s(t) + j\hat{s}(t) \quad (3.5)$$

The desired envelope  $E(t)$  is the magnitude of this complex vector.

$$E(t) = |s_a(t)| = \sqrt{s(t)^2 + \hat{s}(t)^2} \quad (3.6)$$

To provide an intuitive understanding of Envelope Detection, consider a filled-in sine wave: the detection process draws a boundary that connects the wave's peaks, effectively removing its rapid internal oscillations. To fit the massive dynamic range of echoes onto a screen that humans can interpret, logarithmic compression is applied, as shown in Fig. 3.5. Ultrasound echoes vary widely in intensity due to differences in the reflectivity of tissues. The difference can be 100 dB (a ratio of 100,000 : 1). A standard monitor only displays 256 shades of grey (8-bit). If we mapped this linearly, the weak tissues (where the diagnostic info is) would be completely black,

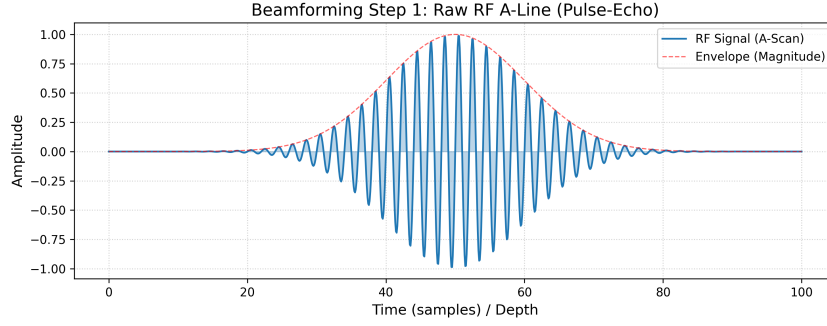


Figure 3.5: Visualisation of the envelope detection process. The oscillating raw RF signal (A-Scan) is demodulated to obtain the signal magnitude, which serves as the basis for the displayable image.

and the bone would be white. Therefore, the logarithmic transformation, Eq 3.7, enhances weak signals while preventing strong signals from saturating.<sup>2</sup>

$$D(t) = \alpha \cdot \log_{10}(\beta \cdot E(t) + 1) \quad (3.7)$$

Where:

- \*  $D(t)$  is the display-ready signal.
- \*  $E(t)$  is the envelope from the previous step.
- \*  $\alpha$  is a scaling factor (gain).
- \*  $\beta$  controls the curvature of the compression (dynamic range).

The last step in forming the image is to transform the stored polar coordinates of the wave into Cartesian coordinates that are representable and interpretable. A set of vectors  $(r, \theta)$ , where  $r$  is depth (time) and  $\theta$  is the angle of the beam, must be represented in a rectangular grid of pixels  $(x, y)$ . If we plot the lines side by side, a sector scan would appear as a rectangle, distorting the anatomy. We must warp the image into the correct fan shape, Eq. 3.8-3.10. We map every pixel  $(x, y)$  on the screen back to the acoustic data  $(r, \theta)$ .

- Geometric Mapping:

$$r = \sqrt{x^2 + y^2} \quad (3.8)$$

$$\theta = \arctan\left(\frac{x}{y}\right) \quad (3.9)$$

- Interpolation: Because the calculated  $(r, \theta)$  usually lands between existing data points (especially deeper in the image where scan lines spread apart), a bilinear Interpolation is commonly used to avoid this problem. It computes the weighted average of the four nearest acoustic samples to determine the screen pixel's value.

$$P(x, y) \approx w_1 S(r_1, \theta_1) + w_2 S(r_2, \theta_1) + w_3 S(r_1, \theta_2) + w_4 S(r_2, \theta_2) \quad (3.10)$$

---

<sup>2</sup>This step is crucial for tissue texture. Without it, organs would look like solid black shapes with bright white outlines.

Finally, the processed data is mapped to a grayscale image, where pixel correlations can be interpreted using the Table in 3.1.

Table 3.1: Mapping between pixel colour space, beamforming signal and anatomical interpretation [76]

| Pixel Value     | Reference   | Interpretation          |
|-----------------|-------------|-------------------------|
| 0-25 (Black)    | No echo     | Fluid, cyst             |
| 26-200 (Grey)   | Echo        | Soft tissue texture     |
| 201-255 (White) | Strong echo | Bones, stones, artefact |

Prior to final visualisation, the raw image data undergoes a suite of post-processing algorithms designed to optimise diagnostic clarity by mitigating artefacts and sharpening structural definitions. A primary focus of this stage is the attenuation of speckle noise a pervasive, granular interference inherent to coherent imaging systems through the application of advanced filtering techniques such as anisotropic diffusion [98] or wavelet-based de-noising. These methods are mathematically sophisticated, distinguishing stochastic noise from actual anatomical structure to smooth homogeneous tissue regions while rigorously preserving the integrity of structural edges, Fig. 3.6.

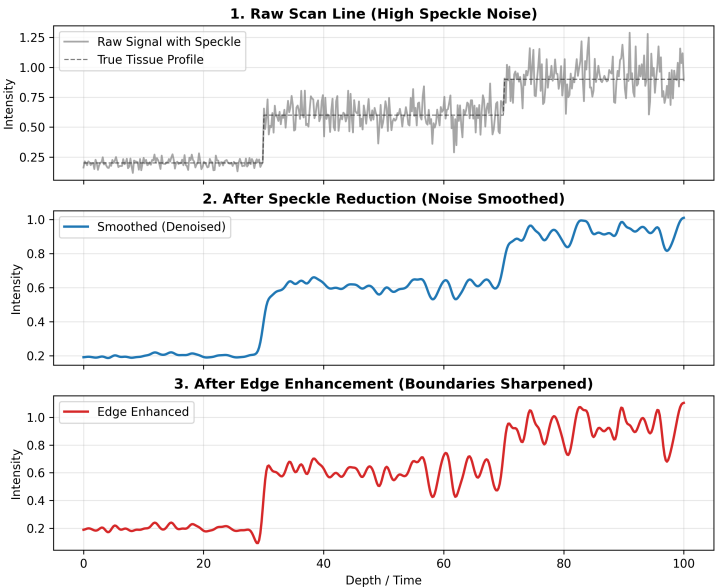


Figure 3.6: Conceptual signal smoothness steps to reduce possible noises

Complementing this noise reduction, edge enhancement protocols are frequently employed, utilizing high-pass filtering mechanisms to amplify high-frequency components of the signal; this process sharpens the contrast at the boundaries between distinct organs or tissue types, effectively delineating anatomical borders that might otherwise be obscured by the inherent fuzziness of the ultrasound modality, ulti-

mately yielding a refined image with improved signal-to-noise ratio and structural visibility.

### 3.1.5 Challenges and Artefacts

Ultrasound imaging is not without its challenges. The images are often noisy and heavily dependent on operator skill. The choice of transducer and machine settings significantly impacts image quality. Modern sonography devices come with different probes; each probe can be accessorised with a transducer of a particular geometry, Fig. 3.7. Initially, transducers were refined to have either linear or convex geometry. The first type of geometry enables the analysis of shallow areas, such as in vascular studies and breast assessment. This modality offers greater detail resolution, but because of its high frequency, it cannot penetrate deeply. Convex transducers are designed for depth analysis. The lower emission frequency (i.e., 2-3 MHz) enables greater tissue penetration at the cost of resolution. Convex probes are suited for deeper abdominal or pelvic work.

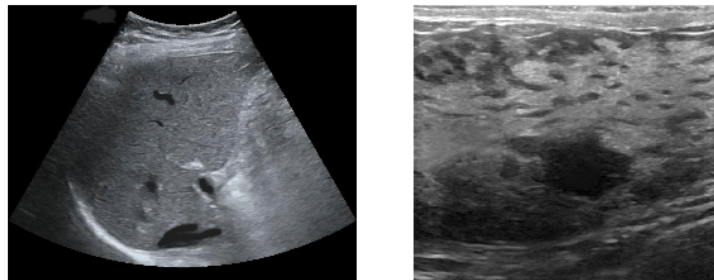


Figure 3.7: Example of the point-of-view of a convex probe exam (left) and a linear probe exam (right)

While the selection of the appropriate probe is primarily a clinical decision, the operator can further refine image quality through specific settings, most notably depth and gain. These parameters heavily rely on operator expertise. The depth setting adjusts how deep the machine listens for returning echoes. To examine a structure at 2 cm, setting the depth to 3 cm will maximise detail, whereas reviewing a particular anatomical part may require a deeper depth setting. Gain is essentially the brightness control; it amplifies the returning electrical signals. If the image is too dark, the gain can be increased; however, excessive gain will make the image too bright, washing out essential details. Modern machines enable gain adjustment at different depths to compensate for the natural attenuation of sound as it travels deeper into the body, thereby achieving various contrast/detail combinations.

Ultrasonic imaging relies on the assumption that a transmitted pulse travels in a straight line at a constant speed, reflects off a boundary, and returns directly to the transducer. Artefacts occur when these assumptions are violated. A common issue is the acoustic shadow, Fig. 3.10 B, which occurs when sound waves encounter a highly reflective surface, such as a rib or an oedema. This surface reflects almost all the sound, leaving a shadow or dark void deep beneath it because sound waves cannot penetrate the structures below. To scan the chest cavity, for instance, operators must orient the transducer between the ribs to avoid these shadows. Proper application



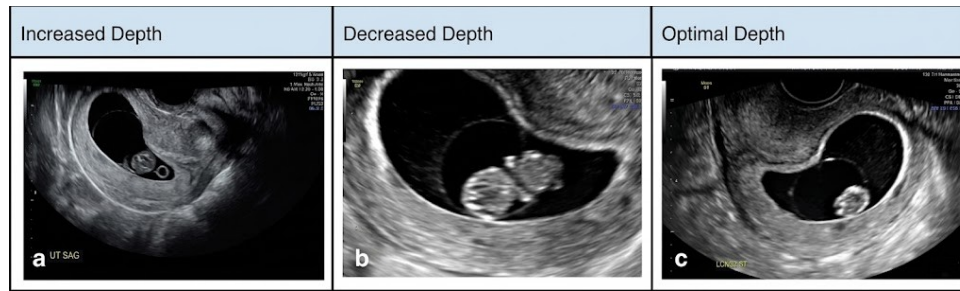


Figure 3.8: Effect on the depth parameter on the image sections and the correct identification of anatomically correct markers

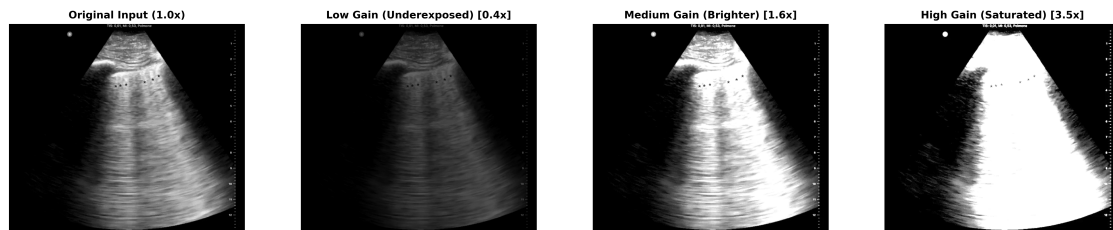


Figure 3.9: Effect on the gain parameter on the image resolution and the correct identification of anatomical correct markers

of ultrasound gel is also critical, as it eliminates air between the probe and the skin, facilitating the passage of sound waves. Another example is comet tail artefacts, Fig. 3.10: a comet tail forms when acoustic energy becomes trapped between surfaces with a significant acoustic impedance mismatch. Specifically, it occurs when sound is trapped between two highly reflective surfaces that are very closely spaced. These interfaces can be anatomical or foreign, such as small pockets of gas within tissues, tiny metal fragments, or cholesterol crystals.

While artefacts can sometimes obscure anatomical details, recognising specific patterns, such as the comet tail, is crucial for accurate diagnosis. The clinical utility of identifying these artefacts was clearly demonstrated during the COVID-19 pandemic, particularly in POCUS. Rather than returning immediately to the probe, the sound bounces repeatedly within this narrow space before eventually returning to the transducer. Because the ultrasound machine calculates depth based on the time elapsed between transmission and reception of the echo, it interprets these delayed, repeated echoes as real structures located deeper within the body. The visual result on the monitor is a bright, tapering white "tail" extending below the object. Consequently, the appearance of numerous, bright, tapering white tails originating from the pleural line often termed "B-lines" in pulmonary settings became a hallmark sonographic sign of COVID-19 interstitial pneumonia [123]. The detection and quantification of this artefact enabled clinicians to rapidly diagnose pulmonary involvement and monitor disease progression without relying solely on ionising radiation imaging, such as CT scans. In this context, the comet tail ceased to be mere "noise" and became a critical diagnostic data point. The inherent variability, noise, and artefact-rich nature of POCUS images make them a challenging but ideal candidate for AI analysis. An AI model capable of discerning subtle pathological patterns



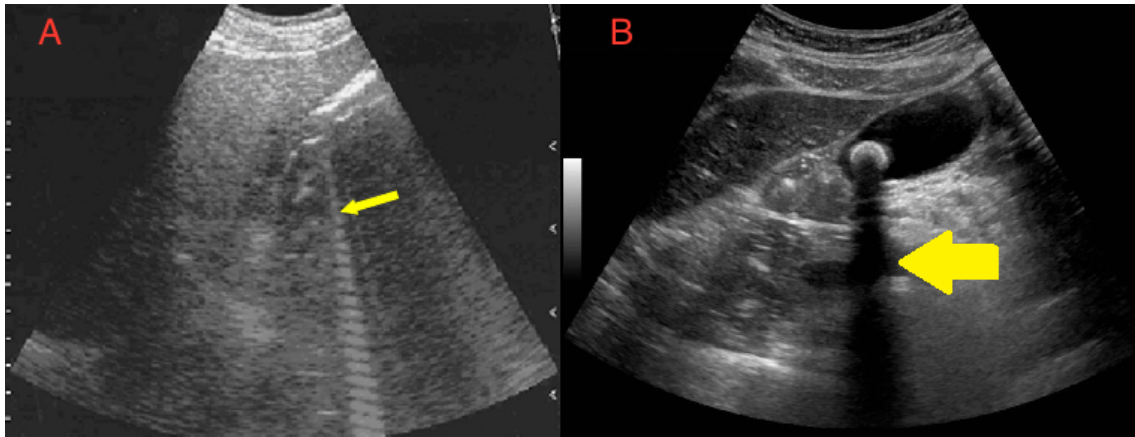


Figure 3.10: The image contains two examples of ultrasound artefacts, the comet tail and the shadowing. It can be seen that the two of them work in complementary ways: the first fills with erroneous information, while the latter obscures the image content.

amidst this noise can serve as a powerful assistive tool, standardising interpretation and augmenting clinicians' capabilities, especially in fast-paced emergencies such as the SARS-CoV-2 pandemic. Techniques and challenges for ultrasound image interpretation using DL methods are discussed in the following sections, with examples including a classification method for SARS-CoV-2 pneumonia and the detection of comet tails.

## 3.2 Deep learning: strategies and architecture

With a clear grasp of what an ultrasound image is, how it is generated, and which patterns are meaningful to examine, it is time to turn to the set of procedures, algorithms, and criteria for interpreting this type of data. The field of DL has evolved at a breathtaking pace; new, more performing models are produced yearly. The following section focuses on the key concepts central to our case studies of reliable LUS classifications. The analysis will introduce the concepts of data handling and pre-processing, architecture selection, training strategies, and evaluation criteria, with a focus on computational efficiency, transfer learning, and XAI visualisation techniques.

### 3.2.1 Data preparation

Preparing training, validation, and test sets for ultrasound adheres to fundamental machine learning principles, but the heterogeneity of the data sources cannot be overlooked. Because a wide range of devices produce LUS, capturing this hardware diversity is essential for robust model performance. However, compiling a dataset from a single source, such as the localised TiAssisto project, where a single specialist acquired the majority of the images, often results in limited data diversity. This can lead to overfitting and spurious correlations, which are classical problems

in DL practice that cannot be overlooked. The trained model, regardless of its precision or accuracy, can produce a pattern that is not a reliable depiction of correct decision-making. This problem is sometimes depicted as shortcut learning, in which the model focuses on the wrong part of the images. This incorrect part is usually a defined pattern, as indicated by the mark or label, derived from the equipment itself, rather than being anatomically reliable. This problem can be primarily fixed with a more diverse dataset, as shown in Fig. 3.11. Without a more varied dataset, the model can become polarised by repeated information, making it easier for the model to learn the patterns it aims to realise. The following figures show a set of different ultrasound images, all sampled with the same device; it appears that some markers are universal. While statistical learning theory suggests that sufficiently



Figure 3.11: A comparison of different ultrasound images acquired and saved through different devices.

large datasets enable DL models to disentangle causal features from background noise, relying on the Law of Large Numbers to marginalise non-robust correlations, this asymptotic assumption fails catastrophically in the data-scarce domain of medical imaging. On the contrary, in large general-domain datasets such as ImageNet [27], the sheer diversity of data acts as a natural regulariser, preventing the model from relying on a single background feature to predict a class. However, in medical imaging, which typically suffers from the "small N, large P" problem (few samples, high dimensionality), models are highly susceptible to the so-called Clever Hans effect for AI [78]. To understand why AI models can succumb to spurious correlations in medical imaging, we must look beyond intuition and examine the breakdown of Empirical Risk Minimisation (ERM). In statistical learning theory, we assume that a model  $f$  with parameters  $\theta$  attempts to minimise the true risk, as given in Eq. 3.11.

$$R(\theta) = \mathbb{E}_{(x,y) \sim P}[\mathcal{L}(f(x; \theta), y)]. \quad (3.11)$$

However, since the true distribution  $P$  is unknown, we minimise the empirical risk  $\hat{R}(\theta)$  over a dataset of size  $N$ . Let us denote the input medical image  $x$  as a composite

of two distinct latent feature sets as described in the following Equation 3.12:

$$x = \Phi(z_c, z_s) \quad (3.12)$$

Where:

- $z_c$ : The invariant causal features (e.g., a comet tail' shape).
- $z_s$ : The spurious features (e.g., a hospital tag, scanner noise, or a universal label artefact).

In a massive dataset where  $N \rightarrow \infty$ , if  $z_s$  is truly spurious, it will eventually appear with different labels, breaking the correlation. The conditional probability  $P(y|z_s)$  approaches the prior  $P(y)$ , rendering  $z_s$  non-predictive.

In the Small  $N$  regime typical of medical imaging, we often encounter a scenario where the spurious feature  $z_s$  exhibits a falsely high mutual information with the label  $y$  within the training distribution  $\mathcal{D}_{\text{train}}$ , such that:

$$I(y; z_s)_{\mathcal{D}_{\text{train}}} \gg I(y; z_s)_{\mathcal{D}_{\text{true}}} \approx 0 \quad (3.13)$$

Mathematically, the model seeks to minimise loss, but often exploits the high mutual information described in Eq. 3.13. If the universal label  $z_s$  provides a linear separability margin  $\gamma_s$  that is larger or simpler than the margin  $\gamma_c$  provided by the complex pathology  $z_c$ , the optimiser (e.g., SGD) will prioritise  $z_s$ . This phenomenon is often described via the Simplicity Bias of neural networks. If learning a function  $f(z_s)$  requires lower complexity (e.g., lower norm weights or lower frequency components) than learning  $f(z_c)$ , gradient descent will converge to a solution reliant on  $z_s$  long before it learns  $z_c$ . The model learns to minimise loss under the training distribution  $P_{\text{train}}(x, y)$ , which differs from the test distribution  $P_{\text{test}}(x, y)$  specifically in the correlation structure. If we model the relationship as the following equation 3.14.

$$y = \alpha \cdot z_c + \beta \cdot z_s + \epsilon \quad (3.14)$$

In the training set due to selection bias or shortcuts  $z_s$  and  $y$  are highly correlated ( $\rho_{sy} \approx 1$ ). Consequently, the model assigns a high weight to  $\beta$ . However, in the real world (Out-of-Distribution),  $z_s \perp y$ . The model, having learned a high  $\beta$ , will predict based on the artefact  $z_s$ , leading to catastrophic failure.

In this context, the robust solution (relying solely on  $z_c$ ) often corresponds to a flatter, harder-to-find minimum in the loss landscape. In contrast, the spurious solution (relying on  $z_s$ ) represents a sharp, easily accessible minimum. Without sufficient data volume to smooth out the spurious minimum, the model gets trapped in the shortcut.

Introducing diversity from an open-access dataset with the same distribution of objective classes can help mitigate these problems, as shown in the previous Fig.3.11; this diversity should act as a natural regulariser.

A possible further step, especially significant in the medical domain, is the removal of all content except that containing the anatomical information. This procedure is possible in all cases where the additional information is not from an anatomical perspective; otherwise, the actual information content would be altered.

A robust countermeasure to the susceptibility of medical models to spurious correlations is the strict enforcement of Anatomical Isolation. This procedure involves the algorithmic removal of all information content extraneous to the biological region of interest (ROI). Mathematically, this can be viewed as applying a binary mask  $M$  to the input  $x$ , such that the processed input  $\hat{x} = M \odot x$  retains only the causal features  $z_c$  while setting the spurious subspace  $z_s$  to zero.

This preprocessing step, valid only when the background confers no pathological information, serves a dual purpose: it acts as a hard inductive bias for model training and a critical safeguard for data privacy.

In cybersecurity, DL models are increasingly viewed as potential vectors for data leakage. A significant threat is the Model Inversion Attack [66], in which an adversary uses the learned weights  $\theta$  to reconstruct the original training samples.

DNNs are known to memorise high-frequency details of the training data. In medical imaging, raw DICOM files often contain burned-in annotations (e.g., patient names, hospital IDs, or acquisition timestamps) within the pixel data itself. If a model is trained on these raw inputs, it essentially compresses this Sensitive Personally Identifiable Information (SPII) into its parameters.

Formally, if an adversary seeks to reconstruct an image  $x$  given access to the model  $f_\theta$  and a target output  $y$ , they optimize:

$$x^* = \arg \min_x \mathcal{L}(f(x; \theta), y) + \mathcal{R}(x) \quad (3.15)$$

in the previous Eq. 3.15 the term  $\mathcal{R}(x)$  is used a regularization ensuring image realism. If the original training set included background text or artefacts, the reconstructed  $x^*$  will often reveal this text, leading to a privacy breach. By enforcing Anatomical Isolation ( $\hat{x}$ ), we ensure that even if a successful inversion attack occurs, the reconstructed image yields only generic biological structures, decoupling the patient's identity from the anatomical data and significantly reducing the privacy risk surface. From a learning dynamics perspective, removing non-anatomical content simplifies the optimisation landscape. DL training can be visualised as a search through a high-dimensional hypothesis space. The inclusion of background artefacts increases the variance of the input distribution  $P(x)$ , effectively widening the search space the optimiser must traverse. By removing  $z_s$  pre-training, we collapse the dimensionality of the manifold that the model must approximate. This leads to two distinct improvements:

1. Accelerated Convergence: with the removal of irrelevant high-frequency components (noise/artefacts), the gradient signal  $\nabla_\theta \mathcal{L}$  becomes less noisy, allowing the optimiser to descend toward the minimum more directly;
2. Prevention of Shortcut Learning: by manually setting  $I(y; z_s) = 0$  via cropping or segmentation, we physically prevent the model from latching onto the Universal Label. The model is forced to exploit gradients solely from  $z_c$ , ensuring that the learned features are pathologically relevant and generalizable to external datasets.

From a learning dynamics perspective, the excision of non-anatomical content fundamentally simplifies the optimisation landscape, as DL training can be conceptualised as a search through a high-dimensional space. Artefact inclusion in the

space can increase the variance of the input distribution  $P(\mathbf{x})$ , effectively widening the search space the optimiser must traverse. By removing the artefact component  $\mathbf{z}_s$  before training, we collapse the dimensionality of the manifold that the model must approximate, leading to accelerated convergence, Fig.3.12, as the gradient signal  $\nabla_{\theta}\mathcal{L}$  becomes less noisy, and simultaneously preventing shortcut learning by effectively forcing the mutual information  $I(\mathbf{y}; \mathbf{z}_s)$  to zero, which compels the model to exploit gradients solely from the clinical content  $\mathbf{z}_c$  to ensure generalisability. Op-

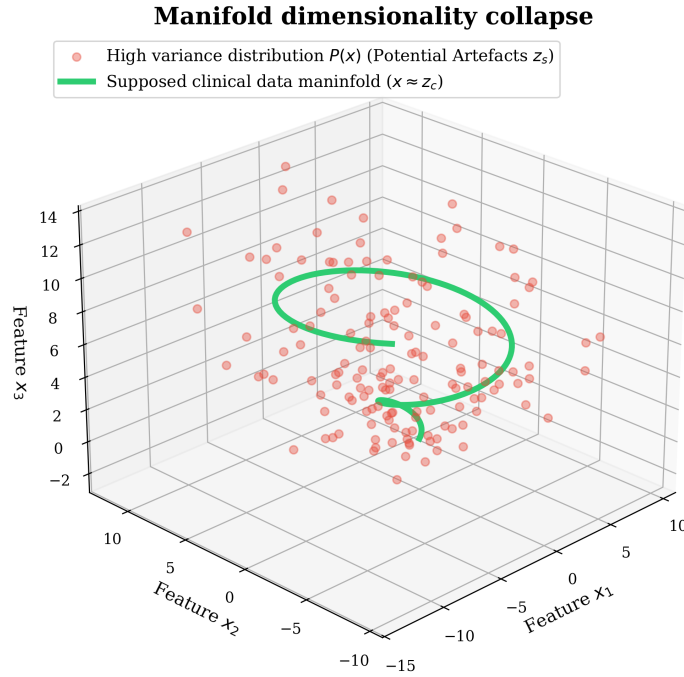


Figure 3.12: Manifold dimensionality collapse, the Artefacts expand the input space into a high-variance cloud (red points). The removal forces the data distribution to collapse onto the lower-dimensional manifold, making clinical details easier to extract.

erationalizing the removal of these annotations can be initially approached through classical computer vision techniques that rely on the statistical distinctiveness of high-intensity alphanumeric overlays compared to grayscale biological tissue; for instance, a binary mask  $M_{bin}(\mathbf{x}, \mathbf{y})$  can be derived from the input image  $I(\mathbf{x}, \mathbf{y})$  via a global threshold  $\tau$ , such that  $M_{bin}(\mathbf{x}, \mathbf{y}) = 1$  if  $I(\mathbf{x}, \mathbf{y}) > \tau$  and 0 otherwise. To resolve the ambiguity between textual artefacts and bright anatomical structures like the pericardium, this thresholding is typically refined using morphological operators, specifically the morphological opening  $M_{open} = (M_{bin} \ominus B) \oplus B$ , where  $\ominus$  and  $\oplus$  denote erosion and dilation respectively by a structuring element  $B$ , theoretically pruning the manifold to isolate high-frequency text components. A second possible approach is the contouring function, which serves as a natural selector for a specific



area with a defined shape and pixel-intensity characteristics. One of the most useful contour detection functions, notably used in the OpenCV C++ and Python library [13], implements the topological structural analysis algorithm proposed by Suzuki and Abe [124]. Mathematically, the algorithm operates on a binary matrix  $I(x, y)$  where pixels take discrete values of 0 (background) or 1 (foreground). The process begins with a linear raster scan (row-by-row) of the matrix until a border transition is encountered. Two types of borders are defined:

- Outer Border: A transition from  $0 \rightarrow 1$ , indicating the entry into a connected component (shape).
- Hole Border: A transition from  $1 \rightarrow 0$ , indicating the entry into a hollow space enclosed within a shape.

Once a transition pixel  $P_0$  is found, the algorithm switches from raster scanning to a border following mode. It traces the perimeter of the component by examining the 8-connected neighbourhood (the 8 pixels immediately surrounding the current pixel) in a clockwise or counter-clockwise direction to locate the next non-zero pixel  $P_1$ .

This tracing process continues recursively until the path returns to the starting pixel  $P_0$ . The output is a vector of coordinates  $C = \{P_0, P_1, \dots, P_n\}$  representing the polygon of the contour. Additionally, the algorithm generates a hierarchy tree that mathematically defines the topological relationships between contours, categorising them as primary shapes (enclosing shapes) or secondary shapes (nested holes); this procedure shows enough reliability, Fig.3.13, and can be used in situations when an exploration dataset search is needed.



Figure 3.13: In this side-by-side comparison, the chosen filter perfectly identifies the area of interest in the image

However, because heuristic functions are brittle under variable ultrasound gain settings, this procedure is often iterative. It may need attention for a specific subset of images in which the algorithm manipulates too much of the information, altering the images' content (Fig.3.14), or simply fails to function correctly (Fig.3.15).

Modern cleaning pipelines increasingly integrate OCR (Optical Character Recognition) systems that utilise deep representational learning to distinguish text from texture. Specifically, these systems employ Convolutional Recurrent Neural Networks (CRNNs), often implemented in frameworks such as the Keras OCR module [19]. In this architecture, a backbone Convolutional Neural Network (CNN) processes the image to generate a feature map sequence, Eq. 3.16, denoted as:

$$X = \{x_1, x_2, \dots, x_T\} \quad (3.16)$$

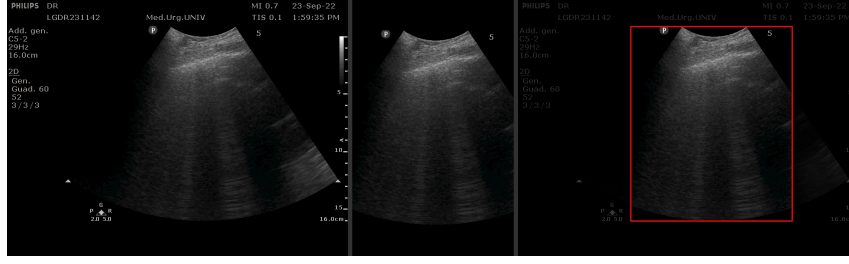


Figure 3.14: The contour detection algorithm employed in this instance exhibits excessive sensitivity to subtle pixel intensity gradients between the region of interest and the background. This results in a substantial distortion of the information content, which is methodologically unacceptable. In medical imaging, the erroneous exclusion of data or the misinterpretation of empty space constitutes a significant source of diagnostic error.

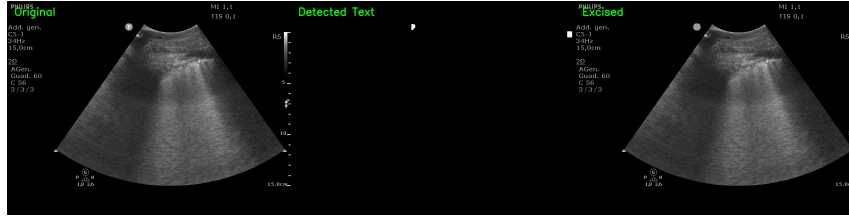


Figure 3.15: Example of the use of OCR detection, the filter, even if really fast, needs a more complex characterisation to work on the given dataset; the methods partially work, recognising only the letter 'P'

This sequence captures local spatial patterns, which are subsequently fed into a bi-directional Long Short-Term Memory (LSTM) network. Within this network, the hidden state  $h_t$  is updated recursively:

$$h_t = \text{LSTM}(x_t, h_{t-1}) \quad (3.17)$$

This mechanism models the syntactic and contextual dependencies between character strokes. Consequently, this hybrid structure allows the system to be optimised via the Connectionist Temporal Classification (CTC) loss function, defined in the following Eq. 3.18.

$$\mathcal{L}_{CTC} = -\log P(Y|X) \quad (3.18)$$

This function marginalises over all possible alignments of the predicted character sequence  $Y$ , thereby providing a robust, data-driven mechanism for detecting and excluding annotations without requiring pixel-perfect pre-segmentation. The clean result that this approach provides can be seen in the following Fig. 3.16.

In the context of preparing sonographic data for deep representational learning, a fundamental distinction must be drawn between statistical normalisation, which is designed primarily to condition the optimisation landscape, and signal enhancement techniques, specifically Contrast Limited Adaptive Histogram Equalisation (CLAHE), which fundamentally alter the local information content of the signal. While both processes transform the pixel intensity distribution, their objectives,

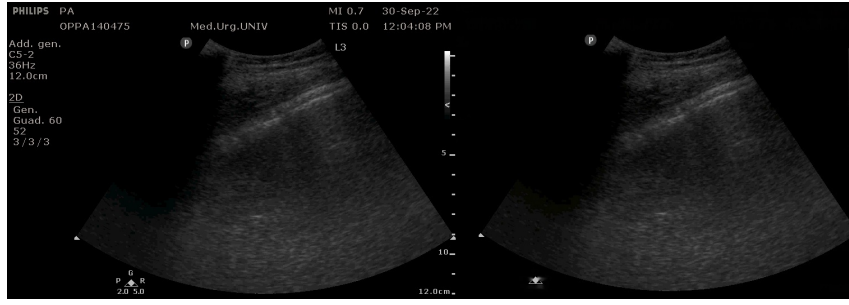


Figure 3.16: The picture shows that the DL-based OCR algorithm has a very high detection precision and rarely needs excessive fine-tuning to work; the image, in fact, is just the result of the application of the OCR-Keras pre-trained pipeline over the two analysed datasets

mathematical formulations, and impacts on the model’s learning dynamics are distinct and often complementary. In ultrasound imaging, where the raw signal is heavily dependent on operator-defined gain settings, depth-dependent attenuation, and the constructive and destructive interference patterns of speckle noise, the raw pixel intensities lack a standardised physical meaning comparable to Hounsfield units in Computed Tomography. Consequently, the input distribution  $P(\mathbf{x})$  often exhibits high intra-class variance that is unrelated to anatomical truth.

The first class of transformations, statistical normalisation, is a prerequisite for numerical stability in gradient descent-based optimisation. In the high-dimensional non-convex optimisation landscape of a DNN, the scale of the input features directly influences the scale of the gradients. If the input features  $\mathbf{x}$  are not standardised, the parameters  $\theta$  associated with features of larger magnitudes will update significantly faster than those associated with smaller magnitudes, leading to pathological curvature in the loss landscape and necessitating a vanishingly small learning rate.

The most common approach is  $z$ -score standardisation, which assumes that the pixel intensities follow a normal distribution. For a given dataset or batch, we calculate the global mean  $\mu$  and standard deviation  $\sigma$ . The normalised input  $\hat{\mathbf{x}}_{norm}$  is computed as Eq. 3.19.

$$\hat{\mathbf{x}}_{norm} = \frac{\mathbf{x} - \mu}{\sigma + \epsilon} \quad (3.19)$$

Where  $\epsilon$  is a small stability constant. This linear transformation shifts and scales the manifold such that it is centred at zero with unit variance, but crucially, it preserves the relative distances between pixel values. The structural relationships and local contrast ratios remain invariant; a dark region remains dark relative to a bright region, and the histogram shape is preserved, merely translated and scaled along the intensity axis.

Alternatively, Min-Max normalisation is employed to bound the inputs within a fixed range, typically  $[0, 1]$  or  $[-1, 1]$ , defined as Eq. 3.20.

$$\hat{\mathbf{x}}_{mm} = \frac{\mathbf{x} - x_{min}}{x_{max} - x_{min}} \quad (3.20)$$

While these operations effectively precondition the Hessian matrix of the loss func-



tion  $\mathcal{L}$ , ensuring that the contours of the error surface are more spherical rather than elliptical, they do not address the inherent limitations of ultrasound visibility, such as poor contrast in soft-tissue interfaces or shadowing artefacts. They are passive standardizers rather than active feature enhancers.

In contrast to the linear, global operations of statistical normalisation, Contrast Limited Adaptive Histogram Equalisation (CLAHE) is a non-linear [99] local signal-processing technique that maximises image entropy to reveal latent morphological features. Standard Histogram Equalisation (HE) seeks a transformation function  $T$  that maps the input intensities  $r$  to output intensities  $s$  such that the output probability density function (PDF) is uniform. Mathematically, for a discrete image with intensity levels  $r_k$ , the transformation is given by the cumulative distribution function (CDF), as shown in Eq. 3.21.

$$s_k = T(r_k) = (L - 1) \sum_{j=0}^k p_r(r_j) \quad (3.21)$$

Where  $L$  is the number of grey levels and  $p_r(r_j)$  is the normalised histogram count of intensity  $r_j$ . While mathematically elegant, applying this globally to ultrasound images is often disastrous; global HE tends to over-amplify the background speckle noise in homogeneous regions (such as fluid-filled cysts or the liver parenchyma) because the histogram is dominated by these low-intensity dark pixels. The transformation stretches the contrast in these dominant regions, rendering the noise as high-contrast features that can confuse a convolutional feature extractor. CLAHE addresses this by introducing two critical modifications: locality and contrast limiting. Instead of computing a global histogram, the image is partitioned into a grid of contextual regions or tiles (e.g.,  $8 \times 8$ ). Within each local tile, the histogram is computed to capture the local intensity distribution. To prevent over-amplification of noise, which mathematically corresponds to a CDF with a very steep slope, CLAHE imposes a clipping limit,  $\beta$ , on the histogram bins. If a histogram bin  $h(i)$  exceeds this limit  $\beta$ , the excess pixels are clipped. Let  $N_{\text{excess}}$  be the total number of pixels clipped from all bins. These pixels are not discarded; instead, they are redistributed uniformly across all histogram bins to ensure the PDF integral remains unity (equal to the total pixel count). The modified histogram  $h_{\text{mod}}(i)$  is thus constructed such that the slope of the resulting mapping function is constrained. The redistribution can be expressed as an iterative process or as a direct addition of a defined ratio to each bin, as given in Eq. 3.22, ensuring that no bin's height exceeds the limit.

$$\frac{N_{\text{excess}}}{L} \quad (3.22)$$

Following this, the standard histogram equalisation transformation is applied using the modified histogram. To prevent blocking artefacts at tile boundaries, the final pixel values are generated by bilinearly interpolating the mapping functions derived from the neighbouring tile centres.

From a manifold perspective, while statistical normalisation preserves the topology of the input data manifold, CLAHE distorts it. By locally flattening the PDF of pixel intensities, CLAHE effectively spreads the data more uniformly across the

available dynamic range. In ultrasound, this enhances the high-frequency components, edges, and subtle tissue interfaces, which are critical for segmentation tasks. For example, the boundary between a tumour and the surrounding tissue, which might be obscured by low contrast in the raw data, becomes more pronounced in the gradient domain  $\nabla I$  after CLAHE processing. This suggests that CLAHE can be viewed as a form of "feature engineering" that injects prior knowledge (that information should be maximally uniform) into the pipeline.

However, the interaction between these two preprocessing steps must be carefully managed. Since CLAHE modifies the histogram, it alters the image's global mean and variance. Therefore, logical ordering dictates that CLAHE should precede statistical normalisation. The pipeline operates as shown in Eq. 3.23.

$$\mathbf{x} \xrightarrow{\text{CLAHE}} \mathbf{x}_{\text{enhanced}} \xrightarrow{\text{Norm}} \hat{\mathbf{x}} \quad (3.23)$$

If one were to normalise first, the non-linear warping of CLAHE would destroy the zero-mean, unit-variance property that normalisation aims to preserve. Furthermore, the application of CLAHE introduces a hyperparameter search space, specifically the tile grid size and the clip limit  $\beta$ . A low  $\beta$  approximates the original image (linear scaling), while a high  $\beta$  approaches standard Adaptive Histogram Equalisation, risking the amplification of specific ultrasound artefacts like acoustic shadowing or reverberation. If the clip limit is too high, the neural network may learn to detect the amplified speckle patterns rather than the anatomical structure, leading to overfitting on the noise distribution of the training set. Formally, we define an augmented sample  $\mathbf{x}'$  as the result of a spatial mapping applied to the pixel coordinates of the original image  $\mathbf{x}$ . For a coordinate vector  $\mathbf{p} = [u, v, 1]^T$  in the source image, the transformed coordinates  $\mathbf{p}'$  are computed via an affine transformation matrix  $M$ . To simulate the realistic jitter of a handheld probe without violating the directional constraints of the acoustic beam, we restrict the rotation angle  $\theta$  to a narrow range (e.g.,  $\theta \in [-10^\circ, 10^\circ]$ ). The mapping is formally expressed as Eq. 3.24.

$$\begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = \begin{bmatrix} s_u \cos \theta & -s_v \sin \theta & t_u \\ s_u \sin \theta & s_v \cos \theta & t_v \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad (3.24)$$

where  $s_u, s_v$  represent scaling factors simulating changes in zoom or depth settings, and  $t_u, t_v$  represent translations corresponding to lateral probe movements. Furthermore, given the non-rigid nature of biological soft tissue, simple affine transformations are often insufficient to capture the variance in organ morphology caused by probe pressure. To address this, we employ elastic deformations, treating the tissue as a continuous medium that can be distorted. This is achieved by generating a random displacement field  $\Delta(u, v)$  typically derived from Gaussian noise convolved with a smoothing kernel  $\sigma$ . The pixel value at the new location is determined by the following pixel mapping Eq. 3.25,

$$\mathbf{x}'(u, v) = \mathbf{x}(u + \Delta_u(u, v), v + \Delta_v(u, v)) \quad (3.25)$$

This process forces the neural network to learn invariant features that are robust to the inherent malleability of organs, effectively simulating the physiological com-

pression that occurs during a live examination. By strictly limiting  $\mathcal{T}$  to these physics-compliant transformations, we ensure that the augmented manifold  $\mathcal{M}_{aug}$  remains a subset of the valid biological manifold, thereby improving the generalisability of the model without introducing noise artefacts that contradict the axioms of acoustic propagation.

### 3.2.2 Architectures

The last decade has witnessed an explosion of novel architectures in computer vision, catalysed by the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). Each of these architectures was deemed a foundational model. This section outlines the historical succession of these models, analysing the architectural and conceptual leaps from the pioneering deep networks (AlexNet) through the race for extreme depth (VGG, ResNet) to the modern era of computational efficiency (EfficientNet). Finally, the 'Efficient Adaptable Ensemble' paradigm is introduced, representing a synthesis of the field's accumulated knowledge prioritising architectural intelligence over raw scale. To understand the magnitude of the shift in modern computer vision, one must consider the transition from Multi-Layer Perceptrons (MLPs) to Convolutional Neural Networks (CNNs). This represents a fundamental change in the inductive bias of machine learning models. In a standard fully connected layer, every input neuron interacts with every output neuron, implying no prior assumption about the spatial structure of the data. While theoretically capable of approximating any continuous function, MLPs suffer from the Curse of Dimensionality. For a standard  $224 \times 224$  pixel RGB image, a single neuron in the first hidden layer would require connections to every input pixel:

$$3 \text{ channels} \times 224 \times 224 \approx 150,000 \text{ weights} \quad (3.26)$$

This explosion in the number of parameters, as in Eq. 3.26, leads to severe overfitting and computational intractability. The solution lay in the CNN, which introduces three structural priors: local connectivity, parameter sharing (translational invariance), and hierarchical pooling. At the heart of the architecture is the discrete cross-correlation. For an input image  $I$  and a learnable kernel  $K$  of dimensions  $k_h \times k_w$ , the feature map  $S$  at position  $(i, j)$  is given by Eq. 3.27.

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(i + m, j + n) K(m, n) \quad (3.27)$$

This operation enforces stationarity, drastically reducing the parameter space from  $\mathcal{O}(W \cdot H \cdot W \cdot H)$  to  $\mathcal{O}(k_h \cdot k_w)$ . A defining characteristic of CNNs is the expansion of the Receptive Field (RF). If  $r_{l-1}$  is the receptive field of layer  $l-1$ , the RF of layer  $l$  with kernel size  $k_l$  and stride  $s_l$  expands as:

$$r_l = r_{l-1} + (k_l - 1) \times \prod_{i=1}^{l-1} s_i \quad (3.28)$$

This hierarchical stacking, Eq. 3.28, allows the network to compose simple geometric primitives into complex semantic concepts. Pooling layers (Max or Average) introduce translation invariance. By summarising a neighbourhood  $\mathcal{R}_{i,j}$ , the network

becomes robust to small shifts given by the following equation 3.29.

$$y_{i,j} = \max_{(p,q) \in \mathcal{R}_{i,j}} x_{p,q} \quad (3.29)$$

AlexNet [74] stands as the milestone moment in modern computer vision, the model that decisively demonstrated the overwhelming superiority of deep convolutional neural networks (ConvNets) over traditional, feature-engineered methods. Its victory in the ILSVRC 2012 competition was not merely an incremental improvement; it was a paradigm shift that ignited an intense and focused research effort into DL for vision tasks. The model's success was not due to a single innovation, but to a strategic combination of architectural choices and novel training techniques that collectively addressed the challenges of training an extensive network at unprecedented scale. A combination of innovations is introduced in the paper that are now staples of the majority of DL architectures for computer vision:

- ReLU Nonlinearity: replacing Tanh/Sigmoid with  $f(x) = \max(0, x)$  mitigated the vanishing gradient problem;
- Dropout: to combat overfitting in a model with  $\approx 60$  million parameters, neurons were randomly dropped with probability  $p = 0.5$ . Mathematically, the expected output becomes as Equation 3.30.

$$\mathbb{E}[y] = (1 - p) \sum w_i x_i \quad (3.30)$$

- Local Response Normalisation : a form of lateral inhibition inspired by neuroscience.

Following the success of AlexNet, architectural exploration turned to a central question: what is the effect of network depth on performance? The VGGNet architecture, from the Visual Geometry Group at Oxford [122], addressed this with a disciplined, almost scientific experiment. Rather than introducing novel components, VGGNet's primary contribution was a thorough evaluation of increasing network depth in a controlled manner. By isolating depth as the key variable and keeping architectural motifs homogeneous, VGGNet validated the hypothesis that substantial accuracy gains could be achieved by simply stacking more layers, cementing the "deeper is better" mantra of the era. The core principles of the VGGNet architecture were its uniformity and its methodical increase in depth. VGGNet exclusively used  $3 \times 3$  filters. A stack of three  $3 \times 3$  layers has the same effective receptive field as a single  $7 \times 7$  layer but offers fewer parameters:

$$\text{Stack: } 3 \times (3^2 \times C^2) = 27C^2 \quad (3.31)$$

$$\text{Single: } 7^2 \times C^2 = 49C^2 \quad (3.32)$$

This reduction acts as regularisation while increasing the non-linear density. As networks deepened, researchers encountered the degradation problem. While VGGNet conclusively demonstrated the value of depth, its brute-force approach also exposed a fundamental barrier: simply stacking more layers indefinitely was not a viable

long-term strategy, as it would soon lead to a new, unexpected problem. The success of VGGNet led to a natural impulse to build even deeper networks. However, researchers quickly encountered a perplexing obstacle: the degradation problem. Counterintuitively, as plain networks became deeper, their performance would saturate and then rapidly degrade. This was not a case of overfitting; the deeper models exhibited higher training error than their shallower counterparts, indicating a fundamental optimisation challenge. The introduction of the Deep Residual Learning framework, or ResNet [42], was a revolutionary breakthrough that elegantly solved this problem and enabled the training of networks of unprecedented depth. The ResNet architecture was built on a simple yet profound insight into how layers could be reformulated to ease training. ResNet introduced Deep Residual Learning to solve this. Instead of learning a mapping  $H(\mathbf{x})$ , the network learns a residual function  $F(\mathbf{x}) := H(\mathbf{x}) - \mathbf{x}$ . The gradient of the loss  $\mathcal{L}$  with respect to input  $\mathbf{x}_l$  becomes:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}_l} = \frac{\partial \mathcal{L}}{\partial \mathbf{x}_{l+1}} \left( 1 + \frac{\partial}{\partial \mathbf{x}_l} F(\mathbf{x}_l, W_l) \right) \quad (3.33)$$

The term one in Eq. 3.33 allows gradients to flow directly through the network (well-guided gradients), enabling the training of models with 150+ layers. After ResNet demonstrated that extreme depth was attainable, the next evolutionary wave in model design pivoted to address computational and parameter efficiency. This shift was not merely an academic exercise; it was a practical necessity driven by the immense computational costs introduced by the "deeper is better" era, exemplified by models like VGG and very deep ResNets. The goal was no longer to build powerful models, but to build models that achieved superior performance with fewer parameters and a smaller computational footprint, making real-world deployment more feasible. This era produced several innovative architectures, including DenseNets and MobileNets, demonstrating that state-of-the-art results did not require monolithic, computationally expensive designs. DenseNet introduced a novel connectivity pattern to maximise information flow and encourage feature reuse throughout the network. The core innovation of DenseNet is that it connects each layer to every other layer in a feed-forward fashion. For any given layer  $l$ , it receives the feature maps of all preceding layers, Eq. 3.34.

$$\mathbf{x}_l = H_l([\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_{l-1}]) \quad (3.34)$$

This encourages feature reuse and parameter efficiency. Designed for mobile devices, MobileNet [48] introduced Depthwise Separable Convolutions. The computational cost reduction, Eq. 3.35, is proportional to:

$$\frac{1}{N} + \frac{1}{K^2} \quad (3.35)$$

where  $N$  is the number of kernels and  $K$  is the kernel size.

### 3.2.3 Efficient Architecture

In 2019, the introduction of EfficientNet marked a significant turning point, challenging the notion that higher accuracy must come at the expense of exponentially

greater complexity. The researchers behind EfficientNet identified a critical flaw in previous approaches: model scaling was often done arbitrarily along a single dimension, such as increasing the network depth (more layers), width (more channels), or input image resolution. They demonstrated that these dimensions are interdependent and that balancing them is key to optimal performance. This led to the development of compound scaling, a principled method for scaling all three dimensions (depth, width, and resolution) in unison using a simple yet effective compound coefficient. EfficientNet [126] moved beyond ad-hoc scaling. The authors proposed a Compound Scaling Method where depth ( $d$ ), width ( $w$ ), and resolution ( $r$ ) are scaled uniformly by a coefficient  $\phi$  as the following Eq. 3.36.

$$d = \alpha^\phi, \quad w = \beta^\phi, \quad r = \gamma^\phi \quad (3.36)$$

Subject to  $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ . This principled approach achieves SOTA accuracy with significantly fewer FLOPs.

The second key innovation of EfficientNet was its foundational building block: the inverted bottleneck MBConv module, first introduced in MobileNetV2 [111]. Traditional convolutional blocks often compress the number of channels, process the data, and then expand them back. The MBConv does the opposite: it first expands the channels, performs a lightweight "depthwise" convolution (which processes each channel independently), and then compresses the channels back down. This structure was found to be significantly more parameter-efficient. The combination of compound scaling and the efficient MBConv block resulted in a family of models (from EfficientNet-b0 to b7) that set new standards for accuracy and efficiency. As detailed in the LUS classification study, EfficientNet-b0, the smallest model in the family, achieves an accuracy of 77.1% on ImageNet with only 5.3 million parameters and 0.39 G-FLOPs, outperforming many larger, more complex models [126]. The largest model, EfficientNet-b7, reached a then-state-of-the-art accuracy of 84.3% with 67 million parameters, still significantly fewer than many of its contemporaries [126]. This remarkable trade-off between accuracy and complexity makes EfficientNet-b0 an ideal choice for applications where efficiency is critical, such as the real-time LUS classification system.

### 3.2.4 U-net and Transformer

While mainstream evolution focused on classification backbones (ResNet, EfficientNet), parallel developments focused on dense prediction and non-convolutional paradigms. U-Net is a model for biomedical image segmentation and differs from traditional CNNs by eliminating the fully connected layer, as its primary objective was segmentation rather than classification. U-Net [106] introduced a symmetric architecture that preserves spatial fidelity. Unlike classification models that reduce dimensionality to a single vector, U-Net requires pixel-level output. It consists of a contracting path (encoder) to capture context and a symmetric expanding path (decoder) that enables precise localisation. The critical innovation is the use of skip connections that concatenate feature maps from the encoder directly to the decoder, an evolution of the previous Fully Convolutional Model, FCC, from which it is derived, Eq 3.37.

$$x_{dec}^l = \text{Up}(x_{dec}^{l-1}) \oplus x_{enc}^l \quad (3.37)$$

This allows the network to recover fine-grained spatial details lost during pooling, making it foundational for segmentation and modern diffusion generation models.

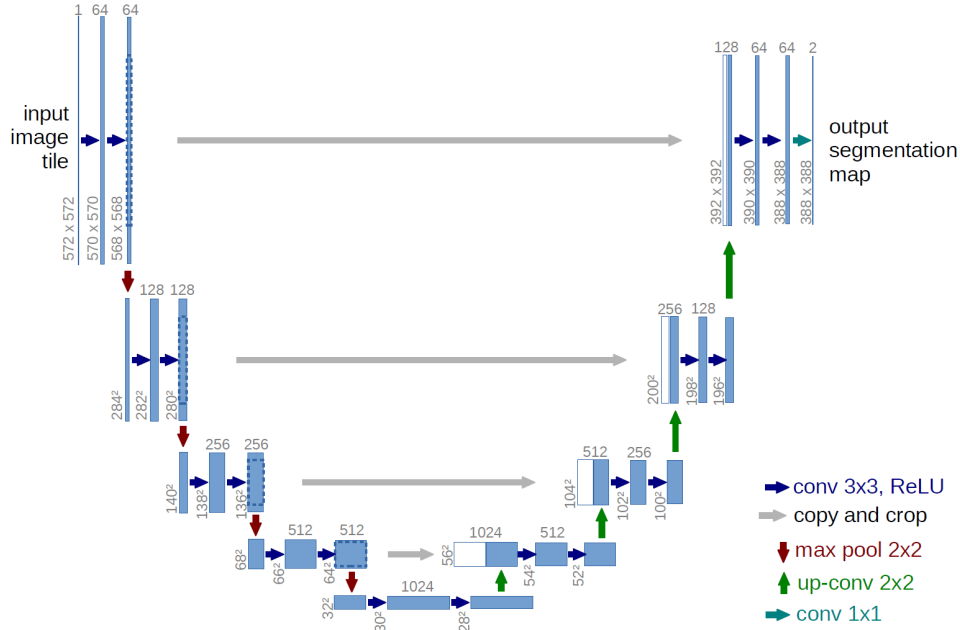


Figure 3.17: The main architecture used for U-net, the encoder and decoder architectures are mirrored structures connected by the skip connection to grant residual learning capacity at the network

The Vision Transformer (ViT) [29] marked a radical departure from CNNs by applying the self-attention mechanism, a dominant NLP technique, directly to images. ViT splits an image into fixed-size patches (e.g.,  $16 \times 16$ ), linearly embeds them, and adds positional embeddings. These token sequences are processed by Transformer encoders using Multi-Head Self-Attention (MSA), Eq. 3.38 [134]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3.38)$$

where  $Q, K, V$  represent queries, keys, and values. This mechanism enables global connectivity among pixels in the very first layer, unlike the slowly expanding receptive fields of CNNs. Despite the rise of Transformers, CNN-based architectures (including ResNets and EfficientNets) remain preferred in many scenarios due to their strong Inductive Bias.

1. Inductive Bias vs. Data Hunger: CNNs have hard-coded assumptions about images: locality (pixels define features with their neighbours) and translation invariance (a cat in the corner is still a cat). Transformers lack these priors and must learn them from scratch. Consequently, ViT requires massive datasets (e.g., JFT-300M) to outperform CNNs; on standard datasets (ImageNet-1k), CNNs often generalise better with less data.
2. Computational Complexity: the self-attention mechanism scales quadratically with the number of tokens ( $O(N^2)$ ), making high-resolution processing computationally prohibitive without modification. In contrast, Convolution scales



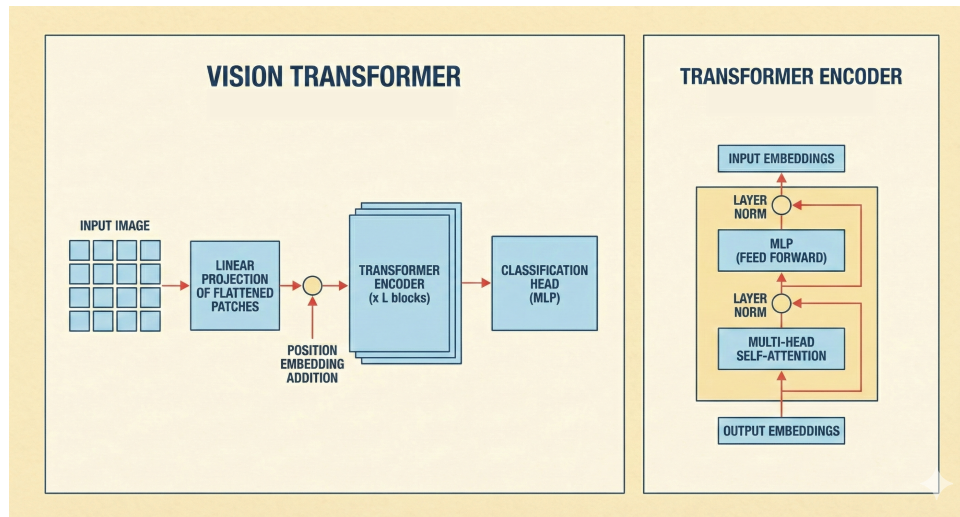


Figure 3.18: A visual transformer uses the same attention mechanism from the basic transformer architecture; the innovation lies in the catching mechanism for the input image

linearly ( $O(N)$ ), making CNNs far more efficient for high-resolution edge deployment.

3. Hierarchical Structure: CNNs naturally build features from simple edges to complex semantics. While hierarchical Transformers exist (e.g., Swin), standard CNNs remain the most optimised and intuitive baseline for feature extraction, where local texture and shape are paramount.

### 3.2.5 The Minimal Adaptive Ensembling Methodology

While efficient architectures and clever training strategies form the building blocks of high-performing AI systems, ensemble learning provides a powerful way to push performance to its absolute limits. By combining the predictions of several individual models, an ensemble can achieve better generalisation and robustness than any of its constituent members. However, this benefit has historically come at a steep price in computational complexity. This chapter details an innovative methodology, the efficient adaptive ensemble, that captures the power of ensembling while sidestepping its traditional complexity drawbacks, representing the core technical contribution of the classification strategy later used to analyse LUS.

Ensemble learning is based on a simple yet powerful idea: the predictive power of a diverse set of algorithms is often greater than that of any single classifier. In machine learning, this translates to combining several models, known as weak learners, to produce a single, superior strong learner (the combination of CART in Random Forest acts very similarly). The key to a successful ensemble is the diversity of its members; if all the weak learners make the same mistakes, combining them yields no benefit. This diversity can be achieved by training models on different subsets of the data (e.g., bagging), by training models sequentially to correct the errors of their predecessors (e.g., boosting), or by training different types of models and using



another model to learn how best to combine their predictions (e.g., stacking)

The most common method of combination is through voting for classification tasks (where the majority-voted class is the final prediction) or averaging for regression tasks. While effective, these standard approaches suffer from significant drawbacks, especially in the context of DL: complexity scales linearly with ensemble size, and aggregation remains static, failing to adaptively weight component model contributions based on their specific feature representations. To address these limitations and reverse the prevailing trend, where minor accuracy improvements require exponential increases in model complexity, this work proposes the Efficient Adaptive Ensembling method.

### 3.2.6 Architecture of the Efficient Adaptive Ensemble

This approach fundamentally reimagines the ensemble process by integrating a specific design choice: replacing the static voting mechanism with a trainable combination layer that operates on extracted features rather than final outputs. The methodology leverages the EfficientNet-b0 architecture as the core weak learner due to its superior trade-off between accuracy and computational resources, achieved through compound scaling, a technique that uniformly scales network width, depth, and resolution. By utilising EfficientNet-b0, the system ensures that the ensemble's foundational building blocks remain lightweight and amenable to rapid training. The training protocol employs a rigorous bagging strategy that partitions the training dataset into disjoint subsets, maintaining class stratification to ensure balanced representation. Distinct weak learners are trained end-to-end on these specific subsets. A counterintuitive yet pivotal aspect of this phase is the deliberate training of weak learners to the point of overfitting. While overfitting is typically avoided in standalone models due to poor generalisation, in this adaptive ensemble context, highly specialised weak learners contribute distinct, high-variance perspectives that the subsequent combination layer can effectively synthesise. This observation aligns with findings suggesting that overfitted component models can enhance the ultimate quality of an ensemble, particularly when the combination mechanism is adaptive. Following the training of weak learners, the adaptive phase constructs the strong learner not by averaging outputs, but by concatenating the feature vectors extracted by the penultimate layers of the frozen weak models, as shown in the following set of Equations 3.39.

$$\begin{aligned} Feat_{concat} &= [Feat_{weak_1}, Feat_{weak_2}, \dots, Feat_{weak_N}] \\ Feat_{comb} &= W \cdot Feat_{concat} + b \end{aligned} \tag{3.39}$$

- $Feat_{weak_i}$ : The feature vector extracted by the  $i$ -th weak learner (EfficientNet-b0 feats specifically);
- $Feat_{concat}$ : The vector resulting from the concatenation of all weak learners' feature vectors;
- $W$  and  $b$ : The trainable weights and bias of the combination layer (which is fine-tuned while the weak learners are frozen).

- $Feat_{comb}$ : The final combined feature vector used for classification.

A trainable linear combination layer is then fine-tuned on top of these concatenated features. This architecture eliminates the redundancy of retaining multiple output classification heads, allowing the model to learn a weighted, non-linear integration of the features provided by the specialised learners. The resulting system works similarly to a fully differentiable version of Gradient Boosting, yet performs ensemble integration at the feature level rather than via decision-tree traversal. Experimental validation across diverse benchmark datasets, including CIFAR-10 [73], CIFAR-100 [73], and Food-101 [10], demonstrates that this method outperforms State-of-the-Art (SOTA) solutions by an average of 0.5% in accuracy while drastically reducing computational overhead. The efficiency gains are substantial: the proposed ensemble requires 5 to 60 times fewer parameters and reduces Floating Point Operations Per Second (FLOPs) by 10 to 100 times compared to competing SOTA architectures. Furthermore, the training process is highly optimised; because the weak learners are independent, they can be trained in parallel, and the final fine-tuning phase involves only a fraction of the total parameters (approximately 100,000), ensuring rapid convergence. This methodology confirms that sustainable "Green AI" paradigms can achieve superior performance without the carbon footprint of massive, monolithic DL models.

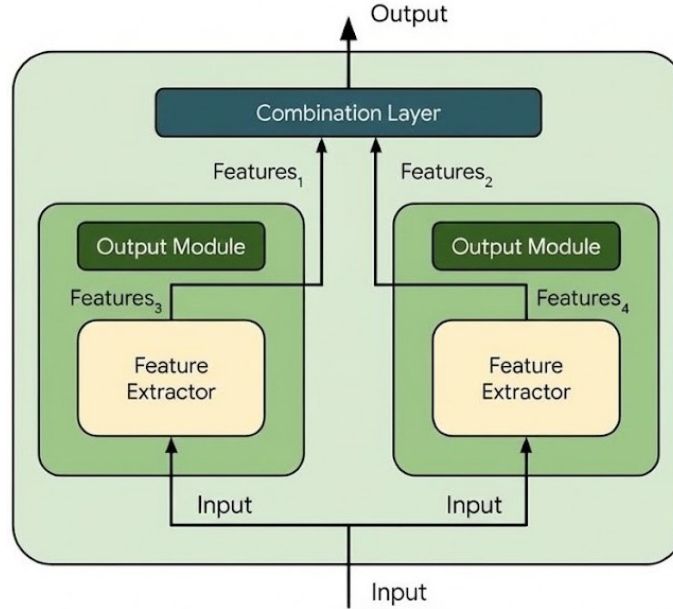


Figure 3.19: The ensemble complete pipeline, after the training of the weak model, it is frozen and ensembles the feature extraction model of the best two. The final linear combination layer is then trained on the same task as the previous one, yielding a more accurate, feature-aware classification output.

The overall validation pipeline, as illustrated in the Fig. 3.19.

### 3.2.7 Training and Validation Process

The two-stage nature of the pipeline translates into a highly efficient training and validation process that fundamentally reimagines the relationship between model complexity and computational cost. The initial training phase, while robust, departs from traditional ensembling dogma by strictly partitioning the dataset into disjoint subsets to train the weak learners. This "bagging" strategy ensures that each weak learner, specifically the EfficientNet-b0 architecture chosen for its optimal accuracy-to-complexity ratio, develops a unique perspective on the data. In a move that might appear counterintuitive to the orthodox practitioner, these weak learners are deliberately trained to the point of overfitting. While conventional wisdom holds that overfitting degrades generalisation, within this adaptive ensemble framework, it serves a distinct purpose: highly specialised weak learners provide high-variance feature representations that the subsequent combination layer can synthesise more effectively than generalised, homogeneous models can. This phase utilises the AdaBelief optimiser, selected for its ability to secure fast convergence and generalisation, alongside a standard cross-entropy loss function appropriate for multiclass tasks.

Following the creation of these specialised feature extractors, the methodology transitions to the ensemble fine-tuning stage, where the design's actual efficiency becomes apparent. Rather than aggregating final classification outputs via static voting or averaging, the system freezes the weak learners' feature extractors and discards their output modules. The ensemble is then constructed by concatenating the feature vectors and passing the resulting concatenation through a trainable linear combination layer. This architectural shift entails that during the fine-tuning phase, the number of trainable parameters is drastically reduced to approximately 100,000, a negligible number compared to the millions found in standard DL architectures. Consequently, the computational burden of backpropagation is virtually eliminated for the vast majority of the network, as gradient calculation halts at the frozen layers. This reduction in memory footprint allows a significant increase in batch size, from 55 during initial training to 200 during fine-tuning, further accelerating convergence of the combination layer. The implications of this design extend well beyond the training phase, offering substantial gains during inference that align with sustainable "Green AI" paradigms. By utilising independent weak learners, the architecture naturally supports parallelisation; the feature extraction processes for the component models can be executed simultaneously on modern hardware. As a result, the time required to generate the concatenated feature vector is roughly equivalent to the inference time of a single weak learner. The total computational cost, therefore, approximates the sum of a single forward pass and the negligible operation of the linear combination layer. Empirically, this results in a system that requires 10 to 100 times fewer Floating Point Operations Per Second (FLOPs) than competing State-of-the-Art (SOTA) models while consistently achieving superior accuracy across diverse benchmarks such as CIFAR-100 and Food-101. For instance, on the CINIC-10 dataset, the proposed method achieves superior accuracy with approximately 0.9 Giga-FLOPs, whereas the most efficient competing SOTA model requires at least 10 GFLOPs, and others demand significantly more. This methodology effectively decouples high performance from exponential complexity,

proving that an adaptive, feature-level ensemble can serve as a potent, low-carbon alternative to monolithic DL models.

### 3.3 Case study: LUS Classification

The actual test of any novel methodology lies in its application to a challenging, real-world problem. This chapter examines the deployment of the minimal adaptive ensembling strategy on the target task of classifying lung pathologies from ultrasound images. This case study not only demonstrates the model’s state-of-the-art performance but also delves into its interpretability, showcasing how Explainable AI is a first critical step that can bridge the gap between algorithmic prediction and clinical trust.

The strategic development of advanced diagnostic tools for pulmonary diseases has become a paramount objective in global health, a reality underscored by the recent SARS-CoV-2 pandemic. In emergency and point-of-care scenarios, the ability to perform rapid, safe, and accessible diagnostics is not merely an operational advantage but a critical component of effective patient management and public health response. This need has catalysed research into imaging modalities that can be deployed at the patient’s bedside, minimising delays and risks associated with transporting critically ill patients to specialised radiology suites. The assessment of pulmonary disease has traditionally relied on well-established imaging modalities, each with distinct clinical trade-offs. POCUS has emerged as a powerful alternative, particularly in high-pressure environments. The following Table 3.2 presents a comparison of the more powerful image modalities, CT and MRI, with POCUS use.

Table 3.2: Comparison of POCUS and CT/MRI Modalities

| Imaging Modality | Key Characteristics & Clinical Trade-offs                                                                                                                                                                                                                                                                                                                                                                      |
|------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| POCUS            | Pros: Highly portable, safe (no ionising radiation), inexpensive, repeatable, and quick. It is a cost-effective tool ideal for bedside examinations and continuous monitoring.<br>Cons: Generated images are often noisy, and their quality is highly dependent on operator experience. Interpretation can be less straightforward compared to other modalities.                                               |
| CT/MRI           | Pros: Considered the gold standard due to high image quality and exceptional diagnostic power. Provides detailed anatomical information that is crucial for definitive diagnosis.<br>Cons: Involves significant ionising radiation (a chest CT delivers ~7 mSv), is more time-consuming, substantially more expensive, and requires stationary equipment, making it impractical for rapid, bedside deployment. |

Artificial Intelligence, and DL specifically, offers a compelling solution to mitigate the primary disadvantages of POCUS. By developing models capable of accurately and reliably interpreting ultrasound images, it is possible to reduce the dependency on expert operators and standardise the diagnostic process. This enhancement of

clinical utility is the central motivation for the present study, which leverages a large, publicly available LUS dataset to develop and validate a new classification model. The methodological framework for this thesis was designed to ensure robust, reproducible, and clinically relevant results. It introduces in detail the benchmark datasets used for training and validation, the specific data preprocessing and augmentation pipeline employed to enhance model generalisation, the novel ensemble architecture at the core of the classification system, and the rigorous validation protocol used to assess its performance.

### 3.4 Methodological Framework: The Efficient Adaptive Ensemble

At the heart of both experiments lies the EfficientNet-b0 architecture, section 3.2.2, selected for its principled "compound scaling" method that balances network depth, width, and resolution to maximise feature extraction capabilities with minimal computational overhead. However, the true innovation lies not in the base model itself, but in the Efficient Adaptive Ensembling strategy that enhances it. To systematically assess architectural efficacy within this specific image domain, preliminary experiments were conducted using established baseline architectures. Although these models precede current large-scale foundational systems, they serve as critical benchmarks that reflect the rapid evolution of DL and establish a necessary performance floor. This experimental suite functions as a practical ablation study to validate the final methodology. The analysis is divided into two phases: an initial benchmarking phase conducted during the thesis proposal stage, and a subsequent evaluation of emergent, state-of-the-art architectures that emerged during the final stages of the research 4.6. The architectures evaluated in this study include VGG-16, ResNet, and GoogLeNet. While the structural nuances of VGG-16 and ResNet were detailed in Section 3.2.2, GoogLeNet (also known as Inception-v1) requires further technical introduction [49] due to its departure from traditional sequential stacking. The introduction of the Inception module distinguishes GoogLeNet's architecture. This multi-level feature extractor performs parallel convolutional operations with varying kernel sizes ( $1 \times 1$ ,  $3 \times 3$ , and  $5 \times 5$ ) within the same layer. This design enables the network to capture spatial information at multiple scales while utilising  $1 \times 1$  convolutions for dimensionality reduction, thereby maintaining computational efficiency. Furthermore, GoogLeNet utilises Global Average Pooling instead of fully connected layers at the final stage to reduce the total parameter count and mitigate overfitting, and auxiliary classifiers during the training phase to address the vanishing gradient problem inherent to its 22-layer depth.

Departing from the brute-force approach of training massive monolithic networks, the definitive protocol begins by training multiple independent weak learners. These learners are trained on disjoint subsets of data (bagging) and are deliberately pushed towards overfitting, as described in section 3.2.6. While conventional wisdom suggests overfitting degrades generalisation, in this context, it forces each weak learner to develop a highly specialised, high-variance perspective on the data. The construction of the final "strong" learner is an exercise in surgical efficiency. Rather than

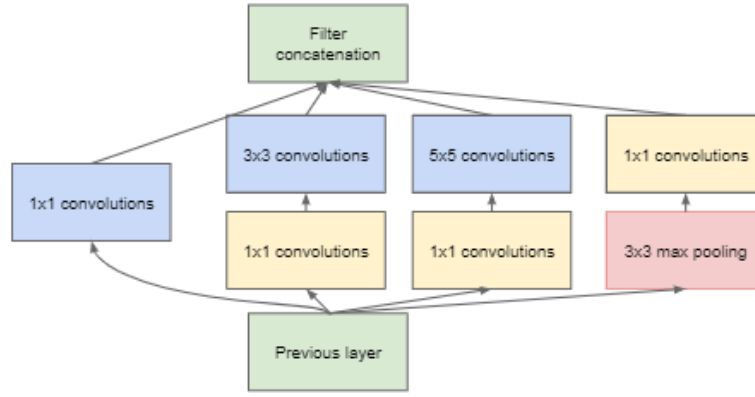


Figure 3.20: Architectural composition of the inception module at the base of GoogLeNet, the naive version is the same but without the dimensional operation block on a convolutional kernel bigger than 1x1.

averaging the final predictions of these specialised models, the architecture discards their classification heads, freezes their feature extraction layers, and concatenates their deep feature vectors. A single, lightweight linear combination layer is then introduced and fine-tuned to synthesise these features. This process is computationally inexpensive yet highly effective, allowing the model to learn the optimal non-linear combination of the specialised representations. The result is a system that achieves the robustness of a large ensemble with inference speed comparable to that of a single small network, effectively decoupling high performance from high computational cost.

### 3.5 Experiment 1: DL Model Validation for POCUS

To establish a rigorous baseline for the LUS classification task, a series of preliminary experiments were conducted using a suite of established Deep Convolutional Neural Network (DCNN) architectures. This stage served as a critical feasibility study to determine how standard feature extractors, initially designed for natural image datasets like ImageNet, respond to the unique speckle noise, varying gain levels, and transient artifacts inherent in pulmonary ultrasound imagery. The models selected for this benchmarking phase included VGG-24, ResNet-256, and GoogLeNet, alongside a basic Vanilla CNN architecture. These models represent different architectural philosophies: from the deep sequential stacking of the VGG family to the residual skip-connections of ResNet and the multi-scale parallel convolutions of the Inception modules in GoogLeNet. The training protocol was standardised across all architectures to ensure an objective comparison of their representational capacities. Images were resized to a resolution of 224 x 244 pixels. Each model was trained for a maximum of 45 epochs using a categorical cross-entropy loss function. However, to prevent deleterious overfitting and minimize unnecessary computational expenditure, an Early Stopping mechanism was implemented. This mechanism monitored the validation loss, terminating the training process if no significant improvement was observed for five consecutive epochs, effectively "freezing" the models at their

most generalizable state. The results of this preliminary phase, summarized in Table 3.3, revealed that while classical models are capable of extracting meaningful features from LUS data, they exhibit a clear performance ceiling and varying degrees of sensitivity to the data distribution. As observed, ResNet-256 achieved the highest

Table 3.3: Preliminary Benchmarking Performance of Classical Architectures on LUS Data

| Architecture | Final Test Accuracy |
|--------------|---------------------|
| Vanilla CNN  | 0.85                |
| VGG-24       | 0.94                |
| ResNet-256   | 0.96                |
| GoogLeNet    | 0.93                |

baseline performance at 0.96, likely due to its residual learning framework, which mitigates the vanishing gradient problem in deep structures. Conversely, the Vanilla CNN plateaued at 0.85, demonstrating that shallower, non-specialised architectures struggle to capture the complex, non-linear relationships between ultrasound artefacts (such as B-lines) and pathological states.

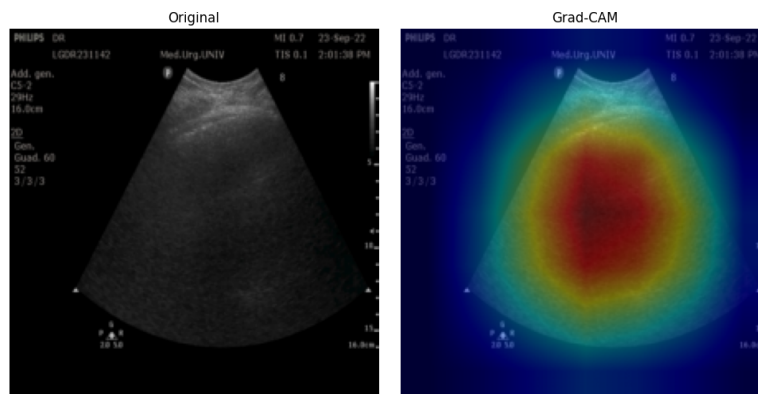


Figure 3.21: An example of correct classification of the trained ResNet-256 network; even if the accuracy is high, the model struggles to focus on meaningful detail and depict an area of interest.

These findings provided empirical justification for moving toward the more sophisticated EfficientNet-b0 based Adaptive Ensemble, as the traditional models either lacked the necessary precision or, in the case of ResNet-256, required a parameter count exceeding the requirements for efficient point-of-care deployment, other than struggle to focus on meaningful details, Fig. 3.21.

## 3.6 Experimental Validation II: Pathology Classification

This experiment was conducted using the largest publicly available LUS dataset [9], a benchmark resource curated and reviewed by medical experts for classifying

pulmonary conditions. Its scale and diversity make it an ideal test for developing and validating advanced machine learning models.

- Total Content: 261 ultrasound videos and images
- Patient Cohort: 216 distinct patients
- Data Sources: 41 different sources

To facilitate a frame-based classification task, the videos were decomposed into individual images, yielding a substantial training dataset. The distribution of frames across the three primary pathology classes is detailed in Table 3.4.

Table 3.4: Dataset Class Distribution

| Class     | Number of Frames |
|-----------|------------------|
| COVID-19  | 1024             |
| Pneumonia | 704              |
| Healthy   | 1326             |

Prior to model training, all images underwent a standardized preprocessing pipeline. The input images were first resized to a consistent dimension of  $512 \times 512$  pixels. The pixel values were then normalized and standardized to ensure the data belonged to a distribution with a mean near zero and unit standard deviation, a practice that improves the stability and convergence of the training process.

To prevent overfitting and improve the model’s ability to generalize to unseen data, a series of data augmentation techniques were applied. These transformations introduce variability into the training set without altering the core diagnostic content of the images:

1. Rotation: Images are subjected to random rotation of up to 20 degrees in either direction, simulating variations in probe orientation;
2. Random Affine Translation: Images are translated vertically or horizontally by up to 10% of the image’s height, mimicking slight shifts in probe placement;
3. Random Shearing: A shearing transformation is applied, skewing the image by up to 10% to introduce further geometric variability;
4. Random Scaling: The image size is randomly adjusted to between 80% and 120% of its original size before being resized back to the standard  $512 \times 512$  dimensions, simulating changes in imaging depth and zoom.

### 3.6.1 Performance and State-of-the-Art Comparison

The results of this experiment were unequivocal. Across a rigorous stratified 5-fold cross-validation protocol carefully designed to prevent data leakage by keeping patient videos intact within folds, the adaptive ensemble achieved a perfect 100% accuracy on the test set in every run. When benchmarked against previously published



results on this dataset, the proposed ensemble model establishes a new state-of-the-art, demonstrating clear superiority in both predictive accuracy and computational efficiency, as detailed in the comparative table 3.5.

Table 3.5: Comparative Analysis with SOTA

| Model             | Accuracy (%) | # Parameters (M) | FLOPs (G) |
|-------------------|--------------|------------------|-----------|
| POCOVID-Net       | 82.1         | 14.7             | 30.7      |
| VGG-16            | 87.8         | 14.7             | 15.3      |
| Light Transformer | 93.4         | 0.3              | ~ 1.0     |
| Weak model (our)  | 98.7         | 5                | 0.39      |
| Ensemble (our)    | 100          | 10               | 0.78      |

The proposed ensemble model is unequivocally the new state-of-the-art in predictive performance, achieving 100% accuracy and surpassing the previous SOTA (Light Transformer) by a significant margin of 6.6 percentage points. Critically, this perfect accuracy is achieved within a highly efficient computational framework. While the Light Transformer model is smaller in parameter count, our ensemble’s computational cost (0.78G FLOPs) is orders of magnitude lower than older, less accurate architectures such as VGG-16 (15.3G FLOPs) and POCOD-Net (30.7G FLOPs). The profound efficiency of our underlying architecture is further evidenced by the Weak model’s performance. A single EfficientNet-b0 instance, with just 5M parameters and 0.39G FLOPs, achieves 98.7% accuracy, outperforming nearly all previous SOTA models on its own while being exceptionally lightweight. This combination of best-in-class accuracy and low computational cost makes the model an exceptionally strong candidate for integration into real-time, point-of-care diagnostic systems where both speed and precision are paramount.

### 3.6.2 Explainability and Clinical Plausibility

To ensure the model’s decisions were grounded in clinical reality, we employed Grad-CAM [115]. The analysis revealed distinct, clinically plausible patterns: the model consistently focused on the upper lung field for COVID-19, the lower field for bacterial pneumonia, and the clear, air-filled regions for healthy cases, Fig. 3.22.

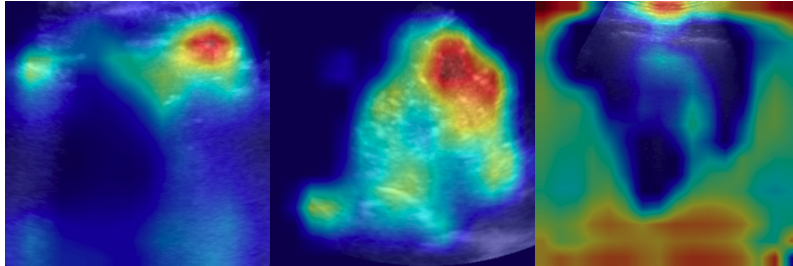


Figure 3.22: Healthy (first column), Pneumonia (second column), COVID (third column), Healthy samples from the dataset and their saliency map. Red (higher) to blue (lower) scale.

Crucially, when comparing the perfect ensemble against the slightly less accurate weak learners, Grad-CAM revealed that the errors in the latter were driven by focus on irrelevant artefacts or acoustic shadows, Fig. 3.23.

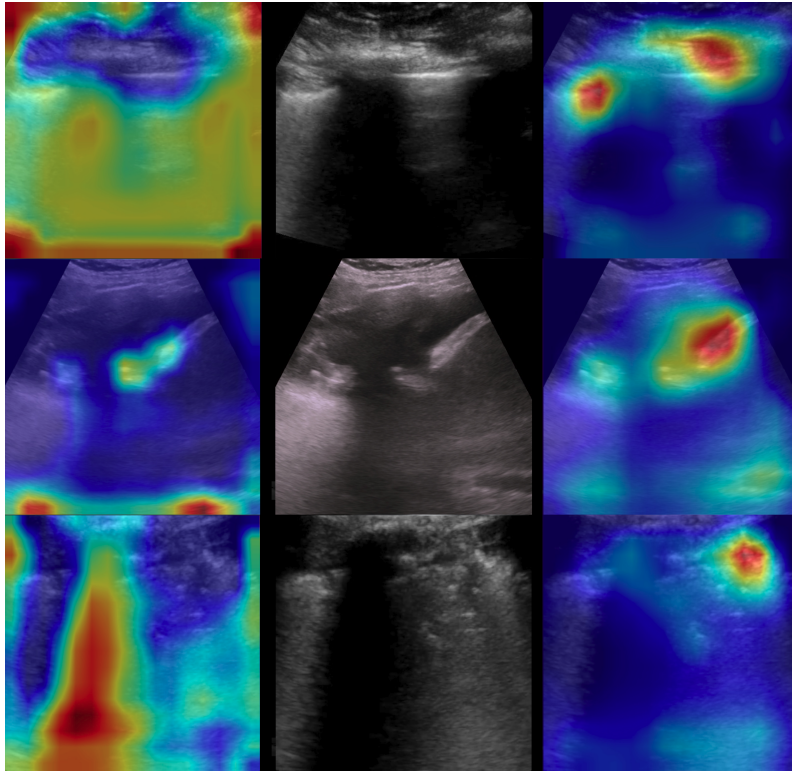


Figure 3.23: Classifications of a LUS image: in the middle, the input image; on the left, the focus of a wrong classification; on the right, the focus of the correct classification obtained by the classifier achieving 100% accuracy. Red (higher) to blue (lower) scale; (real Class  $\rightarrow$  error Class): in order Covid  $\rightarrow$  Pneumonia, Pneumonia  $\rightarrow$  Covid, Healthy  $\rightarrow$  Covid

The ensembling process effectively filtered out this noise, guiding the final model toward robust, pathologically relevant features.

### 3.7 Experimental Validation III: B-Lines

The second experiment shifted focus from broad pathology classification to quantifying B-lines, a hallmark of interstitial lung disease. The utilised dataset comprises 30 LUS videos from 30 distinct subjects enrolled in the TiAssisto project. To create a working dataset for the classification model, individual images were extracted from these videos. The frames were subsequently labelled based on the clinical context provided with each video, which gave insight into the presence and severity of B-lines. This labelling process, agreed upon with the project's physicians, quantifies the extent of the ultrasound field of view (FOV) occupied by these comet-tail artefacts. The final distribution of images across the four defined labels is summarised in Table 3.6.

Table 3.6: Dataset Label Distribution

| Label Description                            | Number of Images |
|----------------------------------------------|------------------|
| Label 0: No B-Lines                          | 183              |
| Label 1: B-Lines occupying ~30% FOV          | 754              |
| Label 2: B-Lines occupying ~50% FOV          | 196              |
| Label 3: B-Lines occupying more than 50% FOV | 76               |
| Total                                        | 1209             |

This unique dataset provides the foundation for the rigorous methods employed to process and analyse the images for pathological signs.

### 3.7.1 A Four-Stage Methodological Pipeline

A comprehensive four-stage pipeline was developed to process the LUS data. This structured approach was designed to ensure patient privacy, model robustness, high classification performance, and ultimately, clinical trustworthiness. The four stages are: (1) Data Anonymisation and Preparation, (2) Image Augmentation, (3) The Adaptive Ensemble Classification Model, and (4) Explainability via Post-Hoc Analysis.

### 3.7.2 Stage 1: Data Preparation and Privacy-Preserving Anonymisation

Initial data handling focused on preparing the images for analysis while strictly adhering to privacy requirements:

- **Annotation:** A custom-implemented Java tool was developed to enable medical staff to perform frame-by-frame video processing and annotation across the entire dataset;
- **Anonymisation:** to ensure patient privacy, a critical step of automated text cleaning was performed on all images. This process utilises an OCR-based method to detect and mask sensitive information. It employs a Convolutional Recurrent Neural Network (CRNN) with a bidirectional Long Short-Term Memory (LSTM) architecture to identify text bounding boxes in ultrasound images. These detected areas are subsequently masked using an inpainting method. This process is carefully designed to operate only outside the primary diagnostic field of view, thereby avoiding any alteration of the core clinical content. The autonomous anonymisation process took approximately 45 minutes for the entire dataset, with an average processing time of 2 seconds per image, demonstrating its feasibility for real-world POCUS applications.

### 3.7.3 Stage 2: Image Augmentation

To enhance the model’s generalisation and prevent overfitting, a sequence of data augmentation techniques was applied to the images before training. This sequence simulates the natural variability in probe handling, rotation, translation, and pressure-induced shearing, making the model more robust to real-world acquisition differences.

1. Resizing: all input images are resized to a consistent dimension of  $512 \times 512$  pixels;
2. Rotation: images are subjected to a random rotation of up to 20 degrees in either direction;
3. Random Affine Translation: the image is moved slightly, with random vertical or horizontal translation up to 10% of the image’s height;
4. Random Shearing: the image is skewed by up to 10% to introduce further distortion;
5. Random Scaling: the image size is randomly adjusted to between 80% and 120% of its current size, after which it is resized back to the standard  $512 \times 512$  pixels;
6. Normalisation : after all geometric transformations, pixel values are converted to tensors and scaled to a  $[0, 1]$  range. They are then normalised using the mean and standard deviation calculated for each of the three colour channels across the entire dataset.

#### Privacy-Preserving Preprocessing

Given the sensitive nature of this clinical data, this experiment necessitated a more complex preprocessing pipeline. Before any analysis could begin, a custom OCR pipeline utilising a CRNN with bi-directional LSTMs was deployed. This system autonomously detected and masked patient metadata embedded in the ultrasound frames via inpainting, processing the entire dataset in approximately 45 minutes. This successful anonymisation step highlights the feasibility of integrating privacy-preserving mechanisms directly into the AI workflow without disrupting clinical throughput.

### 3.7.4 Stage 3-4: Results

The challenge of severity quantification proved to be a more rigorous test of the model’s capabilities than simple classification. Here, the value of the ensembling strategy became starkly apparent. The individual weak learners struggled to reach a clinical-grade benchmark, plateauing at approximately 91-92% test accuracy. However, the adaptive ensemble bridged this gap, consistently achieving near-perfect scores with an average accuracy of 99% and perfect precision and recall across all four severity classes (see Table 3.7).

Table 3.7: Performance Disparity: Weak Learners vs. Adaptive Ensemble (TiAssisto Dataset)

| Run | Weak Learner Performance |      |       | Ensemble Performance |      |       |
|-----|--------------------------|------|-------|----------------------|------|-------|
|     | Test                     | Val  | Train | Test                 | Val  | Train |
| 1   | 0.92                     | 0.89 | 1.00  | 1.00                 | 0.99 | 0.99  |
| 2   | 0.91                     | 0.90 | 1.00  | 1.00                 | 0.99 | 0.99  |
| 3   | 0.91                     | 0.90 | 1.00  | 1.00                 | 0.99 | 0.99  |

This disparity highlights the specific utility of the adaptive ensemble in tasks with subtle inter-class boundaries, such as distinguishing between 30% and 50% B-line occupancy. The Grad-CAM analysis further corroborated this, showing that while the model perfectly segmented B-lines in clear-cut cases, it occasionally produced diffuse activations in the intermediate Label 2 category, suggesting a specific target for future data augmentation and refinement.

The detailed classification report for the best-performing ensemble model further underscores its capabilities. The model achieved perfect scores (1.0) across all key metrics, Precision, Recall, and F1-Score, for every class in the test set, culminating in an overall accuracy of 100% (Table 3.8).

Table 3.8: Detailed Classification Report (Best Ensemble Model)

| Class          | Precision | Recall | F1-Score | Support |
|----------------|-----------|--------|----------|---------|
| 0 (No B-Lines) | 1.0       | 1.0    | 1.0      | 19      |
| 1 (~30% FOV)   | 1.0       | 1.0    | 1.0      | 76      |
| 2 (~50% FOV)   | 1.0       | 1.0    | 1.0      | 20      |
| 3 (>50% FOV)   | 1.0       | 1.0    | 1.0      | 7       |
| Accuracy       |           |        | 1.0      | 122     |
| Macro avg      | 1.0       | 1.0    | 1.0      | 122     |
| Weighted avg   | 1.0       | 1.0    | 1.0      | 122     |

3.7.5 Grad-CAM usage to assess model viability

The qualitative analysis using Grad-CAM provides crucial context to the quantitative results, offering insight into the model’s reasoning and its alignment with clinical expertise, as shown in the following Fig. 3.24 and 3.25. The models focus their decision in a defined area with relevant content, in this case, B-Lines.

The activation maps generated via Grad-CAM provided several critical insights:

- **Physician Feedback:** overall, the team of physicians involved in the study was satisfied with the system’s initial version, agreeing with a significant portion of the visual explanations it provided,

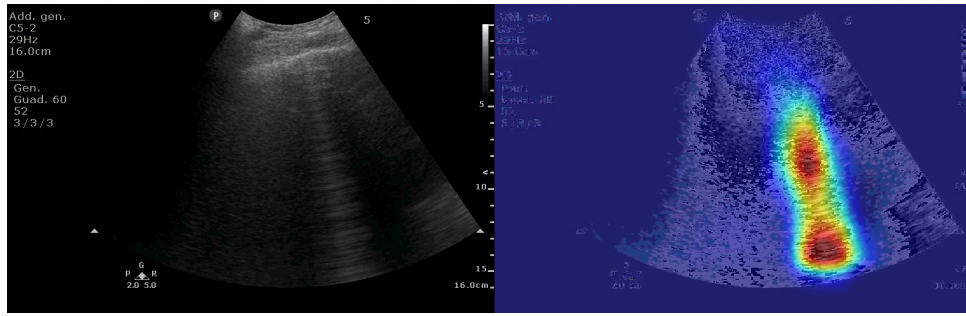


Figure 3.24: A LUS correctly classified from the class 1, the single B-Lines is detected by the model and almost entirely segmented

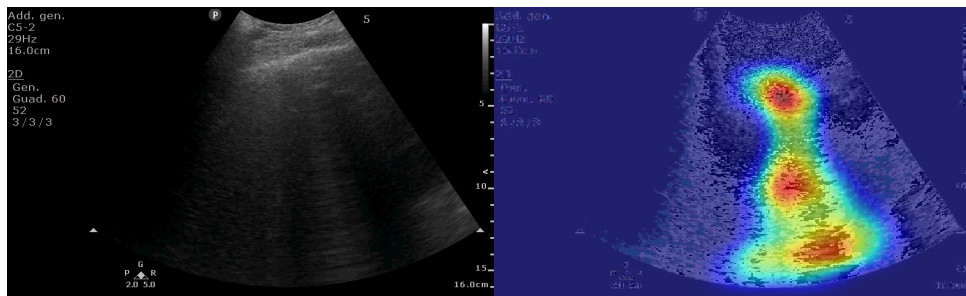


Figure 3.25: A LUS correctly classified from the class 2, the B-Lines are detected by the model and almost entirely segmented, in the wider area where they can be spotted.

- **Successful Detections:** the model demonstrated a strong ability to correctly identify meaningful anatomical details for specific classes. For images corresponding to Label 1 ( $\sim 30\%$  FOV) and Label 3 ( $>50\%$  FOV), the Grad-CAM visualisations showed that the system correctly detected and almost perfectly segmented the B-lines, confirming it was focusing on the correct clinical signs.
- **Identified Weaknesses:** the analysis also revealed one limitation. The model appeared to struggle to correctly identify images corresponding to Label 2 ( $\sim 50\%$  FOV), in some cases, when the B-Lines are separated, the model struggles to narrow the focus on the affected zone, generalising on the small possible affected area, which, though, also includes a healthy lung.
- **Robustness to Annotation Errors:** a notable case of a mislabelled image provided evidence of the model's robustness. An image classified as Label 3 was reviewed; however, the Grad-CAM visualisation, corroborated by a physician, showed no actual comet tails. Critically, the model's feature activations were correctly focused on regions consistent with Label 0 (no B-lines), suggesting an ability to overcome certain types of label noise in the dataset.

These results transition the evaluation from the specific findings to a broader discussion of their implications and the study's overall context.

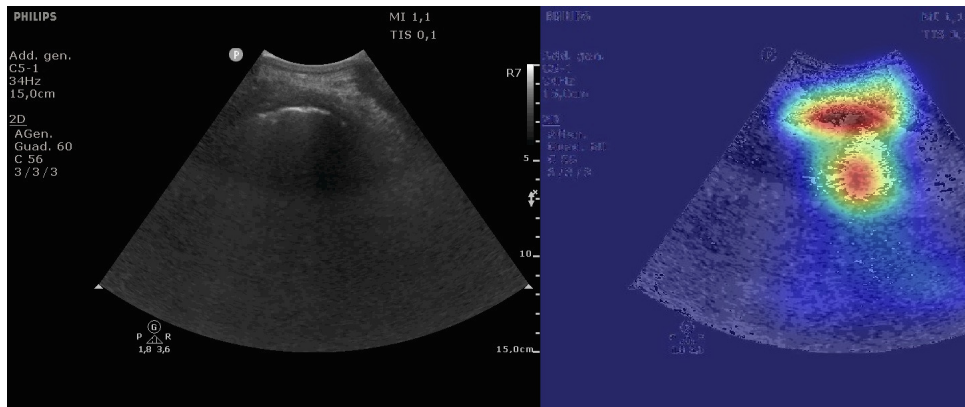


Figure 3.26: This image was originally mislabelled. The original label II was erroneous since there is no presence of B-Lines; the ensemble model at first misclassified this as a label 0, correcting the mistake.

### 3.8 LUS classification final pipeline

This studies presents the result of the development and evaluation of Efficient Adaptive Ensemble methodology, with the primary objective of improving diagnostic accuracy and computational efficiency for DL applications in POCUS settings. By assessing both pathology classification and B-line quantification tasks, the study demonstrates the methodologies transformative potential. Specifically, the proposed approach seeks to balance computational overhead with diagnostic precision, establishing a robust and practical framework suitable for resource-constrained clinical environments. The analysis further underscores the need for medical AI systems to be accurate, sustainable, and transparent to achieve meaningful clinical adoption. The methodology demonstrated a remarkable capacity to generalise across diverse datasets, achieving state-of-the-art (SOTA) performance on two distinct yet complementary diagnostic tasks. The transition from individual "weak" learners to the "strong" ensemble was the pivotal factor in bridging the performance gap. While individual EfficientNet-b0 models showed high sensitivity, they were often prone to shortcut learning, a phenomenon in which a model achieves apparently good performance by exploiting superficial cues or device-specific artefacts, such as digital overlays or gain-related noise, rather than identifying actual anatomical pathology. By implementing a trainable combination layer that operates on deep feature concatenations, the system effectively filtered these spurious correlations. This allowed the ensemble to leverage the high-variance features of overfitted specialised learners to synthesise a decision boundary that is more robust and fundamentally grounded in the physical reality of ultrasound propagation.

Furthermore, the "Green AI" paradigm used in this study challenges the prevailing industry trend, in which minor accuracy gains are bought at the cost of exponential increases in model complexity. By prioritising feature-level integration over massive parameter stacking, the system achieved 100% pathology classification accuracy and 100% B-line FOV classification, with a computational footprint of only 0.78 GFLOPs. This represents a reduction in computational demand of nearly 40-fold compared to foundational models, such as transformers, which, despite their high



performance, lack the efficiency required for low-end hardware. This proving ground suggests that architectural intelligence can effectively replace raw scale, yielding models that are both more accurate and environmentally sustainable.

A cornerstone of this research was the deliberate use of XAI not merely as a retrospective visualisation tool, but as a critical diagnostic instrument for the model itself. The application of Gradient-weighted Class Activation Mapping (Grad-CAM) provided an essential bridge between algorithmic logic and clinical intuition. In classifying COVID-19 and bacterial pneumonia, the saliency maps revealed that the models focus was consistently aligned with the pleural irregularities and lower-lobe consolidations that sonographers use for bedside diagnosis.

This transparency also helped with error correction and improving the dataset. During the B-line quantification study, Grad-CAM revealed cases of inconsistent human annotation. In some outlier cases, the model outperformed its training labels, correctly flagging frames as artefact-free even though reviewers had labelled them as pathological. This suggests the pipeline can act as an automatic quality-control tool. It can flag possible mislabelling in medical datasets and help set a higher standard for future "ground truth." By explaining its predictions, XAI turns the model from a "black box" into a diagnostic partner.

The final pipeline presented in this thesis, integrating Anatomical Isolation, Privacy-Preserving Method, and Adaptive Ensembling, represents a comprehensive end-to-end solution for medical imaging analysis. It respects the unique physical principles of ultrasound, such as the diagnostic utility of comet-tail artefacts, while adhering to modern standards of data privacy and computational efficiency. Looking ahead, several research directions remain critical to the evolution of this technology. However, deploying such AI-driven systems in real-world clinical settings

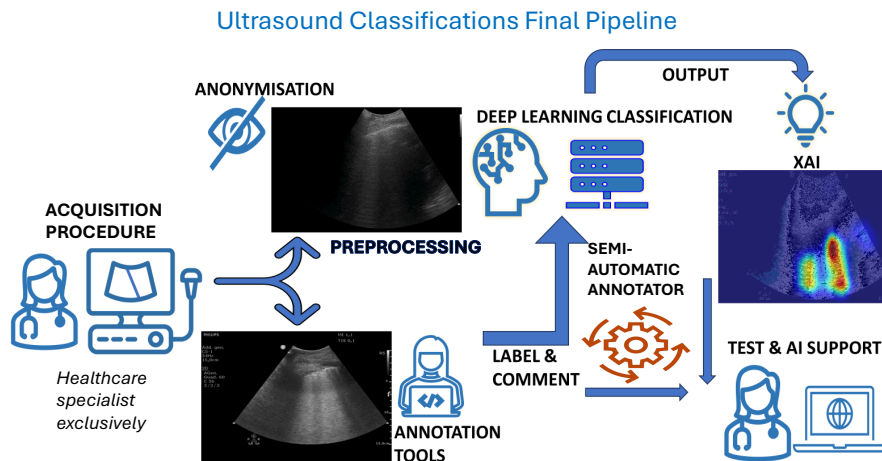


Figure 3.27: Final proposed pipeline for POCUS analysis

poses several challenges that must be addressed for successful translation. These challenges include robust validation across heterogeneous devices and patient populations, integration with existing clinical workflows, and consistent performance under variable imaging conditions. The integration of temporal context is a logical next step; moving from static, frame-based classification to 3D Convolutional



or Transformer-based video analysis will enable the system to capture the pleural line's dynamic shift, a key indicator for ruling out conditions such as pneumothorax. Additionally, the ensemble's lightweight nature makes it a prime candidate for Federated Learning, enabling models to be trained across multiple hospital nodes without ever transferring sensitive patient data. Finally, the ultimate goal remains the direct deployment of these algorithms on handheld POCUS hardware, enabling real-time Edge AI feedback to assist clinicians in the most critical moments of patient care. Ultimately, this work demonstrates that when AI is designed with an acute awareness of the physical and clinical context of its data, it evolves into an indispensable, interpretable, and safe partner in the clinical diagnostic workflow. In the next part of the thesis, a new, mathematically refined approach will be proposed to help the network recognise the necessary pattern to guide the decision towards a reliable outcome.



## Part II

Reliable method for deep learning model  
analysis



# Chapter 4

## Loss function and target classification output

The material presented in this chapter is based on the conference papers and preprints:

- G. Ignesti, M. Martinelli, and D. Moroni. Youve got the wrong number: evaluating deep learning training paradigms using handwritten digit recognition data. *Pattern Recognition and Image Analysis*, 34(4):978–983, 2024,
- G. Ignesti, M. Martinelli, and D. Moroni. Youve got the wrong outfit: Evaluating deep learning paradigms on digit and fashion recognition. In *International Conference on Pattern Recognition*, pages 202–215. Springer, 2024,
- G. Ignesti, D. Moroni, and M. Martinelli. Batch-cam: Introduction to better reasoning in convolutional deep learning models. *arXiv preprint arXiv:2510.00664*, 2025.

The rapid development of Artificial Intelligence (AI) has precipitated a crisis of trust, fracturing the field of AI safety into three distinct epistemological schools: XAI, Trustworthy AI, and Reliable AI. While these schools share the ultimate goal of safe deployment, they represent fundamentally competing mathematical objectives. This divergence stems from the transition from linear statistical methods to high-dimensional DNNs, which has rendered internal model states opaque [40]. The school of XAI posits that trust is a function of transparency, rejecting the black box nature of DL models. In contrast, the school of Trustworthy AI extends beyond mere explanation to a holistic set of requirements, including fairness, privacy, and accountability, often optimising for human alignment rather than raw computational faithfulness [83]. Finally, the school of Reliable AI prioritises consistency and worst-case robustness, arguing that understanding the internal mechanics is secondary to ensuring the system never fails catastrophically, a perspective now being rigorously tested in novel computing paradigms such as Spiking Neural Networks (SNNs) and analogue hardware [77].

## 4.1 eXplainable AI (XAI)

XAI can be identified as the school of transparency, formalised as a distinct engineering program by Gunning et al. They argue that high-stakes deployment requires an accurate proxy of the internal state. The central mechanism for this is often Feature Attribution, best represented by the Shapley Value, which attempts to fairly distribute the prediction output among input features. Mathematically, for a set of features  $N$  and a value function  $v$ , the attribution  $\phi_i$  for feature  $i$  is calculated as the weighted average of marginal contributions across all possible coalitions  $S$ :

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \quad (4.1)$$

The defining quirk of this school is the Accuracy-Interpretability Frontier, which suggests that:

$$\frac{\partial \mathcal{A}}{\partial \mathcal{I}} < 0 \quad (4.2)$$

Which means that a fully faithful explanation of a non-linear system is often as complex and unintelligible as the system itself, as in Equations. 4.1-4.2. This trade-off forces a choice between "post-hoc" explanations (which are intelligible but approximate) and "intrinsic" interpretability (which restricts model complexity) [7].

## 4.2 Trustworthy AI

Moving beyond the specific goal of transparency, we can find the so-called school of Human Alignment; Trustworthy AI focuses on a broader set of ethical and legal requirements. As defined in recent comprehensive surveys, Trustworthy AI is not just about explaining a decision, but about ensuring the entire lifecycle of the AI system, from data collection to deployment, adheres to principles of fairness, privacy, robustness, and accountability [83]. This school operates on the premise that trust is multi-dimensional. A model might be explainable as we know why it discriminated, but untrustworthy because it discriminates.

Mathematical formalisms in this school often involve constrained optimisation where the objective function includes penalties for violating human norms (e.g., fairness metrics). If  $L(f)$  is the standard loss function, Eq 4.3, Trustworthy AI optimises:

$$\min_{\theta} L(f_{\theta}) + \lambda_1 \mathcal{D}_{fair}(f_{\theta}) + \lambda_2 \mathcal{D}_{priv}(f_{\theta}) \quad \text{s.t.} \quad \text{Explanation}(f_{\theta}) \approx \text{Human Prior} \quad (4.3)$$

where  $\mathcal{D}_{fair}$  measures demographic parity or equal opportunity, and  $\mathcal{D}_{priv}$  ensures differential privacy. A defining quirk here is the Deception Trap (originally termed in the context of Plausible AI [61]), where the system may optimise for explanations that are convincing to a human user (high trust) but unfaithful to the model's true computation. For instance, a rationalisation generator might produce text that sounds plausible ("I denied the loan due to debt history"). At the same time, the underlying model actually relied on a proxy variable for race. Trustworthy AI seeks to mitigate this by enforcing low bias and high criteria alignment, both of which are true and convincing.

## 4.3 Reliable AI

Parallel to these interpretability-focused schools is the school of Reliable AI, which emphasises consistency and robustness and argues that understanding the why is a luxury, whereas preventing catastrophic failure is a necessity. Seminally outlined by Amodei et al. (2016) [2], this school treats the model as a statistical object to be bounded. The quirk here is the Ostrich Effect, where the reasoning process is ignored in favour of verifiable safety constraints. This school employs adversarial training to solve the Min-Max Saddle Point Problem, as shown in Eq. 4.4

$$\min_{\theta} \mathbb{E}_{(x,y) \sim D} \left[ \max_{\delta \in \Delta} L(f_{\theta}(x + \delta), y) \right] \quad (4.4)$$

Recent advances have expanded this domain into the physical implementation of AI, prompting questions about whether digital hardware (Turing machines) can ever be truly reliable for continuous-time problems. Boche et al. argue that digital computing introduces an algorithmic solvability gap for inverse problems, in which digital approximations (Turing machines) are less transparent and reliable than analogue models (Blum-Shub-Smale machines) [6]. Furthermore, research into Spiking Neural Networks (SNNs) highlights that while biologically plausible and energy-efficient, discretising these continuous-time dynamics for digital hardware can lead to complexity blowup, rendering the system theoretically unreliable or computationally infeasible [33]. Thus, Reliable AI is increasingly concerned not just with the software model, but with the physics of the compute substrate itself. A brief comparison of the three schools of DL models' interpretation is summarised in the following Table 4.1.

Table 4.1: Comparative Taxonomy of AI Trust Schools

| Feature           | XAI (Transparent)                    | Trustworthy (Aligned)                                                   | Reliable (Robust)                            |
|-------------------|--------------------------------------|-------------------------------------------------------------------------|----------------------------------------------|
| Defining Paper    | Gunning (2017)                       | Li et al. (2023);<br>Jacovi & Goldberg (2020)                           | Amodei et al. (2016);<br>Boche et al. (2025) |
| Optimization Goal | Faithfulness<br>( $\min \ E - f\ $ ) | Alignment<br>( $\min \mathcal{L} + \lambda \mathcal{R}_{\text{fair}}$ ) | Worst-Case Safety<br>( $\min \max L$ )       |
| Defining Quirk    | Complexity                           | Deception                                                               | Opacity                                      |

### 4.3.1 Grad-CAM Post-Hoc Visual Validations: Benefits and Limits

To foster trust and ensure the AI is not relying on spurious correlations, XAI is mandatory. Although modern DL models, notably CNNs and Vision Transformers (ViTs), have demonstrated exceptional performance, their black-box nature fundamentally limits our understanding of their decision-making processes [109]. In high-stakes scenarios such as medical diagnostics, this opacity poses significant risks. For

instance, a model trained to detect pneumonia might achieve high accuracy not by identifying pathological features, but by focusing on a hospital-specific pattern or device tag, rendering it erroneous when deployed in a new environment.

To address this, post-hoc attribution methods are dominant in computer vision. These techniques generate saliency maps that highlight the pixels or regions of an input image that most influence the final prediction. Grad-CAM uses the gradients flowing into the final convolutional layer of a CNN to produce a coarse localisation map of these critical regions. Visual analysis confirms if the model focuses on medically meaningful areas; for example, verifying that COVID-19 classifications rely on the pleural line rather than background artefacts. While incredibly useful for debugging, these methods share a fundamental limitation: they are purely diagnostic. They can reveal that a model relies on spurious correlations, but they offer no direct mechanism for correcting it.

### 4.3.2 Evolution to Intrinsically Guided Training Paradigms

The realisation that diagnostic tools cannot fix the underlying reasoning gap has driven research toward a third paradigm: using interpretability tools as an active component of the training process itself [108]. This approach moves beyond post-hoc analysis and intrinsically interpretable models (such as prototype-based networks [17]) to training-time interventions that guide the model's learning process. The core hypothesis of this evolution is that the model already possesses sufficient information within the training data to be both precise and capable of generalisation, provided it is guided to focus on the correct features.

This guidance is achieved by manipulating the loss function, which serves as the cornerstone of the optimisation problem. Just as Physics-Informed Neural Networks (PINNs) bind input data to physical laws through loss terms [102], or self-supervised models like DINOv3 [121] employ objective functions to prevent feature degradation, the loss function can be engineered to shape the model's internal representations [93]. Early efforts in this direction, such as those by Ross et al., proposed penalising input gradients to suppress sensitivity to irrelevant features, though this required significant human annotation overhead [108]. Subsequent research has refined this by using influence functions or contextual decomposition to align neural networks with prior knowledge [71, 104].

### 4.3.3 A new Approach: Learning for the Right Reasons

Building upon these foundations, the research pathway evolves into a novel training paradigm that fuses the Grad-CAM algorithm directly with a reconstruction loss. Unlike previous methods that relied heavily on explicit human supervision of irrelevant features, this approach leverages the data distribution itself to generate a supervisory signal. By combining Grad-CAM with a prototypical reconstruction loss, the model is forced not only to classify an image correctly but also to detect which part of the image should receive its focus.

This method operates by comparing the generated explanation (the CAM) against a class prototype, effectively providing evidence for which part of the image should be used to return the classification. This compels the model to learn for the "right



reasons" by ensuring that the features used for prediction align with the class's semantic reality. The evolution of this architecture has shifted from inefficient hook-based methods to streamlined batch implementations (Batch-CAM), in which the model's averaged focus across a batch is regularised to achieve simultaneous improvements in accuracy and image reconstruction quality. This represents a shift toward more transparent, explainable, and trustworthy AI systems that bridge the gap between high predictive accuracy and human-interpretable reasoning.

## 4.4 In-Training Interpretability with Contrastive Loss

The genesis of this research trajectory was a response to the primary limitation of post-hoc explanation methods: they can reveal when a model relies on spurious correlations, but they cannot directly correct its reasoning. The initial proposed solution represented a paradigm shift: moving the generation of CAM, specifically Grad-CAM, from a post-training analysis step into the core training loop itself. The hypothesis was that by forcing the model's internal attention map to resemble the input image, it could be guided to focus on the salient object rather than irrelevant background features. The initial methodology proposed in "You've Got the Wrong Number" [55] implemented this concept via a two-part loss function, Eq. 4.5.

- First, a standard Multi-Class Cross-Entropy Loss ( $\mathcal{L}_{CE}$ ) was used for the primary classification task:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(p_{i,c}) \quad (4.5)$$

- Second, a custom Contrastive Loss ( $\mathcal{L}_L$ ) was introduced to enforce visual consistency. This function minimises the Euclidean distance ( $D$ ) between the Grad-CAM-reconstructed image and the original input for correctly matched pairs (label  $Y = 1$ ). For dissimilar pairs ( $Y = 0$ ), it ensures that the distance exceeds a specified margin  $m$ , effectively teaching the model to produce explanations that visually match the input. The formula is summarised as follows: Eq. 4.6 :

$$\mathcal{L}_L = \frac{1}{2} (Y \cdot D^2 + (1 - Y) \cdot \max(0, m - D)^2) \quad (4.6)$$

- The Total Loss ( $\mathcal{L}_{TOT}$ ), Eq. 4.7 was a simple combination of the two:

$$\mathcal{L}_{TOT} = \mathcal{L}_{CE} + \mathcal{L}_L \quad (4.7)$$

The initial experiment on the MNIST dataset was replicated across multiple random seeds, confirming that this new training regime could produce a model with accuracy comparable to a classically trained one, but with markedly different internal representations. The averaged results gave researchers the confidence to proceed to more complex datasets.

Qualitatively, the difference was striking, Table 4.2. The R-model's reconstructed CAMs demonstrated a superior ability to identify the digit's actual shape from the

Table 4.2: Metric (Averaged over 6 seeds) - MNIST

| Metric        | R-Model (Contrastive Loss) | C-Model (Classical) |
|---------------|----------------------------|---------------------|
| Test Accuracy | $\sim 98.44\%$             | $\sim 98.99\%$      |
| Training Time | $\sim 15.5$ hours          | $\sim 20$ hours     |

very first epoch. In contrast, the C-model’s feature maps were more abstract and less directly interpretable, even in later epochs. These promising but preliminary findings on a relatively simple dataset necessitated more rigorous validation and exploration on more complex data, leading to the next stage of research.

## 4.5 Validation, Challenges, and Algorithmic Refinement

Moving a novel method beyond proof of concept requires rigorous testing to assess its generalizability and robustness. This phase of the research served as a critical stress test for the initial paradigm, subjecting it to a more challenging dataset and revealing both its underlying strengths and its critical limitations.

The experimental setup was expanded to enhance robustness by using multiple random seeds for training. More importantly, the Fashion MNIST (FMNIST) dataset was introduced as a more complex challenge than the handwritten digits dataset. This new dataset exposed a critical flaw in the initial training strategy [56].

### 4.5.1 The Initial Failure Case

When applied to FMNIST, the initial R-model’s training was ineffective. The learning curves showed that the model’s validation accuracy plummeted after the first epoch, leading to premature early stopping around epoch 4 or 5. This resulted in a model with poor test performance ( $\sim 80.5\%$ ) and a significant accuracy gap compared to its classically trained counterpart, highlighting that the optimisation strategy was failing to navigate the more complex loss landscape of the FMNIST dataset.

### 4.5.2 A Successful Refinement

The research team gained a critical insight: the dual-objective loss function ( $\mathcal{L}_{CE} + \mathcal{L}_L$ ) created a much more complex and volatile loss landscape than standard classification. The initial, simpler optimiser was insufficient for this landscape. To correct this, they refined the training process with several key changes:

- Optimiser : switched from the standard ADAM to the AdamW optimiser;
- Learning Rate: increased the initial learning rate by an order of magnitude (from  $1e^{-3}$  to  $1e^{-2}$ );

- Scheduler: introduced the `ReduceLROnPlateau` scheduler, which dynamically adjusts the learning rate.

This more adaptive combination was not just a tweak, but a necessary algorithmic refinement. It enabled a much more effective exploration of the loss panorama, allowing the model to escape the poor local minima that had previously halted its training and leading to a dramatic improvement in performance, as shown in Table 4.3.

Table 4.3: Model Performance on FMNIST

|                 |     | Initial R-Model             | Refined R-Model              |
|-----------------|-----|-----------------------------|------------------------------|
| Training        | Be- | Premature Early Stopping (~ | Successful convergence (~ 20 |
| behaviour       |     | 5 epochs)                   | epochs)                      |
| Test Accuracy   |     | ~ 80.5%                     | ~ 89%                        |
| Performance Gap |     | Significant gap vs. C-Model | Competitive with C-Model     |

The paradigm was also extended to more complex architectures, such as ConvNeXT. This revealed the importance of the feature map dimensions; larger final feature maps produced much clearer and more useful reconstructions. However, applying the technique to transfer learning with frozen layers yielded mixed results, with reconstructed images showing a mix of coherent shapes and abstract patterns. This suggested that the methodology is best suited for end-to-end training, in which the entire network can adapt to the dual-objective loss function. While the refined method was now effective, it suffered from computational inefficiencies and a conceptually limited loss target, prompting a complete redesign of the architecture for the next major evolution.

## 4.6 Batch-CAM and Prototypical Reasoning

This stage represents an architectural leap: the research marks the maturation of the initial concept. The primary goals were to address the fundamental efficiency and conceptual limitations of the prior hook-based, per-image approach and to evolve the ad hoc method into a formalised, scalable paradigm named Batch-CAM [59]. A key innovation was the complete redesign of the Grad-CAM generation process to be more computationally efficient. The old method relied on attaching "hooks" to the model's layers, requiring an inefficient, iterative backward pass for every single image in a batch. The new functional architecture leverages PyTorch's `torch.autograd.grad` for a direct, vectorised computation, drastically reducing overhead.

Beyond efficiency, a pivotal conceptual shift occurred, Table 4.4: the reconstruction target was changed from the input image to a class prototype. In this context, a class prototype is the average of all training images for a given class (e.g., the average of all "T-shirt" images). This change offers a profound strategic benefit: instead of forcing the model to explain each unique instance, it enforces semantic

Table 4.4: Evolution of Grad-CAM Generation

|                 | Old Hook-Based Architecture                                                 | New Functional Architecture                                              |
|-----------------|-----------------------------------------------------------------------------|--------------------------------------------------------------------------|
| Gradient Access | Relies on registering forward/backward hooks to the model's layers.         | Uses the direct and stateless <code>torch.autograd.grad</code> function. |
| Computation     | Requires an iterative backward pass for each image in a batch (loop-based). | Performs a single, vectorised computation across the entire batch.       |
| Efficiency      | Higher computational overhead due to graph retention and loops.             | Lower computational overhead and significantly faster execution.         |
| Implementation  | More complex and deeply integrated into the model's class structure.        | Cleaner, self-contained, and does not require modifying the model.       |

consistency in its explanations across all instances of a class. The model now learns what a "T-shirt" generally looks like from an attentional perspective.

This led to two new loss formulations:

- **Prototype Loss (Per-Image Consistency):** for each individual image in a batch, this loss calculates the difference between its generated Grad-CAM and the corresponding pre-computed class prototype;
- **Batch-CAM Prototype Loss (Batch-Level Consistency):** This more aggregated approach first averages the Grad-CAMs for all instances of a specific class within a batch. This single, batch-averaged CAM is then compared to the class prototype, encouraging consistency at a group level.

The final Total Loss functions incorporated these new terms, along with different metrics (L1, L2, SSIM) to measure the reconstruction loss. The lambda ( $\lambda$ ) hyperparameter balances the contributions of the classification and reconstruction loss components, as in Eq. 4.8.

For the Prototype Loss approach:

$$\mathcal{L}_{Total} = \lambda_{class}\mathcal{L}_{Class} + \lambda_{proto}\mathcal{L}_{Metric}^{Proto}(M, P) \quad (4.8)$$

For the Batch-CAM Prototype Loss approach:

$$\mathcal{L}_{Total} = \lambda_{class}\mathcal{L}_{Class} + \lambda_{batch\_proto} \sum_{c \in \mathcal{C}_B} \mathcal{L}_{Metric}^{BatchProto}(\bar{M}_c, P_c) \quad (4.9)$$

This refined, efficient, and conceptually robust new methodology, Eq. 4.9, was then subjected to a comprehensive battery of tests to quantify its advantages.

## 4.7 Comprehensive Experimental Analysis of the Batch-CAM Paradigm

The goal of the final experimental phase was to definitively measure the impact of the mature Batch-CAM training methodology on both quantitative accuracy and the qualitative nature of the model’s internal reasoning. These tests were conducted across multiple modern architectures to validate the approach’s general applicability.

The results showed a nuanced picture. While the proposed methods often achieved the best performance, particularly for ResNet18 and SimpleCNN across both datasets, the baseline training remained competitive and, for ConvNeXt-V2 on FashionMNIST, proved superior. This indicates that while prototype-driven losses provide a powerful inductive bias, their effectiveness can be architecture-dependent (Table 4.5).

Table 4.5: Peak Model Performance Across Architectures

| Model            | MNIST<br>Best Acc. (%) | MNIST<br>Best Method | FMNIST<br>Best Acc. (%) | FMNIST<br>Best Method |
|------------------|------------------------|----------------------|-------------------------|-----------------------|
| ConvNeXt-V2 Tiny | 99.72                  | BCPL L1              | 94.79                   | Baseline              |
| ResNet18         | 99.65                  | BCPL L2              | 94.99                   | BCPL L2               |
| SimpleCNN        | 99.00                  | BCPL L1              | 91.32                   | BCPL L1               |

The most significant impact, however, was in the qualitative improvement of the model’s reasoning. The averaged CAMs produced by baseline models were often unfocused and ambiguous, as they depict different parts of the picture and fail to create a clear representation of the class. In stark contrast, models trained with the Batch-CAM Prototype Loss produced highly coherent and precise CAMs that clearly delineated the defining features of each class.

This methodology also proved to be a powerful diagnostic tool. By examining the generated CAMs for misclassified items in the test set, it is possible to gain invaluable, human-interpretable insights into the model’s specific failure modes. For example, when a model trained with Batch-CAM misclassified a "Pullover," its reconstructed CAM revealed that its attention had shifted away from the defining torso and sleeves to a nondescript region. Similarly, when confusing an "Ankle boot" with a "Sneaker," the CAM showed its focus was on the lower shoe profile, a feature common to both, while ignoring the boot’s defining ankle structure.

The evidence from these experiments demonstrates a powerful dual benefit: a simultaneous improvement in classification accuracy and a profound enhancement in the transparency and interpretability of the model’s decision-making process, directly addressing the core goals of Reliable AI.

## 4.8 Synthesis and Conclusion: A New Paradigm for Reliable AI

This report charts an evolutionary research journey, beginning with a simple yet novel idea: embedding a Grad-CAM-based contrastive loss into the training of a single CNN on the MNIST dataset. Empirical validation challenged this concept with more complex data, prompting an algorithmic refinement of its optimisation strategy. This, in turn, revealed underlying inefficiencies and conceptual limitations that spurred a complete architectural redesign. The culmination of this process is the Batch-CAM paradigm, a sophisticated, efficient, and prototype-driven methodology applicable to state-of-the-art architectures.

This research trajectory provides a powerful, concrete example of the principles of Reliable AI in action. Traditional post-hoc XAI can only diagnose a model’s flawed reasoning after the fact. By embedding an explanatory objective directly into the loss function, this paradigm fundamentally alters the learning process. It forces the model to learn representations that are not only predictive but also semantically consistent and visually robust. This proactive constraint guides the model away from relying on spurious correlations and towards building a more generalizable and reliable internal concept for each class, thereby enhancing its reliability.

Ultimately, the core contribution of this body of work is a compelling argument, backed by empirical evidence, for a new way of thinking about model development: explainability should not be treated as a post-hoc feature but as a fundamental component of the training objective itself. This approach paves the way for AI systems that are inherently more robust, transparent, and trustworthy.

## 4.9 Reliable Learning in Point-of-Care Ultrasound

The transition from benchmark datasets like MNIST to real-world applications is the ultimate test for any reliable AI paradigm. Given the proposed future direction outlined in previous chapters, the Batch-CAM methodology was deployed and rigorously tested in a high-stakes medical domain: LUS analysis for POCUS devices.

### 4.9.1 The Challenge of Medical Sonography

Unlike the clean, object-centric images of fashion items or digits, ultrasound data is characterised by high noise, operator-dependent variability, and affine artefacts [80]. The specific task involved detecting B-lines, which are critical indicators of pulmonary conditions such as oedema or fibrosis. The dataset consisted of 1,208 images categorised into the four severity scores introduced in chapter 3.7. The objective was to train a model designed to be precise, ensuring it focused on the pathological B-lines rather than on spurious correlations in the ultrasound scan. To ensure deployability on embedded medical devices, the lightweight ConvNeXt-Tiny architecture was selected as the backbone [84].

### 4.9.2 Methodological Adaptation: Geometry and Noise

The core Batch-CAM philosophy, regularising training via class prototypes, was adapted to handle the stochastic nature of medical imaging. Three specific variations of the reconstruction loss ( $\mathcal{L}_{recon}$ ) were investigated to determine the most effective way to guide the model:

- **Batch-CAM Prototype Loss:** The standard approach, where the CAM is averaged across the batch to smooth out individual noise before comparison with the prototype;
- **Augmented Prototype Loss:** An attempt to explicitly model ultrasound distortions (rotations, scaling) by randomly augmenting the class prototype before comparing it to the individual image's CAM;
- **Geometric-Aware Model (STN):** A structural variation incorporating a Spatial Transformer Network (STN) to geometrically normalise [62] the input before feature extraction, aiming to make the CAMs invariant to probe orientation.

### 4.9.3 Implicit Regularization Results

The experimental results provided a decisive validation of the Batch-CAM approach in this complex domain. As detailed in Table 4.6, the baseline Batch-CAM method utilising L1 loss achieved the highest test accuracy of 92.86%.

Table 4.6: Test Performance on Lung Ultrasound Dataset

| Experiment Method       | Accuracy (%) |
|-------------------------|--------------|
| BatchCAM_Prototype_L1   | 92.86        |
| STN_Prototype_L1        | 92.31        |
| Augmented_Prototype_L1  | 91.76        |
| BatchCAM_Prototype_SSIM | 91.21        |

Two critical insights emerged from these findings:

1. **Batch Averaging vs. Explicit Augmentation:** the Batch-CAM method outperformed the Augmented Prototype approach (92.86% vs. 91.76%). This suggests that averaging CAMs within a batch acts as a superior "noise reduction" mechanism, implicitly handling patient variability better than manually engineered augmentations can simulate it;
2. **L1 vs. SSIM:** models trained with L1 loss consistently outperformed those using the structural SSIM metric. In the context of B-line detection, the pixel-wise intensity of the activation map appears to be more diagnostically relevant than structural similarity, forcing the model to learn a "constrained and biased attention map" strictly aligned with the pathology.

Qualitative analysis confirmed that the model successfully learned to ignore the conical beam boundaries and background noise, focusing its attention strictly on the vertical B-line artefacts essential for diagnosis. This successful deployment demonstrates that the Batch-CAM paradigm is robust enough to enforce reliable, interpretable reasoning even in the noisy, chaotic domain of medical diagnostics (Figures 4.1-4.2).

## 4.10 Future Directions

Based on the findings and limitations identified in this research, several key avenues for future work have been proposed to extend and validate this new paradigm:

- **Complex Real-World Datasets:** extend and validate the Batch-CAM paradigm on more complex domains where spurious correlations are prevalent, and the cost of opaque reasoning is high, such as in medical imaging;
- **Architectural Redesign:** investigate the redesign of network backbones, potentially towards U-Net-like structures, to further enhance reconstruction capabilities and improve the interpretation of classification problems;
- **Advanced Prototype Generation:** explore more sophisticated methods for generating class prototypes beyond simple averaging, such as using generative models to capture richer and more varied representations of each class archetype.

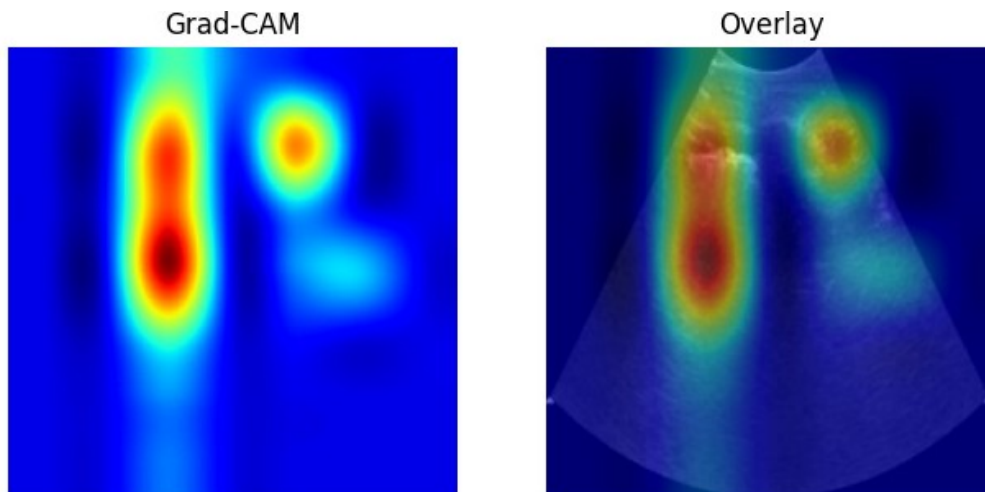


Figure 4.1: Example of good prototype identification using the Batch-CAM prototype training approach



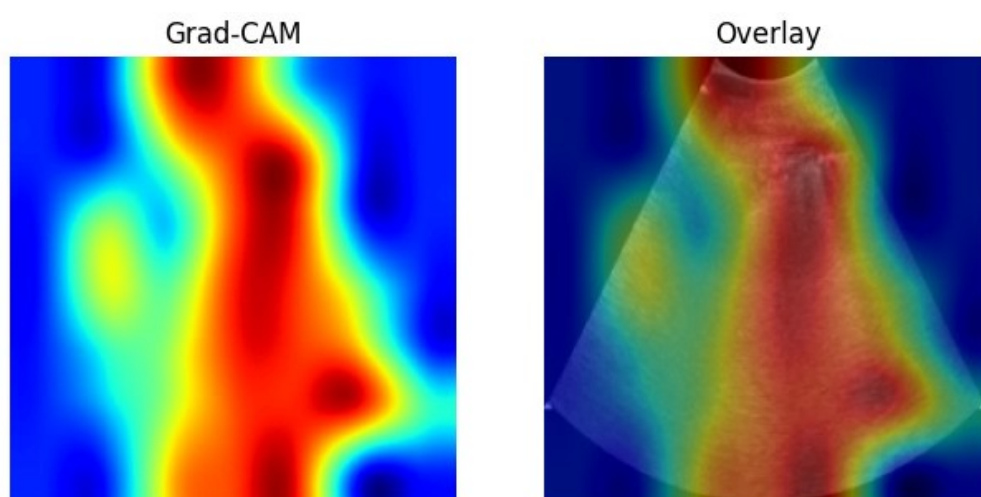


Figure 4.2: Example of good prototype identification using the Batch-CAM prototype training approach with the geometric aware network



## Chapter 5

# Advanced AI Integration

The material presented in this chapter is based on the conference papers and preprints:

- G. Ignesti, D. Moroni, and M. Martinelli. Plant-traits: how citizen science and artificial intelligence can impact natural science. In 2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS). IEEE, 2024,
- G. Ignesti, D. Moroni, and M. Martinelli. Towards the actual deployment of robust, adaptable, and maintainable ai models for sustainable agriculture. Position Papers of the 19thConference on Computer Science and Intelligence Systems (FedCSIS), page 33, 2024,
- S. Daoudagh, G. Ignesti, D. Moroni, L. Sebastiani, and P. Paradisi. Assessment of dance movement therapy outcomes: A preliminary proposal. In International Conference on Computer-Human Interaction Research and Applications, pages 382–395. Springer, 2024.

This chapter outlines the strategy for transitioning plausible LUS models into robust, long-term clinical tools. While the previously described pipelines establish a solid foundation, the final pipeline presents two critical areas for improvement to ensure clinical viability. A critical issue is that DL algorithms inherently suffer from catastrophic forgetting [70], in which models lose previously acquired knowledge when trained on new data. Consequently, they often function as one-trick tools, lacking an accepted strategy for contextualising updated training weights without degrading the network’s baseline accuracy. To address this, CL techniques should be integrated to enable the model to adapt to new data streams incrementally while retaining prior diagnostic capabilities [46]. Second, there is a distinct lack of competent multimodal data fusion models. While LUS images are a critical source of information, they represent only a fraction of the diagnostic picture from a clinical perspective. A fully developed DL model for objective use must be capable of simultaneously processing disparate signal types and integrating visual data (LUS images) with anamnestic and clinical metadata (Electronic Health Records) [51]. This requirement specifies the architectural support provided by the DSS as referred to in Chapter 2. Furthermore, the TiAssisto framework outlines strategies to maintain performance in these dynamic environments and explores novel AI techniques to expand diagnostic capabilities beyond simple image classification [53, 54].

## 5.1 Integration within Hybrid DSS

Starting from the latter topic, the specificity of deployed DL models and their widespread clinical adoption require that they be embedded in a functional ecosystem rather than operating in isolation [116]. We propose a hybrid DSS that effectively utilises both deductive (rule-based) and inductive (DL) approaches. In this architecture, the DSS intervenes both before AI models are executed, by selecting appropriate models based on historical data and specific clinical contexts, and subsequently, by synthesising AI outputs with contextual rules to offer refined suggestions. This bidirectional intervention converts raw predictive insights into actionable clinical knowledge, aligning with the broader requirement for robust systems capable of managing high-stakes medical data [69].

### 5.1.1 Multimodal Feature Fusions

While the current LUS model achieves high accuracy in identifying pathological signs from ultrasound frames, it operates in isolation from the patient’s broader physiological context. To establish a proven baseline for this multimodal approach, we can draw on empirical validation from environmental monitoringsspecifically, recent work on plant trait prediction [68, 113], where visual data proved insufficient for regressing complex biological traits when used in isolation [57]. In that study, we utilise a Feature Fusion architecture that concatenates deep visual features extracted by an EfficientNet-b0 backbone with structured tabular data, processed either through a separate fully connected stream or via a Feature Tokeniser Transformer (FT-Transformer) [37]. The results demonstrated that this multimodal approach, which fused visual embeddings with tabular covariates, yielded superior generalisation compared to unimodal baselines, effectively grounding visual patterns within their environmental context. Crucially, empirical results from our environmental research (Table 5.1) indicate that while FT-Transformers offer theoretical advantages, simpler Artificial NNs (ANNs) achieve performance comparable to state-of-the-art tree ensemble models [38]. This justifies the use of lightweight ANNs for TiAssisto’s mobile deployment, minimising computational overhead on edge devices. Furthermore, our experiments highlighted that while standard inductive transfer learning failed to adapt to noisy, decentralised domains (yielding negative  $R^2$  scores in Table 5.2), the feature fusion approach remained robust [140, 143]. Anchoring visual patterns to tabular covariates is thus necessary for stabilising predictions in high-variability environments.

The  $R^2$  metric presented in Table 5.1 serves as a benchmark for a series of experiments designed to address the inherent complexity of data fusion. The primary objective of this analysis is to determine which architectural combination yields superior performance when integrating tabular data with previously developed models. Among the various strategies evaluated, the fusion of the FT-Transformer with the minimal ensemble emerges as the most promising approach, as it effectively leverages the strengths of both data modalities. However, the relatively low  $R^2$  metric observed on the test set indicates that predicting complex tabular information solely from images remains a significant challenge. Despite these limitations, the results suggest that this fitting framework provides a viable baseline for predicting plant trait in-

Table 5.1: DL model results: R2 metric for performance evaluation across varying architectures.

| Model                     | Validation R2 Metric | Test R2 Metric |
|---------------------------|----------------------|----------------|
| TF weak                   | 0.10                 | 0.08           |
| TF Ensemble               | 0.15                 | 0.10           |
| Trained Weak              | 0.40                 | 0.15           |
| Trained Ensemble          | 0.53                 | 0.18           |
| I-Ensemble ANN            | 0.56                 | 0.247          |
| I-Ensemble ANN (2023)     | //                   | -0.08          |
| I-Ensemble FT-Transformer | 0.56                 | 0.248          |
| Minimal I-Ensemble ANN    | 0.45                 | 0.21           |

formation, laying the groundwork for further refinements in multi-modal predictive modelling. Table 5.2 delineates a series of winning strategies applied across different

Table 5.2: Comparison of model performance: The 2023 winning model vs. adaptive ensemble adjustments for the PT2024 challenge.

| Model                 | Classification Accuracy [%] | Test R2 Metric |
|-----------------------|-----------------------------|----------------|
| 2023 Model 1          | 0.81                        | 0.66           |
| 2023 Model 2          | 0.81                        | 0.62           |
| 2023 Model Bagging E  | 0.83                        | 0.69           |
| 2023 Model Adaptive E | 0.83                        | 0.68           |
| 2024 Model 1          | 0.81                        | -39            |

model architectures, highlighting a critical discrepancy in domain adaptability. It is particularly notable that the specific model architecture which successfully adapted to the 2023 challenge dataset failed to reproduce comparable performance when applied to the 2024 test set. While the 2023 models achieved robust classification accuracy and regression performance, applying these architectures directly to the 2024 data yielded a negative  $R^2$  value, underscoring a fundamental failure to learn the new task. This contrast reveals that features optimised for the specific conditions of the 2023 dataset do not generalise effectively to the 2024 challenge, indicating a significant task shift that prevents straightforward transfer learning between these seemingly similar biological domains. The empirical findings from the plant trait analysis provide critical validation of the architectural design of the TiAssisto LUS framework, particularly the shift from a purely Weak AI image classifier to a holistic diagnostic engine. The feature fusion architecture concatenates the flattened feature vector from the LUS ensemble with the tabular stream of vital signs (oxygen saturation, temperature, blood pressure). By feeding these fused representations into a unified regression or classification head, the system learns non-linear correlations between visual artefacts (e.g., B-lines) and physiological distress (e.g., desaturation events). This mimics the clinical reasoning process, in which a physician synthesises ultrasound scans with the patient’s vital signs and history.

### 5.1.2 The Citizen Science Paradigm: Decentralised Data Governance

The deployment of POCUS devices to non-expert operators (such as patients in home care or general practitioners) fundamentally mirrors the structure of Citizen Science (CS) initiatives [8, 120]. As analysed in our position paper on sustainable AI [58], CS relies on crowd-sensing to gather vast amounts of distributed data [20, 87], but suffers from inherent variability in data quality and acquisition conditions [64]. In the agricultural domain, we observed that while data abundance from citizen scientists is beneficial, the lack of authoritative supervision necessitates rigorous automated filtering and uncertainty quantification to prevent model degradation [85].

The TiAssisto platform faces an identical task shift, in which the transition from expert-generated hospital-quality imagery to home-acquired POCUS scans introduces noise and operator variability. To address this, we propose adopting the Crowd-Sensing Quality Assurance protocols developed for environmental monitoring. This involves a pre-processing gatekeeper module, derived from the GranoScan application architecture, which enforces a dual-layer validation strategy [22]:

- **Blind Quality Assessment:** A lightweight DL filter evaluates technical parameters, checking for acoustic coupling quality and gain settings;
- **Semantic Consistency Check:** An object detection layer verifies the presence of requisite anatomical markers, principally the pleural line.

This filter ensures that only clinically interpretable frames reach the diagnostic ensemble. Furthermore, the CS perspective reframes the retraining process through CL. Just as environmental models must adapt to seasonal shifts, medical AI must dynamically update its batch normalisation statistics and feature weights to account for the specific noise profiles of different portable probes used in the field.

### 5.1.3 Synthesis: Interconnecting Fusion and Crowd-Sensing

The symbiosis between Multimodal Feature Fusion and Citizen Science principles provides a comprehensive solution for decentralising medical diagnostics. While the Citizen Science paradigm provides the necessary filters to ensure data integrity at the point of acquisition (the "Gatekeeper") [133], Feature Fusion provides the biological grounding required to interpret that data accurately.

This interconnection is critical: the gatekeeper module filters out the noise inherent in crowd-sensed data, ensuring the Feature Fusion stream receives high-fidelity input. In turn, the fusion of tabular vitals provides a "safety net" for the Citizen Science data; even if an image is acquired with minor suboptimal orientation by a non-expert, the presence of correlated physiological data (e.g., low  $SpO_2$ ) constrains the AI's manifold of plausible predictions, preventing the catastrophic failures common in unimodal, noisy environments. Together, these methodologies transform TiAssisto from a simple image analysis tool into a robust, context-aware diagnostic ecosystem.

## 5.2 Lifelong Learning

The dynamic nature of healthcare demands that models adapt continuously rather than relying on static validation. The paradigm of fixed datasets must be superseded by continuous or lifelong learning frameworks [96]. This approach enables the AI system to incrementally acquire new information and adapt to emerging data streams without succumbing to catastrophic forgetting. CL ensures that the model remains robust and reliable as patient demographics shift or treatment protocols evolve. Furthermore, decentralised data management, facilitated by federated learning, offers a pivotal solution for deploying models across distributed datasets. This methodology preserves patient privacy while enabling large-scale, blind studies essential for verifying clinical utility [105].

### 5.2.1 Expanding Plausible Analysis and Alternative AI

Future research aims to extend plausible analysis beyond simple classification to include localisation and quantification tasks. Additionally, an investigation into new forms of AI integration is required for objective assessment in medical contexts. For instance, in related health domains, Human Pose Estimation (HPE) techniques, specifically architectures such as YOLO-NAS-Pose, are being examined as objective methods for analysing movement and physiological signals. These have proven more robust than traditional computer vision methods, such as MediaPipe, particularly in complex multi-participant scenarios, suggesting potential cross-domain applications for complex clinical monitoring [23].

## 5.3 The Stability-Plasticity Dilemma LUS

The transition of DL models from static validation to real-world deployment, as in the TiAssisto platform example, introduces a fundamental challenge known as the stability-plasticity dilemma. As medical protocols evolve and patient demographics shift, the AI system must remain sufficiently plastic to acquire new knowledge, such as identifying novel SARS-CoV-2 variants or adapting to different ultrasound probe manufacturers, while remaining stable enough to preserve its diagnostic capabilities [96, 97].

In standard DL training, optimising parameters  $\theta$  for a new task  $\mathcal{T}_{new}$  typically degrades performance on the original task  $\mathcal{T}_{old}$ . This phenomenon, known as Catastrophic Forgetting, occurs because the gradient updates for the new task may be orthogonal or opposing to the solution manifold required for the previous task [31]. In the context of LUS, this risk is clinically unacceptable; a model that learns to identify a new viral pneumonia pattern but subsequently fails to detect basic pulmonary oedema (B-lines) renders the DSS unreliable.

## 5.4 EWC for Ensemble Maintenance

To mitigate catastrophic forgetting [70] within the TiAssisto framework, we propose integrating Elastic Weight Consolidation (EWC) [79] directly into the fine-tuning stage of the Efficient Adaptive Ensemble defined in Chapter 3.

Recall that the architecture consists of frozen weak learners (EfficientNet-b0 feature extractors) and a trainable linear combination layer defined by weights  $W$  and bias [3]. This structural separation offers a distinct advantage for CL. Since the high-dimensional feature extractors are pre-trained and frozen, the phenomenon of forgetting is naturally constrained to the combination layer [41].

We formalise the update rule for the combination-layer weights  $\theta$  when encountering a new clinical dataset  $\mathcal{D}_B$  in a Bayesian framework. The loss function  $\mathcal{L}_{CL}(\theta)$  is defined as the following Equation 5.1.

$$\mathcal{L}_{CL}(\theta) = \mathcal{L}_{\mathcal{D}_B}(\theta) + \sum_i \frac{\lambda}{2} F_i (\theta_i - \theta_{A,i}^*)^2 \quad (5.1)$$

In this equation,  $\mathcal{L}_{\mathcal{D}_B}$  represents the standard Cross-Entropy loss on the new data [114], while  $\theta_{A,i}^*$  denotes the optimal weights of the combination layer derived from the previous task (Task A) [70, 131]. The term  $F_i$  refers to the diagonal of the Fisher Information Matrix, which represents the importance of parameter  $i$  for the previous diagnosis [52]. Finally,  $\lambda$  acts as the regularisation coefficient, controlling the trade-off between plasticity and stability [86]. In the context of the Efficient Adaptive Ensemble, calculating  $F_i$  is computationally highly efficient. The calculation is restricted to the combination layer, comprising approximately 100,000 parameters, rather than the millions of parameters found in backbone architectures [63]. This efficiency allows the system to quantify the importance of specific feature combinations on-device without incurring the computational debt of full network retraining. Consequently, if a specific feature vector, for example, one representing a thickened pleural line, is crucial for the COVID-19 class, its corresponding weight will possess a high Fisher value, penalising significant deviations during subsequent training phases [25].

## 5.5 Prototype Rehearsal via Batch-CAM

While regularisation methods like EWC protect model weights, they do not obviate the need for data retention. Traditional Replay strategies involve storing a buffer of past training samples to mix with new data. However, retaining raw patient ultrasound videos raises significant privacy and storage concerns within telemedicine platforms such as TiAssisto. To address this, we propose a novel Prototype Rehearsal strategy leveraging the Batch-CAM paradigm developed in Chapter 4. Instead of storing raw patient data  $x$ , the system retains the Class Prototypes ( $P_c$ ) generated during the training of the Explainable AI module [59]. These prototypes represent the ideal activation map for a given pathology, such as the canonical vertical artefact pattern for B-lines [54]. During the retraining phase for a new task, these prototypes serve as regularisers, preventing the model’s internal attention mechanisms from drifting. This approach aligns with the Reliable AI objective, ensuring



the model retains not only the correct classification output but also the correct reasoning process[55, 56, 59].

### 5.5.1 Experimental Validation of CL Protocols

To validate the efficacy of the proposed integration of the EWC and Prototype Rehearsal, we implemented a sequential learning experiment using a modified EfficientNet-b0 architecture. The experimental framework simulated a clinical update cycle in which a model trained on an initial hospital dataset (Task A) is subsequently fine-tuned on a new dataset (Task B) from a different domain or device. We compared three configurations:

- Baseline ( $\lambda_{EWC} = 0, \lambda_{Proto} = 0$ ): Unconstrained fine-tuning, serving as a control for catastrophic forgetting,
- EWC Only ( $\lambda_{EWC} = 2000$ ): Standard weight consolidation without prototype guidance,
- Proposed Hybrid ( $\lambda_{EWC} = 5000, \lambda_{Proto} = 2000$ ): The complete framework integrating both weight constraints and prototype replay.

The results, summarised in Table 5.3, demonstrate the protective effect of the proposed methodology. The Baseline model exhibited signs of instability, with accuracy on the historical test set (Old Test) dropping to 89.3% as it prioritised the new task (99.8%). The introduction of EWC significantly mitigated this forgetting, recovering stability to 94.2%. Crucially, the Proposed Hybrid approach achieved a balanced performance profile. By enforcing semantic consistency through Prototype Rehearsal ( $\lambda_{Proto} = 2000$ ), the model maintained high retention of prior knowledge (90.9% Old Test Accuracy) while adapting successfully to the new clinical targets (99.2% New Test Accuracy). Furthermore, qualitative monitoring via Grad-CAM confirmed that the hybrid model retained its focus on clinically relevant features, such as pleural lines. In contrast, the baseline showed greater susceptibility to attention drift.

Table 5.3: Comparative Analysis of CL Strategies. Old Train memory retention, Old Test stability, and New Test plasticity.

| Configuration      | $\lambda_{EWC}$ | $\lambda_{Proto}$ | Stability | Plasticity |
|--------------------|-----------------|-------------------|-----------|------------|
| Baseline           | 0               | 0                 | 0.89      | 0.99       |
| EWC Only           | 2000            | 0                 | 0.942     | 0.99       |
| Low Regularisation | 100             | 100               | 0.87      | 1.0        |
| Proposed Hybrid    | 5000            | 2000              | 0.91      | 0.99       |

These findings indicate that while EWC provides a strong foundation for parameter stability, integrating Prototype Rehearsal offers a complementary mechanism to enforce interpretability and semantic robustness, ensuring the diagnostic logic remains consistent across updates. These findings can be further investigated by analysing the shift in the mean activation map between the original and new training sets. Two samples of the activation map are illustrated in Fig. 5.1 and Fig. 5.2,

which visually contrast the improved detail and high stability (91%) achieved by the proposed method with the baseline EWC regularisation.

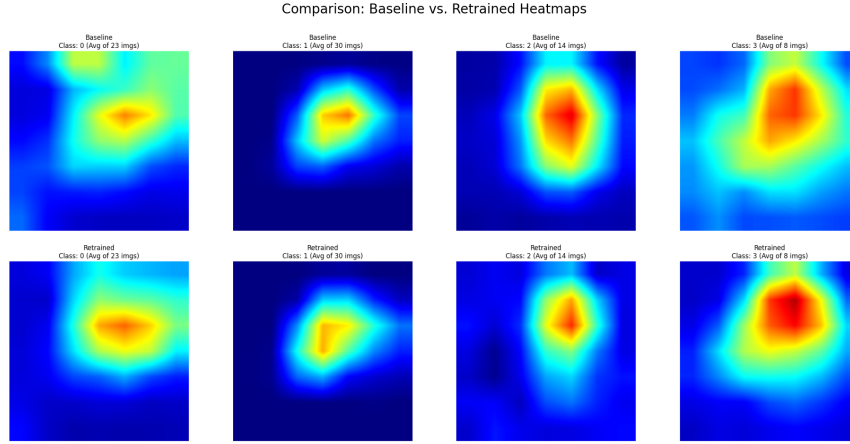


Figure 5.1: The comparison of the map of the trained network with the experiment with the proposed combination of EWC and Prototype regularisation, in this case, the extracted Grad-CAM seems to improve in analysing the detail, and even the stability is fairly high (91%).

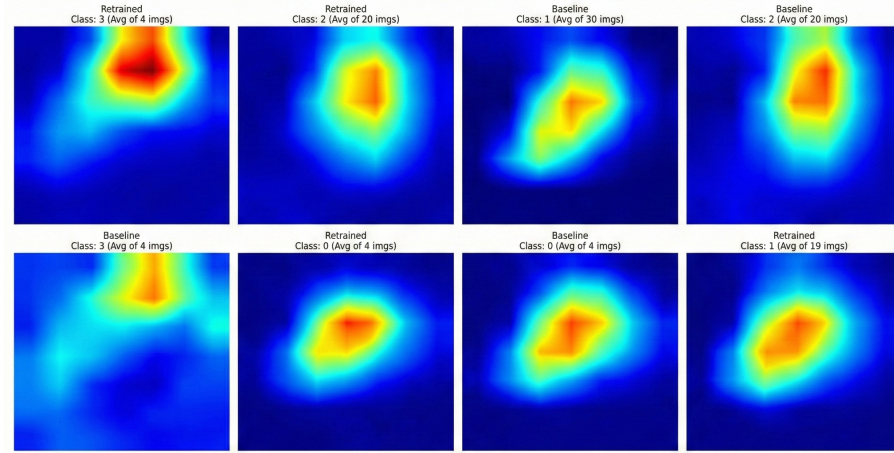


Figure 5.2: The comparison of the map of the original trained network with the experiment with only EWC regularisation, in this case, the Grad-CAM is very similar.

## 5.6 Federated Continual Learning Architecture

The final component of this integration addresses the decentralised nature of the TiAssisto platform, in which POCUS devices operate at the edge, ranging from static remote settings in patients' homes to dynamic mobile edge environments in ambulances [54]. We outline a Federated Continual Learning (FCL) protocol to facilitate model updates while strictly preserving privacy. The process begins with Local Adaptation, where POCUS devices collect new LUS sequences; if the local model

detects low confidence, i.e., via entropy thresholding [43, 72], the sample is flagged for learning. Subsequently, during Gradient Computation, the device computes gradients  $\nabla\theta$  for the combination layer locally, rather than uploading raw video data, which would consume bandwidth and compromise privacy [89, 118]. These gradients then undergo Secure Aggregation upon transmission to the central TiAssisto server. The server aggregates updates from multiple devices, using the EWC constraints described previously, to ensure that the global model does not suffer from catastrophic forgetting when integrating diverse local updates [70, 130]. Finally, a Global Update broadcasts the updated weights  $\theta^*$  back to all edge devices, ensuring the Weak AI component evolves organically with the epidemiological landscape. This architecture ensures that the AI model remains a robust, adaptive, and privacy-preserving partner in the Medicine 5.0 ecosystem.



## Part III

### Future Work and Conclusion



# Chapter 6

## Discussion

The research presented in this thesis, focusing on POCUS imaging features classification, establishes the foundational perception layer for a broader vision: developing a fully integrated, AI-driven Point-of-Care (PoC) diagnostic ecosystem. While current models successfully improve classification accuracy and pattern extraction in LUS images by introducing multimodal data fusion (integrating EfficientNet visual features with FT-Transformer tabular embeddings), translating these findings into deployable clinical systems demands a progression beyond binary pathology prediction. Next-generation Lung Ultrasound (LUS) diagnostic tools must advance three critical capabilities: (1) localisation of pathological signs within anatomical regions, (2) reasoning about physiological constraints and clinical context, and (3) explainability through natural language articulation of diagnostic evidence. To achieve this, future work must interconnect the efficient architectures proposed here with four emerging technological paradigms: advanced Object Detection, Physics-Informed Deep Learning, Graph-Based Reasoning, and Multimodal LLMs. Consequently, the central claim of this thesis is that realising such a robust diagnostic ecosystem requires a fundamental paradigm shift in DL training: moving from passive, post-hoc analysis to active, intrinsically guided learning that prioritises reliability and efficiency over raw model scale. The research contends that standard "black-box" models are inherently susceptible to the spurious correlation effect, in which they rely on background noise rather than anatomy, rendering them unsafe for high-stakes clinical decision-making. To overcome this, the thesis establishes the Efficient Adaptive Ensembling methodology, demonstrating that architectural intelligence (specifically, synthesising deep features from lightweight EfficientNet-b0 learners) can achieve state-of-the-art accuracy with a computational footprint small enough for real-time edge deployment (0.78 G-FLOPs). Crucially, the work argues that trust must be mathematically enforced during the training loop itself; by introducing the Batch-CAM paradigm, which embeds interpretability directly into the loss function via prototype-based constraints, the system is compelled to learn meaningful details, ensuring that diagnostic classifications are strictly aligned with causal anatomical features like B-lines or pleural irregularities. Ultimately, we expand the thesis's architectural vision beyond clinical diagnostics by validating Multimodal Feature Fusion in precision agriculture, specifically for predicting plant traits. This research demonstrates that concatenating EfficientNet visual embeddings with tabular environmental data processed via FT-Transformers significantly enhances generalisation

compared to unimodal baselines. This finding provides the empirical foundation for the TiAssisto hybrid DSS, which essentially mimics clinical reasoning by fusing visual ultrasound data with tabular vital signs (e.g., blood pressure, oxygen saturation). Furthermore, the chapter addresses the necessity of maintainability in these dynamic environments by proposing CL protocols. Specifically, it integrates EWC to anchor critical weights with Batch-CAM-based Prototype Rehearsal. This approach manages the stability/plasticity dilemma by replaying learned prototypes rather than raw data, thereby preventing catastrophic forgetting while strictly preserving patient privacy in decentralised ecosystems.

## 6.1 Beyond Classification

The first step in upgrading the current framework is transitioning from broad image classification to granular tasks. While the current system can flag a frame as pathological it lacks the spatial awareness to quantify specific biomarkers. Integrating state-of-the-art detection strategies into the architecture, such as YOLO [138] variants or Transformer-based detectors like RT-DETR (Real-Time Detection Transformer) [16], will allow the system to precisely bound and segment A-lines, B-lines, and consolidations. This granular output is critical. It pushes the AI from an algorithm that outputs a probability score into a more complex tool that highlights the pathology area and class. By feeding these localised features back into the reliable module developed in this thesis, we can create a feedback loop where the model explains its classification based on the density and location of detected artefacts. The further possibility of incorporating clinical information with standard signal and image analysis opens the way to a complete new analysis pipeline that provides the physician and the patient with a 360-degree view, Figure 6.1.

## 6.2 Beyond Reliability: Physical Reality

A critical limitation of pure computer vision approaches, including the ones used in this thesis, is that they treat ultrasound images merely as arrays of pixels, ignoring the underlying acoustic physics. LUS artefacts like B-lines are not random noise; they are reverberation artefacts governed by wave equations. Future iterations of this work should incorporate Physics-Informed Neural Networks (PINNs) [101]. By embedding the governing equations of ultrasound wave propagation into the neural network's loss function, PINNs can constrain the model to generate physically plausible predictions. This is particularly valuable for denoising scans or distinguishing between actual pathological artefacts and equipment interference. A PINN-enhanced backbone would ensure that efficient feature extractors (such as the EfficientNet-b0 used in this work) remain robust even when trained on smaller, noisy datasets, as they would be guided by physical laws rather than solely statistical correlations.



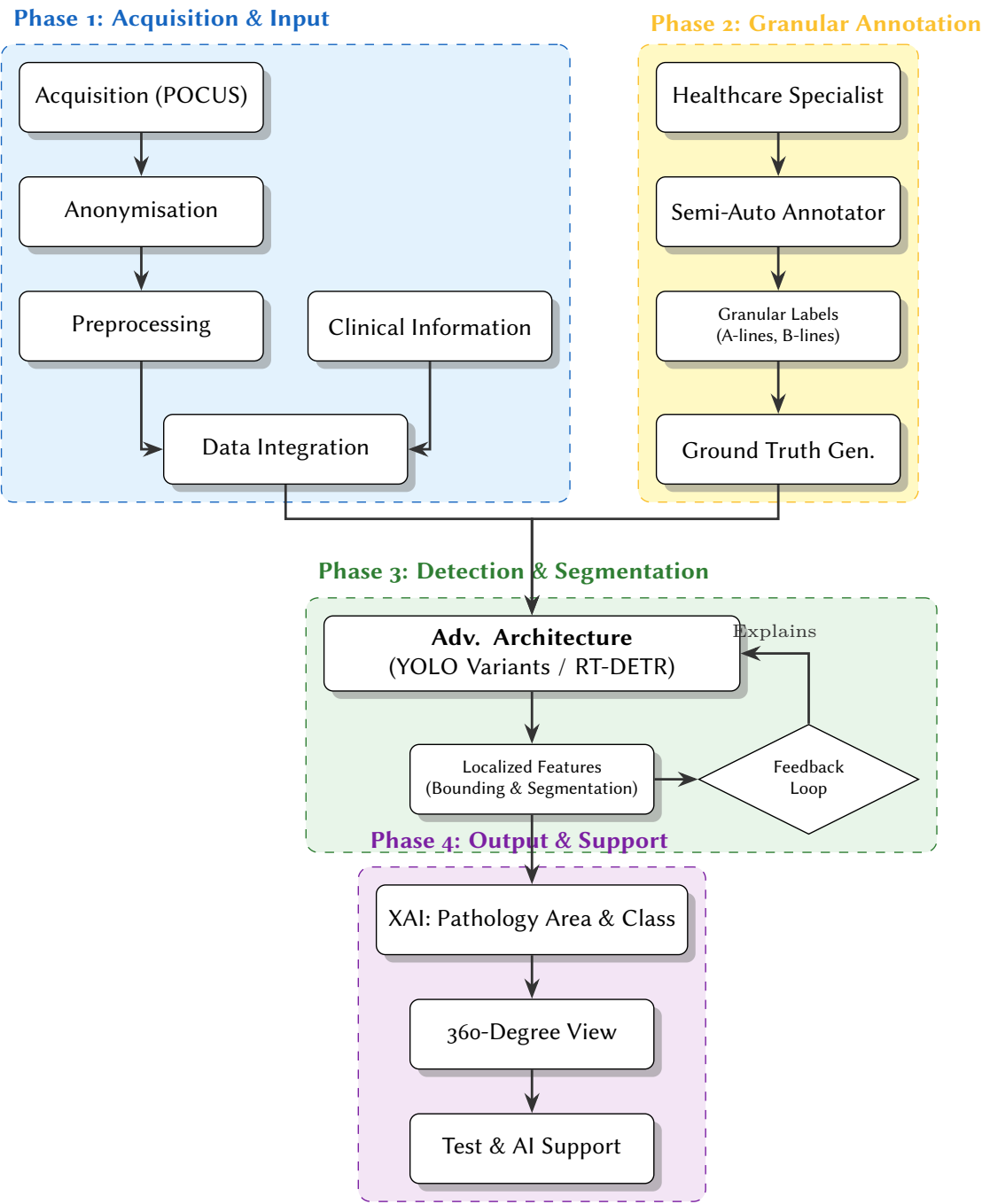


Figure 6.1: Proposed Advanced Ultrasound Analysis Pipeline.

## 6.3 Contextual Reasoning: Graph Neural Networks (GNNs)

Clinical diagnosis is rarely based on a single snapshot; it is a relational process. A complete LUS examination involves scanning multiple zones (usually 12) across the thorax. The current independent frame processing ignores this anatomical connectivity. Future work will leverage Graph Neural Networks (GNNs) to model the patient as a structured graph. In this topology, each ultrasound video clip serves as a node, and the spatial relationships between lung zones (e.g., upper vs. lower lobes) form the edge. By processing the embeddings generated by our multimodal architecture through a GNN, the system can detect patterns that only emerge at the patient level, such as the gradient of oedema (gravity-dependent distribution) typical in heart failure, which distinguishes it from the patchy distribution of ARDS/COVID-19. This moves the system from image analysis to patient analysis.

## 6.4 Chat-Bot: LLM for Reasoning

Finally, the successful deployment of these systems hinges on the interaction between AI and the clinician. This is where Multimodal Large Language Models (LLM + Vision) become the ultimate integrator. The fusion architecture proposed in this thesis can serve as the "vision encoder" for a larger LMM. Instead of outputting a class label, the system would feed its rich visual and tabular embeddings into an LLM (such as GPT-4V [141] or open-source equivalents like LLaVA [81]), allowing the AI to generate a comprehensive clinical report. This DL-based DSS would function as an intelligent layer, intervening to:

- **Select Algorithms:** dynamically choose the right sub-model (e.g., detection vs. classification) based on the context;
- **Synthesise Findings:** convert the detection of "multiple B-lines in Zone 1" and "clinical history of hypertension" into a coherent suggestion "High probability of cardiogenic pulmonary oedema; recommend correlating with BNP levels";
- **Facilitate Dialogue:** allow the clinician to query the scan (e.g., "Show me the consolidations in the lower lobe") and receive visual and textual responses.

## 6.5 Challenges

Deploying this complex Medicine 5.0 architecture poses significant challenges, primarily because foundation models serve as static snapshots of the world at the time of training. In the dynamic nature of clinical environments, these models face two distinct types of degradation: task-shift forgetting, in which adapting to new downstream objectives overwrites general knowledge, and time-shift forgetting, in which data distributions shift due to evolving external factors (e.g., new protocols or changing patient demographics). To counter this, modern CL paradigms [4] must be adopted, allowing the system to incrementally acquire and refine knowledge without

erasing prior capabilities. This involves specific strategies such as Continual Pre-Training (CPT) and Continual Fine-Tuning (CFT). These strategies are essential for mitigating knowledge staleness and distribution shifts, enabling the models to specialise and personalise to specific hospital domains or patient needs without the computational expense of full retraining. Furthermore, Federated Learning (FL) will be essential to train these data-hungry models (particularly LLMs and GNNs) across decentralised hospitals. FL frameworks allow for the adaptation of large models across organisations while preserving patient data privacy and optimising computational resources. However, this presents challenges in terms of communication costs and conflicting optimisation objectives due to varying data distributions across clients. Emerging solutions like FATE-LLM [30] enable collaborative training via parameter-efficient fine-tuning, mitigating the resource burden of communicating large parameter sets. They are applicable across high-stakes fields where objective data must augment subjective expert judgment. For instance, in Dance Movement Therapy (DMT), assessment has traditionally relied on subjective observation. Researchers are now developing frameworks that integrate AI-based Human Pose Estimation (HPE) with physiological signals, a direct parallel to our fusion of LUS video and clinical data. Advanced tools like YOLO-NAS-Pose [23] are being used to objectify movement analysis, mirroring our need for granular object detection in ultrasound. Similarly, in Sustainable Agriculture, AI must overcome environmental variability and sparse data [58]. Research here leverages adaptive ensembling and transfer learning to manage uncertainty techniques that validate our pursuit of robust, efficient architectures (like the EfficientNet-b0 backbone) capable of operating in resource-constrained environments. In conclusion, this thesis provides the necessary, efficient multimodal backbone for a future diagnostic ecosystem. By extending this work with physically grounded (PINNs), relationally aware (GNNs), and linguistically capable (LMs) technologies, we can bridge the gap between algorithmic performance and actual clinical utility.



# List of Figures

|     |                                                                                                                                                                                                                                                                                 |    |
|-----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1 | A husky dog wrongly classified as a wolf due to the snow in the image, example of the clever Hans effect in DL study, image from Ribeiro et al [103] . . . . .                                                                                                                  | 2  |
| 1.2 | A 3-D example of a nonlinear decision boundary . . . . .                                                                                                                                                                                                                        | 4  |
| 1.3 | An ideal flow chart to DL network issues . . . . .                                                                                                                                                                                                                              | 5  |
| 1.4 | A side-to-side comparison of chest medical imaging scans between US, CT, and MRI; the clear difference is that anatomical marks from the latter are clear, while US need further interpretation. . . . .                                                                        | 9  |
| 2.1 | This infographic illustrates the amount of data gathered per subject during the TiAssisto trial period, specifically detailing the reports for three of the main GPs enrolled in the project. . . . .                                                                           | 29 |
| 2.2 | In the figure, the new 5.0 telemedicine approach is represented, in which the platform works as an intermediate step between hospital care and care-home assistance. . . . .                                                                                                    | 32 |
| 2.3 | The figure portrays the Italian TiAssisto website, in which the involved physicians can monitor vitals, images, and notes from the related subjects; the system automatically detects anomalies and visually alerts the physicians. . . . .                                     | 32 |
| 3.1 | Different types of probe shapes for ultrasound devices, each probe corresponds to a type of exam with a designed anatomical target . . .                                                                                                                                        | 36 |
| 3.2 | A modern example of an ultrasound device, its structure resembles that of a modern notebook, each key in the keyboard represents a setting, like Gain, Depth, etc. . . . .                                                                                                      | 38 |
| 3.3 | Example of direct and indirect piezoelectric effect. The stress material produces a voltage in the indirect effect, while the powered material generates stress in the direct effect, as shown in the figure: left (indirect), right (direct). . . . .                          | 39 |
| 3.4 | Representation of beamforming, the reflected soundwave from a generic focal point impacts the transducer back, transducers that, in function of the position, apply a series of corrections to correctly sum the multiple parts of the soundwave into a cleaned signal. . . . . | 41 |
| 3.5 | Visualisation of the envelope detection process. The oscillating raw RF signal (A-Scan) is demodulated to obtain the signal magnitude, which serves as the basis for the displayable image. . . . .                                                                             | 43 |
| 3.6 | Conceptual signal smoothness steps to reduce possible noises . . . . .                                                                                                                                                                                                          | 44 |

|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |    |
|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.7  | Example of the point-of-view of a convex probe exam (left) and a linear probe exam (right) . . . . .                                                                                                                                                                                                                                                                                                                                                   | 45 |
| 3.8  | Effect on the depth parameter on the image sections and the correct identification of anatomically correct markers . . . . .                                                                                                                                                                                                                                                                                                                           | 46 |
| 3.9  | Effect on the gain parameter on the image resolution and the correct identification of anatomical correct markers . . . . .                                                                                                                                                                                                                                                                                                                            | 46 |
| 3.10 | The image contains two examples of ultrasound artefacts, the comet tail and the shadowing. It can be seen that the two of them work in complementary ways: the first fills with erroneous information, while the latter obscures the image content. . . . .                                                                                                                                                                                            | 47 |
| 3.11 | A comparison of different ultrasound images acquired and saved through different devices. . . . .                                                                                                                                                                                                                                                                                                                                                      | 48 |
| 3.12 | Manifold dimensionality collapse, the Artefacts expand the input space into a high-variance cloud (red points). The removal forces the data distribution to collapse onto the lower-dimensional manifold, making clinical details easier to extract. . . . .                                                                                                                                                                                           | 51 |
| 3.13 | In this side-by-side comparison, the chosen filter perfectly identifies the area of interest in the image . . . . .                                                                                                                                                                                                                                                                                                                                    | 52 |
| 3.14 | The contour detection algorithm employed in this instance exhibits excessive sensitivity to subtle pixel intensity gradients between the region of interest and the background. This results in a substantial distortion of the information content, which is methodologically unacceptable. In medical imaging, the erroneous exclusion of data or the misinterpretation of empty space constitutes a significant source of diagnostic error. . . . . | 53 |
| 3.15 | Example of the use of OCR detection, the filter, even if really fast, needs a more complex characterisation to work on the given dataset; the methods partially work, recognising only the letter 'P' . . . . .                                                                                                                                                                                                                                        | 53 |
| 3.16 | The picture shows that the DL-based OCR algorithm has a very high detection precision and rarely needs excessive fine-tuning to work; the image, in fact, is just the result of the application of the OCR-Keras pre-trained pipeline over the two analysed datasets . . . . .                                                                                                                                                                         | 54 |
| 3.17 | The main architecture used for U-net, the encoder and decoder architectures are mirrored structures connected by the skip connection to grant residual learning capacity at the network . . . . .                                                                                                                                                                                                                                                      | 61 |
| 3.18 | A visual transformer uses the same attention mechanism from the basic transformer architecture; the innovation lies in the catching mechanism for the input image . . . . .                                                                                                                                                                                                                                                                            | 62 |
| 3.19 | The ensemble complete pipeline, after the training of the weak model, it is frozen and ensembles the feature extraction model of the best two. The final linear combination layer is then trained on the same task as the previous one, yielding a more accurate, feature-aware classification output. . . . .                                                                                                                                         | 64 |
| 3.20 | Architectural composition of the inception module at the base of Googlenet, the naive version is the same but without the dimensional operation block on a convolutional kernel bigger than 1x1. . . . .                                                                                                                                                                                                                                               | 68 |

|      |                                                                                                                                                                                                                                                                                                                                                                         |     |
|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.21 | An example of correct classification of the trained ResNet-256 network; even if the accuracy is high, the model struggles to focus on meaningful detail and depict an area of interest. . . . .                                                                                                                                                                         | 69  |
| 3.22 | Healthy (first column), Pneumonia (second column), COVID (third column), Healthy samples from the dataset and their saliency map. Red (higher) to blue (lower) scale. . . . .                                                                                                                                                                                           | 71  |
| 3.23 | Classifications of a LUS image: in the middle, the input image; on the left, the focus of a wrong classification; on the right, the focus of the correct classification obtained by the classifier achieving 100% accuracy. Red (higher) to blue (lower) scale; (real Class -> error Class): in order Covid -> Pneumonia, Pneumonia -> Covid, Healthy ->Covid . . . . . | 72  |
| 3.24 | A LUS correctly classified from the class 1, the single B-Lines is detected by the model and almost entirely segmented . . . . .                                                                                                                                                                                                                                        | 76  |
| 3.25 | A LUS correctly classified from the class 2, the B-Lines are detected by the model and almost entirely segmented, in the wider area where they can be spotted. . . . .                                                                                                                                                                                                  | 76  |
| 3.26 | This image was originally mislabelled. The original label II was erroneous since there is no presence of B-Lines; the ensemble model at first misclassified this as a label 0, correcting the mistake. . . . .                                                                                                                                                          | 77  |
| 3.27 | Final proposed pipeline for POCUS analysis . . . . .                                                                                                                                                                                                                                                                                                                    | 78  |
| 4.1  | Example of good prototype identification using the Batch-CAM prototype training approach . . . . .                                                                                                                                                                                                                                                                      | 94  |
| 4.2  | Example of good prototype identification using the Batch-CAM prototype training approach with the geometric aware network . . . . .                                                                                                                                                                                                                                     | 95  |
| 5.1  | The comparison of the map of the trained network with the experiment with the proposed combination of EWC and Prototype regularisation, in this case, the extracted Grad-CAM seems to improve in analysing the detail, and even the stability is fairly high (91%). . . .                                                                                               | 104 |
| 5.2  | The comparison of the map of the original trained network with the experiment with only EWC regularisation, in this case, the Grad-CAM is very similar. . . . .                                                                                                                                                                                                         | 104 |
| 6.1  | Proposed Advanced Ultrasound Analysis Pipeline. . . . .                                                                                                                                                                                                                                                                                                                 | 111 |





# List of Tables

|     |                                                                                                                             |     |
|-----|-----------------------------------------------------------------------------------------------------------------------------|-----|
| 2.1 | Pros and Cons of Lung Ultrasound for SARS-CoV-2 Assessment . . .                                                            | 30  |
| 3.1 | Mapping between pixel colour space, beamforming signal and anatomical interpretation [76] . . . . .                         | 44  |
| 3.2 | Comparison of POCUS and CT/MRI Modalities . . . . .                                                                         | 66  |
| 3.3 | Preliminary Benchmarking Performance of Classical Architectures on LUS Data . . . . .                                       | 69  |
| 3.4 | Dataset Class Distribution . . . . .                                                                                        | 70  |
| 3.5 | Comparative Analysis with SOTA . . . . .                                                                                    | 71  |
| 3.6 | Dataset Label Distribution . . . . .                                                                                        | 73  |
| 3.7 | Performance Disparity: Weak Learners vs. Adaptive Ensemble (TiAsisto Dataset) . . . . .                                     | 75  |
| 3.8 | Detailed Classification Report (Best Ensemble Model) . . . . .                                                              | 75  |
| 4.1 | Comparative Taxonomy of AI Trust Schools . . . . .                                                                          | 85  |
| 4.2 | Metric (Averaged over 6 seeds) - MNIST . . . . .                                                                            | 88  |
| 4.3 | Model Performance on FMNIST . . . . .                                                                                       | 89  |
| 4.4 | Evolution of Grad-CAM Generation . . . . .                                                                                  | 90  |
| 4.5 | Peak Model Performance Across Architectures . . . . .                                                                       | 91  |
| 4.6 | Test Performance on Lung Ultrasound Dataset . . . . .                                                                       | 93  |
| 5.1 | DL model results: R2 metric for performance evaluation across varying architectures. . . . .                                | 99  |
| 5.2 | Comparison of model performance: The 2023 winning model vs. adaptive ensemble adjustments for the PT2024 challenge. . . . . | 99  |
| 5.3 | Comparative Analysis of CL Strategies. Old Train memory retention, Old Test stability, and New Test plasticity. . . . .     | 103 |
| A.1 | Comparison of computational complexity. $N$ is batch size, $K$ is number of classes in batch. . . . .                       | 135 |



# Bibliography

- [1] N. Aijaz, H. Lan, T. Raza, M. Yaqub, R. Iqbal, and M. S. Pathan. Artificial intelligence in agriculture: Advancing crop productivity and sustainability. *Journal of Agriculture and Food Research*, page 101762, 2025.
- [2] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
- [3] B. Antonio, D. Moroni, and M. Martinelli. Efficient adaptive ensembling for image classification. *Expert Systems*, 2022.
- [4] J. Bell, L. Quarantiello, E. N. Coleman, L. Li, M. Li, M. Madeddu, E. Piccoli, and V. Lomonaco. The future of continual learning in the era of foundation models: Three key directions. *arXiv preprint arXiv:2506.03320*, 2025.
- [5] L. Benos, A. C. Tagarakis, G. Dolias, R. Berruto, D. Kateris, and D. Bochtis. Machine learning in agriculture: A comprehensive updated review. *Sensors*, 21(11):3758, 2021. doi: 10.3390/s21113758.
- [6] H. Boche, A. Fono, and G. Kutyniok. Mathematical algorithm design for deep learning under societal and judicial constraints: The algorithmic transparency requirement. *Applied and Computational Harmonic Analysis*, 77:101763, 2025.
- [7] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5):1719–1778, 2023. ISSN 1573-756X. doi: 10.1007/s10618-023-00933-9. URL <https://doi.org/10.1007/s10618-023-00933-9>.
- [8] R. Bonney, C. B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K. V. Rosenberg, and J. Shirk. Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, 2009. doi: 10.1525/bio.2009.59.11.9.
- [9] J. Born, G. Brandle, M. Cossio, M. Disdier, J. Goulet, J. Roulin, and N. Wiedemann. Accelerating detection of lung pathologies with explainable ultrasound image analysis. *Applied Sciences*, 10(19):6755, 2020.
- [10] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101—mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014.

- [11] L. Bottou. Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics Paris France, August 22-27, 2010 Keynote, Invited and Contributed Papers, pages 177–186. Springer, 2010.
- [12] G. E. Box and D. A. Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association*, 65(332):1509–1526, 1970.
- [13] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
- [14] A. Bruno, G. Ignesti, and M. Martinelli. Explaining ensemble models for lung ultrasound classification. *ERCIM News*, 2023(134):16–18, 2023.
- [15] A. Bruno, G. Ignesti, O. Salvetti, D. Moroni, and M. Martinelli. Efficient lung ultrasound classification. *Bioengineering*, 10(5):555, 2023.
- [16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- [17] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.
- [18] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [19] F. Chollet et al. Keras. <https://github.com/fchollet/keras>, 2015. GitHub repository.
- [20] F. Cozim Melges, R. Ripoll-Bosch, G. Veen, P. Oggiano, F. Bianchi, W. Putten, and H. Zanten. Farming practices to enhance biodiversity across biomes: a systematic review. *npj Biodiversity*, 2024. doi: 10.1038/s44185-023-00034-2.
- [21] J. Curie and P. Curie. Développement par compression de l'électricité polaire dans les cristaux hémihédres à faces inclinées. *Bulletin de la Société Minéralogique de France*, 3(4):90–93, 1880.
- [22] R. Dainelli, A. Bruno, M. Martinelli, D. Moroni, L. Rocchi, S. Morelli, E. Ferrari, M. Silvestri, S. Agostinelli, P. La Cava, et al. Granoscan: an ai-powered mobile app for in-field identification of biotic threats of wheat. *Frontiers in Plant Science*, 2024. doi: 10.3389/fpls.2024.1298791.
- [23] S. Daoudagh, G. Ignesti, D. Moroni, L. Sebastiani, and P. Paradisi. Assessment of dance movement therapy outcomes: A preliminary proposal. In *International Conference on Computer-Human Interaction Research and Applications*, pages 382–395. Springer, 2024.

- [24] J. Davis, S. Smith, and R. Johnson. Integrating artificial intelligence with portable ultrasound devices for advanced point-of-care diagnostics. *Journal of Mobile Technology in Medicine*, 12(1):15–22, 2023. doi: 10.3390/diagnostics14151669.
- [25] M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(7):3366–3385, 2021.
- [26] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.
- [27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [28] T. S. Desser and R. B. Jeffrey. Tissue harmonic imaging techniques: physical principles and clinical applications. In *Seminars in Ultrasound, CT and MRI*, volume 22, pages 1–10. Elsevier, 2001.
- [29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [30] T. Fan, Y. Kang, G. Ma, W. Chen, W. Wei, L. Fan, and Q. Yang. Fate-llm: A industrial grade federated learning framework for large language models. *arXiv preprint arXiv:2310.10049*, 2023.
- [31] M. Farajtabar, N. Azizan, A. Mott, and A. Li. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pages 3762–3773. PMLR, 2020.
- [32] M. K. Feldman, S. Katyal, and M. S. Blackwood. Us artifacts. *Radiographics*, 29(4):1179–1189, 2009.
- [33] A. Fono, M. Singh, E. Araya, P. C. Petersen, H. Boche, and G. Kutyniok. Sustainable ai: Mathematical foundations of spiking neural networks. *arXiv preprint arXiv:2503.02013*, 2025.
- [34] A. Genova and S. Lombardini. General practitioners in front of covid-19: Italy in european comparative perspective. *Frontiers in Sociology*, 9:1365517, 2024.
- [35] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feed-forward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

- [36] A. Goffi, A. Zanforlin, T. Perrone, V. Moscariello, R. Copetti, M. B. Stone, and G. Soldati. Novel concepts in lung ultrasound diagnostic protocols in acute respiratory failure. *Critical Ultrasound Journal*, 5(1):1–7, 2013. doi: 10.1186/2036-7902-5-S1-S9.
- [37] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 2021. doi: 10.48550/arXiv.2106.11959.
- [38] L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.
- [39] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First conference on language modeling*, 2024.
- [40] D. Gunning and D. Aha. Darpas explainable artificial intelligence (xai) program. *AI magazine*, 40(2):44–58, 2019.
- [41] T. L. Hayes and C. Kanan. Lifelong machine learning with deep streaming linear discriminant analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 220–221, 2020.
- [42] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [43] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of ICLR*, 2017.
- [44] High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy ai. Technical report, European Commission, Brussels, 2019. URL <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.
- [45] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [46] J. Hofmanninger, M. Perkonigg, J. A. Brink, O. Pianykh, C. Herold, and G. Langs. Dynamic memory to alleviate catastrophic forgetting in continuous learning settings. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 359–368. Springer, 2020.
- [47] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
- [48] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

- [49] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [50] H. Huang, L. Zhang, C. Sun, R. Fang, X. Yuan, and D. Wu. Distributed pruning towards tiny neural networks in federated learning. In *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*, pages 190–201. IEEE, 2023.
- [51] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):136, 2020.
- [52] F. Huszar et al. On the computation of the fisher information in continual learning. *arXiv preprint arXiv:2502.11756*, 2025.
- [53] G. Ignesti, A. Bruno, C. Deri, G. D’Angelo, L. Bastiani, L. Pratali, S. Memmini, D. Cicalini, A. Dini, G. Galesi, F. Pardini, M. Tampucci, A. Benassi, O. Salvetti, D. Moroni, and M. Martinelli. An intelligent platform of services based on multimedia understanding and telehealth for supporting the management of sars-cov-2 multi-pathological patients. In *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 553–560. IEEE, 2022.
- [54] G. Ignesti, C. Deri, G. D’Angelo, L. Pratali, A. Bruno, A. Benassi, O. Salvetti, D. Moroni, and M. Martinelli. Deep learning methods for point-of-care ultrasound examination. In *2023 17th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 435–440. IEEE, 2023.
- [55] G. Ignesti, M. Martinelli, and D. Moroni. Youve got the wrong number: evaluating deep learning training paradigms using handwritten digit recognition data. *Pattern Recognition and Image Analysis*, 34(4):978–983, 2024.
- [56] G. Ignesti, M. Martinelli, and D. Moroni. Youve got the wrong outfit: Evaluating deep learning paradigms on digit and fashion recognition. In *International Conference on Pattern Recognition*, pages 202–215. Springer, 2024.
- [57] G. Ignesti, D. Moroni, and M. Martinelli. Plant-traits: how citizen science and artificial intelligence can impact natural science. In *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2024.
- [58] G. Ignesti, D. Moroni, and M. Martinelli. Towards the actual deployment of robust, adaptable, and maintainable ai models for sustainable agriculture. *Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, page 33, 2024.
- [59] G. Ignesti, D. Moroni, and M. Martinelli. Batch-cam: Introduction to better reasoning in convolutional deep learning models. *arXiv preprint arXiv:2510.00664*, 2025.

- [60] A. Ivanković, I. Lukić, et al. Moving towards society 5.0 and medicine 5.0: The pivotal role of human-centric ai in healthcare. *Annals of Biomedical Engineering*, 51:1–5, 2023. doi: 10.1007/s10439-022-03083-x.
- [61] A. Jacovi and Y. Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*, 2020.
- [62] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *Advances in neural information processing systems*, volume 28, 2015.
- [63] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, pages 709–727. Springer, 2022.
- [64] A. Johnston, E. Matechou, and E. B. Dennis. Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, 2023. doi: 10.1111/2041-210X.13834.
- [65] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 2021. doi: 10.1038/s41586-021-03819-2.
- [66] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima Jr, J. Mancuso, F. Jungmann, M.-M. Steinborn, et al. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence*, 3(6):473–484, 2021.
- [67] A. Kamilaris and F. X. Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90, 2018. doi: 10.1016/j.compag.2018.02.016.
- [68] T. Kattenborn. Planttraits2023, 2023. URL <https://kaggle.com/competitions/planttraits2023>.
- [69] S. Kierner, J. Kucharski, and Z. Kierner. Taxonomy of hybrid architectures involving rule-based reasoning and machine learning in clinical decision systems: A scoping review. *Journal of biomedical informatics*, pages 104428–104428, 2023.
- [70] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.
- [71] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.



- [72] N. Koutsoubis, Y. Yilmaz, R. P. Ramachandran, M. Schabath, and G. Rasool. Privacy preserving federated learning in medical imaging with uncertainty estimation. *arXiv preprint arXiv:2406.12815*, 2024.
- [73] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical Report, University of Toronto, 2009.
- [74] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [75] M. Lafond, A. Payne, and C. Lafon. Therapeutic ultrasound transducer technology and monitoring techniques: a review with clinical examples. *International Journal of Hyperthermia*, 41(1):2389288, 2024.
- [76] B. K. Lal, R. W. Hobson II, P. J. Pappas, R. Kubicka, M. Hameed, E. Y. Chakhtura, Z. Jamil, F. T. Padberg Jr, P. B. Haser, and W. N. Durán. Pixel distribution analysis of b-mode ultrasound scan images predicts histologic features of atherosclerotic carotid plaques. *Journal of vascular surgery*, 35(6):1210–1217, 2002.
- [77] C. Lammie, J. Büchel, A. Vasilopoulos, M. Le Gallo, and A. Sebastian. The inherent adversarial robustness of analog in-memory computing. *Nature Communications*, 16(1):1–15, 2025.
- [78] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.
- [79] Y. Li, R. Zhang, J. Lu, and E. Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020.
- [80] D. A. Lichtenstein and G. A. Mezière. Relevance of lung ultrasound in the diagnosis of acute respiratory failure: the blue protocol. *Chest*, 134(1):117–125, 2008.
- [81] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.
- [82] S. Liu, Y. Wang, X. Yang, B. Lei, L. Liu, S. X. Li, D. Ni, and T. Wang. Deep learning in medical ultrasound analysis: A review. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 66(2):261–275, 2019. doi: 10.1109/TUFFC.2018.2889028.
- [83] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li. Trustworthy llms: a survey and guideline for evaluating large language models’ alignment. *arXiv preprint arXiv:2308.05374*, 2023.

- [84] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11976–11986, 2022.
- [85] E. López-Guillén, I. Herrera, B. Bensid, C. Gómez-Bellver, N. Ibáñez, P. Jiménez-Mejías, M. Mairal, L. Mena-García, N. Nualart, M. Utjés-Mascó, et al. Strengths and challenges of using inaturalist in plant research with focus on data quality. *Diversity*, 2024. doi: 10.3390/d16010042.
- [86] A. Lu, H. Yuan, T. Feng, and Y. Sun. Rethinking the stability-plasticity trade-off in continual learning from an architectural perspective. *arXiv preprint arXiv:2506.03951*, 2025.
- [87] G. Maccani, M. Goossensen, V. Righi, J. Creus, and M. Balestrini. Scaling up citizen science: what are the factors associated with increased reach and how to lever them to achieve impact. JRC: Ispra (VA), Italy, 2020. doi: 10.2760/00926.
- [88] N. Maslej, L. Fattorini, R. Perrault, Y. Gil, V. Parli, N. Kariuki, E. Capstick, A. Reuel, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, J. C. Niebles, Y. Shoham, R. Wald, T. Walsh, A. Hamrah, L. Santarlasci, J. B. Lotufo, A. Rome, A. Shi, and S. Oak. Artificial intelligence index report 2025, 2025. URL <https://arxiv.org/abs/2504.07139>.
- [89] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [90] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.
- [91] C. L. Moore and J. A. Copel. Point-of-care ultrasonography. *New England Journal of Medicine*, 364(8):749–757, 2011.
- [92] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 151–159, 2020. doi: 10.1145/3368555.3384468.
- [93] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, H. Daniel, B. Adila, M. Blyo, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [94] P. K. Panda and B. Sahoo. Pzt to lead free piezo ceramics: a review. *Ferroelectrics*, 474(1):128–143, 2015.
- [95] S. Pappa, V. Ntella, T. Giannakas, V. G. Giannakoulis, E. Papoutsis, and P. Katsaounou. Prevalence of depression, anxiety, and insomnia among health-care workers during the covid-19 pandemic: A systematic review and meta-analysis. *Brain, behavior, and immunity*, 88:901–907, 2020.

- [96] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.
- [97] M. Perkonigg, J. Hofmanninger, C. J. Herold, J. A. Brink, O. Pianykh, H. Prosch, and G. Langs. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. *Nature communications*, 12(1):5678, 2021.
- [98] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on pattern analysis and machine intelligence*, 12(7):629–639, 1990.
- [99] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.
- [100] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. URL [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf).
- [101] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017.
- [102] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.
- [103] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- [104] L. Rieger, C. Singh, W. Murdoch, and B. Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pages 8116–8126. PMLR, 2020.
- [105] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.
- [106] 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*, pages 234–241. Springer, 2015.
- [107] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation, 2015.

- [108] A. S. Ross, M. C. Hughes, and F. Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2662–2670, 2017.
- [109] C. Rudin. Stop explaining black boxes and start using interpretable models for high-stakes decisions and understanding. *Nature Machine Intelligence*, 1(5):206–215, 2019.
- [110] C. Rumack, S. Wilson, J. Charboneau, and D. Levine. Diagnostic Ultrasound. Number v. 1 in *Diagnostic Ultrasound*. Elsevier/Mosby, 2011. ISBN 9780323053976. URL <https://books.google.it/books?id=ekaISQAACAAJ>.
- [111] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [112] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1): 61–80, 2009. doi: 10.1109/TNN.2008.2005605.
- [113] C. Schiller, S. Schmidtlein, C. Boonman, A. Moreno-Martínez, and T. Kattenborn. Deep learning and citizen science enable automated plant trait predictions from photographs. *Scientific Reports*, 2021. doi: 10.1038/s41598-021-95616-0.
- [114] R. Scholz et al. Imbalance-aware loss functions improve medical image classification. In *Proceedings of Machine Learning Research*, volume 250, pages 1–15, 2024. Benchmarks standard Cross-Entropy loss in medical contexts.
- [115] 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. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [116] M. P. Sendak, J. DArcy, S. Kashyap, M. Gao, M. Nichols, K. Corey, W. Ratliff, and S. Balu. A path for translation of machine learning products into health-care delivery. *EMJ Innov*, 10:19–00172, 2020.
- [117] A. Sewify, M. Antico, L. Alzubaidi, H. A. Alwzway, J. Roots, P. Pivonka, and D. Fontanarosa. Systematic review of commercially available clinical cmut-based systems for use in medical ultrasound imaging: Products, applications, and performance. *Sensors (Basel, Switzerland)*, 25(7):2245, 2025.
- [118] M. I. Sharif, M. Mehmood, M. P. Uddin, et al. Federated learning for analysis of medical images: a survey. *Journal of Computer Science*, 20(12):1610–1621, 2024.
- [119] J. Shaver. The state of telehealth before and after the covid-19 pandemic. *Primary care*, 49(4):517, 2022.

- [120] J. Silvertown. A new dawn for citizen science. *Trends in ecology & evolution*, 2009. doi: 10.1016/j.tree.2009.03.017.
- [121] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khali-dov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al. Dinov3. arXiv preprint arXiv:2508.10104, 2025.
- [122] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL <https://arxiv.org/abs/1409.1556>.
- [123] G. Soldati, A. Smargiassi, R. Inchingolo, D. Buonsenso, T. Perrone, C. Briganti, S. Perlini, E. Torri, A. Mariani, E. Mossello, et al. Proposal for international standardization of the use of lung ultrasound for patients with covid-19: a simple, quantitative, reproducible method. *Journal of ultrasound in medicine*, 39(7):1413–1419, 2020.
- [124] S. Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, 30(1): 32–46, 1985.
- [125] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [126] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 2019. doi: 10.48550/arXiv.1905.11946.
- [127] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method, 2000. URL <https://arxiv.org/abs/physics/0004057>.
- [128] E. J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019. doi: 10.1038/s41591-018-0300-7.
- [129] E. J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019. doi: 10.1038/s41591-018-0300-7.
- [130] A. Usmanova, F. Portet, P. Lalanda, and G. Vega. Federated continual learning: Concepts, challenges, and solutions. arXiv preprint arXiv:2502.07059, 2025.
- [131] A. Usmanova, F. Portet, P. Lalanda, and G. Vega. Federated continual learning: Concepts, challenges, and solutions. arXiv preprint arXiv:2502.07059, 2025.
- [132] R. J. van Sloun and L. Demi. Automated quantification of b-lines on lung ultrasonography by using a deep learning method. *IEEE Transactions on*

- Ultrasonics, Ferroelectrics, and Frequency Control, 67(2):296–305, 2019. doi: 10.1109/TUFFC.2019.2947658.
- [133] A. J. Van Strien, J. S. Van Zweden, L. B. Sparrius, and B. Odé. Improving citizen science data for long-term monitoring of plant species in the netherlands. *Biodiversity and Conservation*, 2022. doi: 10.1007/s10531-022-02457-y.
- [134] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [135] G. Volpicelli and L. Gargani. Sonographic signs and patterns of covid-19 pneumonia. *The Ultrasound Journal*, 12(1):1–3, 2020.
- [136] G. Volpicelli, M. Elbarbary, M. Blaivas, D. A. Lichtenstein, G. Mathis, A. W. Kirkpatrick, L. Melniker, L. Gargani, V. E. Noble, G. Via, et al. International evidence-based recommendations for point-of-care lung ultrasound. *Intensive Care Medicine*, 38(4):577–591, 2012. doi: 10.1007/s00134-012-2513-4.
- [137] C. Wang, H. M. Liao, I. Yeh, Y. Wu, P. Chen, and J. Hsieh. Cspnet: A new backbone that can enhance learning capability of CNN. *CoRR*, abs/1911.11929, 2019. URL <http://arxiv.org/abs/1911.11929>.
- [138] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023.
- [139] J. Wosik, M. Fudim, B. Cameron, Z. F. Gellad, A. Cho, D. Phinney, S. Curtis, M. Roman, E. G. Poon, J. Ferranti, et al. Telehealth transformation: Covid-19 and the rise of virtual care. *Journal of the American Medical Informatics Association*, 27(6):957–962, 2020.
- [140] M. Xu, S. Yoon, Y. Jeong, and D. S. Park. Transfer learning for versatile plant disease recognition with limited data. *Frontiers in Plant Science*, 13: 4506, 2022.
- [141] Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu, and L. Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023.
- [142] A. W. K. Yeung, A. Torkamani, A. J. Butte, B. S. Glicksberg, B. Schuller, B. Rodriguez, D. S. Ting, D. Bates, E. Schaden, H. Peng, et al. The promise of digital healthcare technologies. *Frontiers in Public Health*, 11:1196596, 2023.
- [143] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 2019. doi: 10.48550/arXiv.1911.02685.

# Appendix A

## Supplementary Material for Chapter 4

### A.1 Proofs

This section provides a formal analysis of the Batch-CAM training paradigm introduced by Ignesti et al. We derive the mathematical proof for the parallelization efficiency of the Batch-CAM Prototype Loss compared to standard per-sample Grad-CAM approaches. We demonstrate how utilizing class-wise aggregation reduces the computational complexity of the explanation mechanism within the training loop.

#### A.1.1 Mathematical Preliminaries

Let  $B$  be a batch of  $N$  input images. Let  $f(\cdot)$  be the convolutional neural network with a feature extraction backbone and a fully connected classifier.

- $A \in \mathbb{R}^{N \times D \times H \times W}$ : The feature maps from the final convolutional layer for the batch, where  $D$  is the channel depth.
- $S \in \mathbb{R}^{N \times C}$ : The output scores (logits) for  $C$  classes.
- $y_i$ : The ground truth class index for sample  $i$ .

#### A.1.2 Standard Per-Sample Grad-CAM Analysis

In the standard (or "Old Architecture") approach, the generation of CAMs is iterative and hook-based.

#### A.1.3 Computational Steps

For each sample  $i \in \{1, \dots, N\}$ :

1. Forward Hook: Store  $A_i$ .
2. Backward Hook: Perform backpropagation on  $S_{i,y_i}$  to obtain gradients  $G_i = \nabla_{A_i} S_{i,y_i}$ .
3. Weight Computation:  $\alpha_{i,k} = \frac{1}{HW} \sum_{u,v} G_{i,k,u,v}$ .

4. Map Generation:  $M_i = \text{ReLU}(\sum_k \alpha_{i,k} A_{i,k})$ .

Complexity: This requires maintaining the computational graph for individual backward passes or looping through the batch, scaling linearly with batch size  $N$  for the gradient retrieval step.

#### A.1.4 Batch-CAM: Parallelisation and Aggregation Proof

The Batch-CAM approach leverages the linearity of the gradient operator and the statistical distribution of the batch to reduce complexity.

Step 1: Class-wise Grouping

Let  $K$  be the set of unique classes present in the current batch  $B$ . For each class  $c \in K$ , define the subset of indices  $B_c = \{i \mid y_i = c\}$  with cardinality  $N_c$ .

Step 2: Aggregated Feature Maps

Instead of processing maps individually, Batch-CAM computes the centroid of the feature space for class  $c$ :

$$\bar{A}_k^c = \frac{1}{N_c} \sum_{i \in B_c} A_{i,k}$$

This operation is a simple tensor reduction (mean) along the batch dimension for the sliced indices, which is highly optimized on GPUs.

Step 3: Aggregated Gradient Computation (The Proof)

The standard method computes  $G_i = \frac{\partial S_{i,c}}{\partial A_i}$ . Batch-CAM requires the average gradient for the class.

$$\bar{G}_k^c = \frac{1}{N_c} \sum_{i \in B_c} \frac{\partial S_{i,c}}{\partial A_{i,k}}$$

Proof of Vectorization Efficiency: By the linearity of the derivative, the sum of gradients is equal to the gradient of the sum. Therefore, we can express the numerator of the averaged gradient as:

$$\sum_{i \in B_c} \nabla_A S_{i,c} = \nabla_A \left( \sum_{i \in B_c} S_{i,c} \right)$$

This implies that instead of  $N_c$  separate backward calls, one can sum the scores of all images belonging to class  $c$  and perform a single gradient calculation with respect to the feature map tensor  $A$ . This utilizes the stateless ‘`torch.autograd.grad`’ on the aggregated scalar sum, significantly reducing overhead.



## Step 4: Batch-CAM Generation

The neuron importance weights  $\bar{\alpha}^c$  are computed via Global Average Pooling (GAP) on the aggregated gradients:

$$\bar{\alpha}_k^c = \frac{1}{H \times W} \sum_{u=1}^H \sum_{v=1}^W (\bar{G}_k^c)_{u,v}$$

The final Batch-CAM for class  $c$  is:

$$\bar{M}_c = \text{ReLU} \left( \sum_k \bar{\alpha}_k^c \bar{A}_k^c \right)$$

A comparison between the proposed algorithm and the original is depicted in Table A.1.

Table A.1: Comparison of computational complexity.  $N$  is batch size,  $K$  is number of classes in batch.

| Step                 | Standard Grad-CAM        | Batch-CAM                      |
|----------------------|--------------------------|--------------------------------|
| Gradient Calculation | $O(N)$ Backprops         | $O(K)$ Backprops ( $K \ll N$ ) |
| Feature Aggregation  | None                     | $O(N)$ (Vectorized Mean)       |
| CAM Generation       | $N$ Map Computations     | $K$ Map Computations           |
| Loss Computation     | $N$ Pairwise Comparisons | $K$ Pairwise Comparisons       |

## A.1.5 Algorithm Analysis

The algorithm described in the paper transitions from a stateful, hook-based architecture to a functional, stateless architecture, Algorithm 1. This structure ensures that the computationally expensive operations specifically gradient computation and map generation scale with the number of classes  $K$  rather than the batch size  $N$ .

Below is the formal definition of the optimized Batch-CAM processing logic using the proposed vectorization method.

The complexity reduction relies on the fact that the loop at line 2 iterates  $K$  times. In a standard mini-batch of size  $N = 64$  with  $K = 10$  classes, the number of backpropagation calls is reduced by a factor of  $N/K \approx 6.4$ . This structure ensures that the computationally expensive operations (gradient computation and map generation) scale with the number of classes  $K$  rather than the batch size  $N$ , providing the speed improvements noted in the experimental results.

---

Algorithm 1: Batch-CAM Processing Logic (Parallelized)

---

Input: Logits  $L$ , Feature Maps  $A$ , Labels  $Y$   
 Output: Accumulated Loss with Batch-CAM penalty  
 Identify unique classes  $K \leftarrow \text{unique}(Y)$   
 for  $c \in K$  do  
      $Indices_c \leftarrow \{i \mid Y_i = c\}$   
     // Step 2: Aggregated Features (Vectorized Mean)  
      $\bar{A}^c \leftarrow \text{mean}(A[Indices_c])$   
     // Step 3: Aggregated Gradient (Single Backward Pass)  
      $\bar{S}^c \leftarrow \text{sum}(L[Indices_c, c])$   
      $\bar{G}^c \leftarrow \text{autograd}(\bar{S}^c, A[Indices_c])$   
     // Step 4: Generate Class-Level Explanation  
      $\bar{M}^c \leftarrow \text{GenerateCAM}(\bar{A}^c, \bar{G}^c)$   
     // Compute Prototype Loss  
      $Loss \leftarrow Loss + \mathcal{L}_{metric}(\bar{M}^c, Prototype_c)$   
 end  
 return  $Loss$

---