

Universidade do Minho

Statistical and Geometrical priors for Deep Learning with applications in Biomedical modalities

Miguel L. Martins

Doctoral Program in Computer Science - MAP joint programme

Faculty of Sciences of the University of Porto, Faculty of Engineering of the University of Porto, University of Aveiro and University of Minho

2026

PhD

3rd
CYCLE

FCUP
FEUP
UA
UM
2026

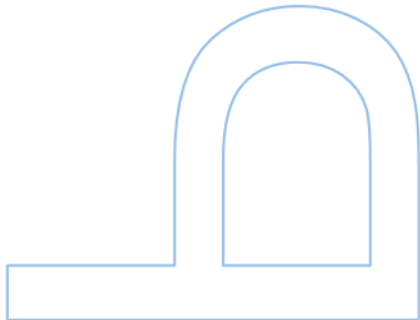
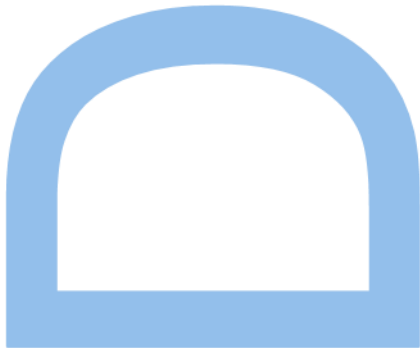


Statistical and Geometrical priors for Deep Learning
with applications in Biomedical modalities

Miguel L. Martins



Universidade do Minho



Statistical and Geometrical priors for Deep Learning with applications in Biomedical modalities

Miguel L. Martins

Thesis carried out as part of the Doctoral Program in Computer Science - MAP joint programme
Department of Computer Science
2026

Supervisor

Francesco Renna, Assistant Professor, Faculty of Sciences of the University of Porto

Co-supervisor

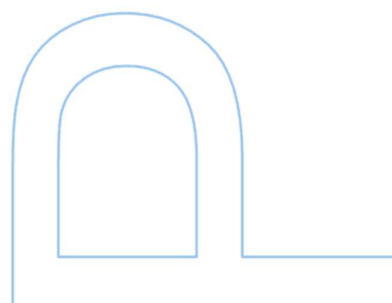
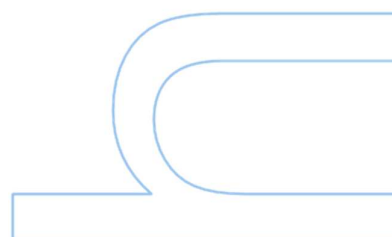
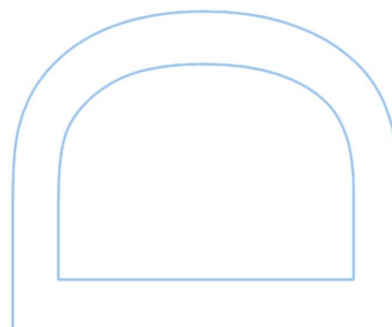
Miguel T. Coimbra, Full Professor, Faculty of Sciences of the University of Porto

All the corrections determined by
the jury, and only those, were
made.

The Supervisor,

Assinado por: **Francesco Renna**
Num. de Identificação: PASIT-YB2887093
Data: 2026.08.13 22:30:36 +0100

Porto, ____/____/____



UNIVERSIDADE DO PORTO

DOCTORAL THESIS

Statistical and Geometrical Priors for Deep Learning with Applications to Biomedical Modalities

Author:

Miguel L. MARTINS

Supervisor:

Francesco RENNA

Co-supervisor:

Miguel COIMBRA

*A thesis submitted in fulfilment of the requirements
for the degree of MAP-i Joint Doctoral Programme in Computer Science
at the*

Faculdade de Ciências da Universidade do Porto
Departamento de Ciências de Computadores

August 5, 2026

“What your son needs is tender, loving, ... medication. Heavy medication.”

Dr. Janus to Butters’ parents, *South Park*, Season 15, Episode 6. Written by Trey Parker.

Acknowledgements

To my family for its unwavering support. To my friends and loved ones, without explicit enumeration.

To my advisors for their patience and friendship, and for the scientific freedom they gave me.

To my hosts at CentraleSupélec.

In Nelson Piquet's fashion: to myself, since I was the one writing this thing.

To my cats, Santino and Sasha.



My cats: Sasha on top, and Santino, in the bottom. Sasha passed away during my PhD. I will never forget him.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto

Departamento de Ciências de Computadores

MAP-i Joint Doctoral Programme in Computer Science

Statistical and Geometrical Priors for Deep Learning with Applications to Biomedical Modalities

by Miguel L. MARTINS

The advent of Deep Learning has fundamentally transformed the landscape of artificial intelligence, offering universal function approximation capabilities that solve complex tasks with unprecedented accuracy. However, this engineering revolution—driven largely by scale and gradient descent—faces critical limitations regarding interpretability, stability, and generalisation, particularly in high-stakes environments such as those involving biomedical data. This dissertation focuses on the role of domain knowledge as a necessary condition for overcoming the fragility of fully data-driven methods.

We make the learning process itself depend on explicit parameterisations that encode different statistical or geometric priors for the domain, thus defining the so-called hybrid or model-based deep learning methods. We first address the challenge of pathological detection in medical imaging by leveraging principles from (multi)fractal geometry. We articulate the self-seeding and cascading nature of pathological tissue abnormalities, showing that optimisation informed by these geometrical priors can guide networks to capture textural regularities that purely convolutional approaches may overlook. We also analyse regularities across time, and propose a statistical prior for sequential data. By jointly optimising a deep neural network with a Markovian prior, we demonstrate that this yields superior performance and robustness compared to unconstrained sequence modelling, highlighting its effectiveness in modeling the rhythmic nature of the cardiac cycle for the task of fundamental heart sound segmentation.

Finally, we investigate the fundamental trade-offs within contemporary Self-Supervised Learning (SSL), a paradigm often touted as agnostic to downstream tasks. We demonstrate that models designed to maximise invariance to subject transformations are subject

to inherent bounds: the more degrees of variation a model captures, the more unstable the representation becomes regarding downstream generalisation. Ultimately, this work provides both methodological contributions from a Computer Science and Machine Learning perspective, illustrated by their impact on biomedical modalities.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto

Departamento de Ciências de Computadores

Doutoramento em Engenharia Informática (MAP-i)

Incorporação de Restrições Estatísticas e Geométricas em Aprendizagem Profunda com Aplicações em Modalidades Biomédicas

por Miguel L. MARTINS

A Aprendizagem Profunda (Deep Learning) transformou fundamentalmente o panorama da inteligência artificial, oferecendo capacidades de aproximação universal de funções que resolvem tarefas complexas com uma precisão sem precedentes. No entanto, esta revolução de engenharia — impulsionada largamente por avanços de infraestrutura e de métodos de optimização — enfrenta limitações críticas no que respeita à interpretabilidade, estabilidade e generalização, particularmente em ambientes de alto risco, como nas modalidades biomédicas. Esta dissertação foca-se no papel do conhecimento do domínio como condição necessária para superar a fragilidade dos métodos puramente baseados nos dados observados durante o treino.

Para este efeito, forçamos o processo de aprendizagem a depender de parametrizações explícitas que codificam diferentes restrições estatísticas ou geométricas do domínio, definindo, assim, os chamados métodos de aprendizagem profunda híbridos ou baseados em modelos *a priori*. Primeiro, abordamos o desafio da deteção patológica em imagiologia médica, tirando partido dos princípios da geometria (multi)fractal. Articulamos a natureza de auto-propagação em cascata das anomalias tecidulares patológicas, demonstrando que a otimização informada por estes parâmetros geométricos pode guiar as redes na captura de regularidades texturais que abordagens puramente convolucionais podem ignorar. Analisamos também regularidades ao longo do tempo e propomos um modelo

estatístico para dados sequenciais. Ao otimizar conjuntamente uma rede neuronal profunda com um *prior* Markoviano, demonstramos que tal resulta num desempenho e robustez superiores em comparação com a modelação de sequências sem restrições, destacando a sua eficácia em forçar a representação a depender da natureza rítmica do ciclo cardíaco para a tarefa de segmentação de sons cardíacos fundamentais.

Finalmente, investigamos os compromissos fundamentais no âmbito da Aprendizagem Auto-Supervisionada (Self-Supervised Learning - SSL) contemporânea, um paradigma frequentemente apontado como agnóstico em relação a tarefas a jusante. Demonstramos que os modelos desenhados para maximizar a invariância a transformações no domínio estão sujeitos a limites inerentes: quanto mais graus de variação um modelo destes captura, mais instável se torna a representação relativamente à generalização a jusante. Em última análise, este trabalho fornece contribuições metodológicas tanto da perspetiva da Ciência de Computadores como da Aprendizagem Computacional, ilustradas pelo seu impacto nas modalidades biomédicas.

Contents

Acknowledgements	v
Abstract	vii
Resumo	ix
Contents	xi
List of Figures	xv
1 Introduction	1
1.1 Thesis Overview	3
1.2 Notation	4
2 Deep Learning with prior knowledge	9
2.1 Background	9
2.1.1 Maximum Likelihood Estimation	10
2.1.2 Empirical Risk Minimisation (ERM)	10
2.2 The Necessity of Inductive Bias	11
2.2.1 Parametric Models (Hypothesis Space)	11
2.2.2 Structural Inductive Bias (Architecture)	12
2.2.3 Loss Function	12
2.2.4 Data Processing	13
2.3 Prior knowledge as an intervention on the Learning Triplet	13
2.3.1 Example: ResNet on ImageNet	14
2.4 Application: Upper gastrointestinal endoscopy	15
2.4.1 Deep Learning in Gastric Lesion Characterization	16
2.4.2 Characterisation of metaplasia for patient risk stratification	17
2.4.3 The GIM self-similarity hypothesis	19
2.4.4 Endoscopic grading of gastric intestinal metaplasia (EGGIM)	20
2.5 Data collection and annotation	21
2.5.1 Patch extraction and annotation	21
2.5.2 Configuration instantiation for patch GIM score prediction	21
2.6 Addressing probability shifts for cross-validation	22
2.7 Experiments	23
2.7.1 Validation of self-similarity hypothesis	23

2.7.2	Model evaluation over highly heterogeneous and limited patient data	24
2.7.3	Patient risk stratification validation	24
2.7.4	Discussion and conclusion	25
2.8	Chapter conclusion	27
2.8.1	Outcomes	27
3	Multifractal Recalibration	29
3.1	Informal introduction	30
3.2	Background	31
3.2.1	Lebesgue Measure	32
3.3	Objects with Lebesgue measure 0	32
3.4	Measuring sets of Lebesgue measure 0	33
3.4.1	Fractal sets and dimension	35
3.4.2	Multifractal formalism	37
3.4.3	Multiplicative cascades and $f(\alpha)$	38
3.5	Channel attention functions as inductive prior	39
3.5.1	Squeeze-and-Excite (SE) [92]	40
3.5.2	Spatial and Channel SE[93]	41
3.5.3	Style-based Recalibration (SRM) [94]	41
3.5.4	Frequency Channel Attention (FCA) [95]	41
3.6	Literature review	42
3.6.1	Fractals in classical computer vision	42
3.6.2	Fractals in Modern Computer Vision	43
3.6.3	Statistical Channel Attention Functions	44
3.7	Methods	45
3.7.1	Differentiable scaling exponent computation	46
3.7.2	Monofractal Recalibration	47
3.7.3	Multifractal recalibration	48
3.7.4	The Learning Triplet for Fractal Recalibration	50
3.8	Application: medical image segmentation	51
3.8.1	Multifractal analysis in Medical Imaging	52
3.8.2	Implementation	53
3.9	Experiments	56
3.9.1	Experimental setup	56
3.9.2	Results	58
3.9.3	Multifractal hyperparameter and aggregation analysis	60
3.9.4	Analysis of Global Channel Attention Functions	62
3.9.5	Discussion and Conclusions	65
3.10	Chapter conclusion	65
3.10.1	Outcomes	66
4	Markov-based Neural Networks	67
4.1	Markov and hidden Markov models	67
4.2	Markov-Based Neural Networks	69
4.2.1	Training	70
4.2.2	Parameter estimation	71
4.3	The Learning Triplet for Markov-Based Neural Networks	75

4.4	Markov models in the deep learning era	76
4.5	Application: heart sound segmentation	76
4.5.1	Literature review on heart sound segmentation	78
4.5.2	Implementation	79
4.5.3	Encoding domain knowledge in the Markov model	79
4.5.4	Numerical computation of forward-backwards algorithm	80
4.5.5	Preprocessing	81
4.5.6	Inference and post-processing	82
4.6	Experiments	82
4.6.1	Materials	83
	The PhysioNet 2016 Dataset (PhysioNet'16)	83
	The CirCor DigiScope Dataset (CirCor'22)	84
4.6.2	Metrics and evaluation	85
4.6.3	MNN baseline benchmark	85
4.6.4	Model performance comparison	87
4.6.5	Fine-tuning MNN on unseen data	89
4.6.6	Discussion and conclusion	93
4.7	Chapter conclusion	96
4.7.1	Outcomes	96
5	No free lunches for self-supervised learning	99
5.1	Generalised contrastive learning problems	100
5.2	No free lunches for Self-supervised Learning	103
5.3	Related work	107
5.3.1	Identifiability of self-supervised learning	108
5.3.2	Identifiability theory contradicts well-established contrastive learning design principles	108
5.3.3	Oversimplified effect of data augmentations in identifiability theory Instantiation of the Learning Triplet $\mathcal{L}$	108 109
5.4	Experiments	109
5.4.1	Numerical simulations	109
5.4.1.1	Inverting the wrong DGP leads to instability	110
5.4.1.2	On the relation of identifiability and best possible down- stream performance	112
5.4.2	Label flipping augmentations have a regularising effect	113
5.4.3	DisLib benchmark	114
5.4.3.1	There is a trade-off between identifiability and downstream performance	116
5.4.3.2	Miss-specified augmentations lead to instability	116
5.4.4	Relationship with Equivariant-SSL	117
5.5	Chapter conclusion	120
5.5.1	Outcomes	120
6	Conclusion	121
A	Deep Learning with prior knowledge	127
A.1	Extended literature review	127

A.1.1	Classical computer vision	127
B	Multifractal recalibration	133
B.1	Additional results for Multifractal formalism	133
B.2	Properties of dimension operator	133
B.3	Repeated iterated subdivision of measures	134
C	Markov-based neural networks	137
C.1	Experimental pipeline	137
C.2	Pre-processing	137
C.2.1	Denoising	138
	Wavelet denoising	139
	Universal thresholding	139
C.3	Feature extraction	143
C.3.1	Hilbert Envelopogram	143
C.3.2	Frequency space envelopes	145
C.3.3	Full set of envelopes	146
C.4	Markov likelihood computation limitations	146
D	No free lunches for self-supervised learning	149
D.1	Additional visualizations	149
	Bibliography	153

List of Figures

2.1	(Top) Two EGD frames of the antrum with EGGIM scores of 1 ($\leq 30\%$ GIM), and 2 ($\geq 30\%$ GIM). The dashed square indicates a illustrative example of a patch. Translucent arrows in the middle EGD frame indicate focal GIM. This example of EGGIM 1 is specially illustrative of the periodic and self-similar character of GIM distribution. Data adapted from [60]. (Bottom) A section of an EGD frame observed at different scales. Note how the textural attributes remain identical throughout.	19
2.2	Representative examples of EGD frames of healthy anatomical landmarks used in the EGGIM protocol.	20
2.3	For each landmark, the probability shift between the retrospective (top) and the prospective (bottom) datasets.	23
2.4	Performance metrics of total EGGIM score prediction using different cutoff values. Notice that the number of mistakes decreases substantially as one approaches the cutoff of ≥ 5 prescribed by [60]. Note that this trade-off is computed in a way that maximises sensitivity to high-risk patients.	25
3.1	Two fractional Brownian motion surfaces embedded in 3-dimensional space with known theoretical values of d_f . For reference, a plane with $d_f = 2$ was plotted at $z = 0.5$ for each of the graphs.	30
3.2	(Top-left) The realization of binomial measure μ with $p = 2/3$ supported in 2-dimension Euclidean space. (Top-right) The associated multifractal spectrum. Notice that f is a concave parabola and μ is singular, but $f(E_x[\alpha(x)])$ will still be associated with the dimension of the support of μ which is 2. (Bottom-left+right) 3D visualization of distribution of the μ at two different scales 2^{-k} , $k = \{2, 4\}$. Notice how the surface of μ appears to be more irregular as k increases.	38
3.3	Sample images from images (top) and respective class-specific segmentation masks (bottom) from (a) ISIC-18, (b) KvasirSeg, and (c) COCO datasets. Note how the biomedical regions of interest are hardly described by their shape. In fact, looking at the masks in isolation, the only plausible inference one can make is that of the cat in (c).	52
3.4	(a) The typical channel attention model where some statistics is pooled directly from the encoder output. (b) Our proposed approaches use statistics derived from the scaling exponents of the features maps.	53
3.5	Visualization of the joint encoding of p_l at encoder g_l . Each level set is characterized by a Gaussian $p_l^{(q)}$ localized around $\alpha_l^{(q)}$	55
3.6	The intensity distribution of the input is highly irregular. Conversely, the target variable is very homogeneous and only relates information of the general region of interest. Sample taken from the ISIC18 dataset.	57

3.7	Dice score (%) inter and intra-model standard deviation per dataset. The dashed gray line marks a ratio of inter/intra model variability of 1.	58
3.8	(Left-most) The input KvasirSEG image (from validation set) and target region of interest (highlighted in red). (Right) Layer depth $l \in \{1, 2, 3\}$ versus normalized response of g_l , $\mathbf{H}_l$, and $ g_l - g_l^{\text{Multi}} $. Note how $\mathbf{H}_l$ encodes complementary textural information of g_l for $l = 1$. For $l = 2, 3$ the preferred singularities relate more to luminance changes that highlight anatomical structures. This also illustrates the theory set forth by Vehel et al. [91], where each level set is associated with a distinct visual primitive.	60
3.9	(Top) Excitation analysis of excitation magnitude on a ImageNet-1000 classification task, adapted from [92]. (Bottom) The number of principal components that capture $\delta = 0.95$ of the excitation variance on a binary segmentation task on the ISIC-18 dataset for the first encoder head of a U-Net deprived of its skip-connections.	62
3.10	Test set excitation responses of cSE, SRM, FCA, and Monofractal recalibration for $l \in \{1, 2, 3\}$ in the first fold of the ISIC dataset. Each blue segment marks a response of the layer given an instance. SRM displays very heterogeneous behavior per instance. On the opposite end of the spectrum, FCA is almost instance-agnostic. cSE and Monofractal recalibration display similar instance variability, but markedly different excitation responses around the average.	63
3.11	Number of principle components required to explain 95% of the variance for each model for every encoder depth l for each dataset. cSE-NS stands for cSE in an U-Net without skip-connections. FCA is not included due to its almost heterogeneous behavior. The U-Shape architecture makes it so that become less class-dependent as $l \rightarrow L$	64
4.1	Examples of reducible/irreducible and periodic/aperiodic Markov chains.	72
4.2	A segment of 0.51 seconds of a PCG sample from the CirCor DigiScope 2022 dataset displaying an entire heart cycle.	77
4.3	The left-to-right HMM parameterized by λ for PCG segmentation containing 4 states: S1 (0), systole (1), S2 (2) and diastole (3).	80
4.4	The feasible set for the columns of $\Gamma \in \mathbb{R}^{3 \times 3}$, where $\mathbb{U}^3 = \{\mathbb{U}_1^3, \mathbb{U}_2^3, \mathbb{U}_3^3\}$. Note how $\mathbb{U}^3 \subset \mathbb{K}^3$. More specifically, it is comprised of the edges of $\mathbb{K}^3$ while excluding its vertices.	80
4.5	The template of the one-dimensional CNN backbone for the MNNs implemented in our experiments, assuming a mini-batch size B . The output layer is a 4-way softmax that yields posterior estimates for each PCG state.	82
4.6	Measured sensitivity (S) and positive predictive value (PPV) in the 10-fold cross-validation experiment of the PhysioNet'16 dataset for different static/hybrid MNNs, under different loss functions $\mathcal{L} = \mathcal{L}_{\text{MMI}} / \mathcal{L} = \mathcal{L}_{\text{CL}}$	86
4.7	Pairwise t -test ($\alpha = 0.05$) between S and PPV of $\text{MNN}_{\mathcal{L}_{\text{MMI}}}^{\Theta}$, $\text{MNN}_{\mathcal{L}_{\text{MMI}}}^{\Psi}$, $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Theta}$ and $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Psi}$ in the PhysioNet'16. White cells signify that we can reject the null hypothesis for a pair of models, black cells mark the converse. The grey colour denotes uninformative cells.	87

- 4.8 Measured sensitivity (S) and positive predictive value (PPV) statistics in 10-fold cross-validation on the CirCor'22 dataset between the HSMM by Springer et al. [206], the U-Net by Renna et al. [207], the Bi-LSTM+A by Fernando et al. [209], and $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Theta}$. The green edged triangles mark the mean of the distribution, while the black crosses signify outliers. Note how the MNN with fixed λ during gradient descent using the complete likelihood loss ($\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Theta}$) yields the best trade-off between the two metrics. . . . 88
- 4.9 Pairwise t -test between $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Theta}$, the HSMM by Springer et al. [206], the U-Net by Renna et al. [207], and the Bi-LSTM+A by Fernando et al. [209] with significance $\alpha = 0.05$ in the PhysioNet'16 dataset. White cells signify that we can reject the null hypothesis for a pair of models, black cells mark the converse. The gray color denotes uninformative cells. . . . 89
- 4.10 Measured sensitivity (S) and positive predictive value (PPV) statistics in 10-fold cross-validation on the CirCor'22 dataset between the HSMM by Springer et al. [206], the U-Net by Renna et al. [207], the Bi-LSTM+A by Fernando et al. [209], and $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Theta}$. The green edged triangles mark the mean of the distribution, while the black crosses signify outliers. Note how the MNN with fixed λ during gradient descent using the complete likelihood loss ($\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Theta}$) yields the best trade-off between the two metrics. . . . 90
- 4.11 Pairwise t -Test between $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Theta}$, the HSMM by Springer et al. [206], the U-Net by Renna et al. [207], and the Bi-LSTM+A by Fernando et al. [209] with significance $\alpha = 0.05$ in the CirCor'22 dataset. White cells signify that we can reject the null hypothesis for a pair of models, black cells mark the converse. The gray color denotes uninformative cells. . . . 91
- 4.12 Fine-tuning average performance of $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Psi_0}$ compared with $\text{MNN}_{\mathcal{L}_{\text{FT}}}^{\Psi_i}$, for $1 \leq i \leq 20$. The baseline is pre-trained on PhysioNet'16 and fine-tuned throughout different numbers of epochs on the CirCor'22 dataset. (Top) PPV. (Bottom) Sensitivity. A pairwise t -test between the baseline metric distribution and the outcome after i epochs was made to assess statistically significant differences. The left vertical axis displays average relative changes in percentage, and the right horizontal axis the average measured values. . . . 91
- 4.13 Fine-tuning average performance of $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Psi_0}$ compared with $\text{MNN}_{\mathcal{L}_{\text{FT}}}^{\Psi_i}$, for $1 \leq i \leq 20$. The baseline is pre-trained on CirCor'22 and fine-tuned throughout a different number of epochs on the PhysioNet'16 dataset. (Top) PPV. (Bottom) Sensitivity. A pairwise t -test between the baseline metric distribution and the outcome after i epochs was made to assess statistically significant differences. The left vertical axis displays average relative changes in percentage, and the right horizontal axis the average measured values. . . . 92
- 4.14 Predicted state sequences of an unseen CirCor'22 sound using a pre-trained $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Psi_0}$ on PhysioNet'16 as baseline (dotted red line) and an MNN fine-tuned before the sensitivity significance threshold $\text{MNN}_{\mathcal{L}_{\text{FT}}}^{\Psi_2}$ (dashed blue line with star markers). Note how $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Psi_0}$ deviates from the underlying regime, specifically from the diastole in the interval [28.4, 28.8] seconds onward. This behavior is mitigated by $\text{MNN}_{\mathcal{L}_{\text{FT}}}^{\Psi_2}$, which effectively reduces the differences in rhythmic structure between the real and predicted state sequences. . . . 92

4.15	Distributions of the systole and diastole heart duration in the PhysioNet'16 (dark gray) and CirCor'22 (light gray). To improve readability, we plotted only values within two standard deviations of the mean.	94
5.1	We study the relationship between identifiability and downstream performance of a model trained using contrastive learning. Illustrated is a particularly problematic mapping for the contrastive learning task: if the model achieves high accuracy on the classes in $\mathcal{X}$, it likely comes at the expense of inverting certain latent factors. On the other hand, when the model is very capable of inversion, then downstream performance is likely compromised. If the model is simultaneously capable of achieving high accuracy and identifying its latent factors, the only remaining outcome is instability, i.e., sensitivity to small perturbations.	105
5.2	Downstream accuracy pairs of $(\kappa_{\text{pre}}, \kappa_{\text{down}})$. Models with higher identifiability scores (Table 5.1) tend to be very accurate on the clean validation set. However, there are no free lunches, so they are sensitive to small perturbations.	110
5.3	Relationship between Identifiability with an empirical lower bound for class separation. Since samples are drawn by first selecting $\mathbf{z}_c$, we compute the average cosine similarity between $\mathbf{z}_i$ and $\mathbf{z}_{c_i}$	111
5.4	identifiably-stability (downstream) trade-off. As predicted by Theorem 5.9, one at the extremes, unconstrained CL learns solutions that manifest the trade-off, as a consequence of the interplay of $\mathcal{T}$ on Z and X	114
5.5	F1 measure between the MCC of the latents in the clean and noisy datasets versus the MCC of the remainder of the latent factors.	115
5.6	SimCLR color augmentations induce very high Pearson correlation coefficient scores on the validation set with respect to the latent elevation . However, it varies drastically in the noisy dataset, correlating with the drop in the coefficient controlling the shape.	116
C.1	A schematic of the experimental pipeline of Markov-based neural networks.	137
C.2	The bandpass filter used comprised of two second-order Butterworth filters.	138
C.3	A slice of a recording of the Circor'22 dataset [212] and its decomposition using 5 levels using the Haar mother wavelet.	140
C.4	A slice of a <i>noisy</i> recording of the Circor'22 dataset [212] and its decomposition using 5 levels using the Haar mother wavelet. Note how most noise concentrates at lower detail levels, and is effectively mitigated suppressed through adaptative threshold.	141
C.5	An example of a <i>noisy</i> recording of the Circor'22 dataset [212] and the output of the 5 level Haar wavelet denoiser.	142
C.6	(top) A portion of a <i>noisy</i> recording of the Circor'22 dataset [212] and its representation as the analytical signal, comprised by the real $x(t)$ and imaginary $x^*(t)$ parts. (b) The resulting envelopogram $H(t)$	144
C.7	The homomorphic envelopogram is a smooth version of the Hilbert envelope, attained by applying a low pass filter to $H(t)$. In this case, an order 2 low-pass Butterworth filter at 640 Hz as prescribed by Springer et al. [206]. . . .	145

C.8	(Top row) A denoised recording from the CirCor'22 dataset. (Remaining rows) The extracted envelopes used as input features for our models. Note how the envelopes simplify the morphology of the signal for the purposes of our task: one expects the energy to be more concentrated during the onset of the two fundamental sounds (S1/S2), which become exaggerated in the envelopograms.	146
D.1	Identifiably-stability (downstream) trade-off cross plots. As predicted by Theorem 5.9, at the extremes, unconstrained CL learns solutions that manifest the trade-off, as a consequence of the interplay of $\mathcal{T}$ on Z and X . F1 measure is computed as the harmonic mean of the predictions on the clean with those of the noisy dataset.	150
D.2	t-SNE [309] for the validation set of one of the seeds for dSprites with increasingly stronger augmentations. The augmentations decorrelate the samples with respect to their semantics (shape) into $\mathcal{T}$ -connected clusters. .	151

Chapter 1

Introduction

The overarching theme of this thesis is the study of how domain knowledge can be incorporated into deep learning systems, and what should be expected when learning is constrained by such knowledge. Rather than treating deep models as generic function approximators whose performance emerges solely from scale, optimization, and data availability, this thesis studies learning as the outcome of a configuration of modeling choices. These choices determine i) which aspects of the data are exposed to the model, ii) which functions the model is allowed to represent, and iii) which errors the optimization procedure is instructed to penalize.

In Chapter 2, this perspective is introduced through the notion of a learning configuration, expressed as the triplet $\mathfrak{C} = (\Phi, \mathcal{F}, \mathcal{L})$, where Φ denotes the data-processing map (e.g., patchification or filtering of an arbitrary input signal), $\mathcal{F}$ the hypothesis space or model family, and $\mathcal{L}$ the loss function. Under this formulation, prior knowledge is understood as an intervention on one or more components of this triplet. The chapter asks the following research question:

RQ1: To what extent can domain knowledge, expressed through the processing of the input signal, simplify the learning problem and make a clinically relevant downstream task feasible?

This question is addressed in an application-oriented setting, where the burden of incorporating prior knowledge is placed primarily on Φ . The purpose is not to propose a universally optimal deep learning architecture, but to show that the way in which data are presented to the model can materially alter the complexity of the task. In this sense,

although we mainly study the specific domain of metaplasia characterization in endoscopic images, the chapter motivates the broader methodological thesis: domain knowledge does not merely improve learning after the fact; it determines the structure of the learning problem itself.

Chapter 3 then moves from data processing to geometrical inductive bias. In this chapter, the focus is on pathological phenomena in medical imaging that exhibit self-seeding, cascading, and spatially heterogeneous behavior. These phenomena are not naturally captured by standard local or grid-based assumptions alone and are particularly relevant in for dense prediction, i.e., segmentation, tasks. The chapter therefore investigates whether principles from fractal and multifractal geometry can be used to define inductive biases that are sensitive to the scale-dependent structure of the input signal. The corresponding research question is:

RQ2: Can geometrical priors derived from fractal and multifractal analysis improve the segmentation of pathological patterns whose visual manifestation is self-seeding, cascading, and scale-dependent?

This chapter articulates domain knowledge at the level of the signal geometry. The prior is not introduced only as a preprocessing heuristic, but as a way of constraining what the model should be sensitive to. The underlying assumption is that, for certain medical imaging tasks, diagnostically relevant structure is distributed across scales and cannot be fully characterized by purely local appearance. Chapter 3 therefore studies how geometrically informed representations can guide learning toward pathological regularities that are expected from the domain itself.

In Chapter 4, the thesis moves from spatial to temporal structure, specifically through the lens of statistical sequential modeling. Here, prior knowledge is articulated both in the model and in the optimisation objective. The target domain is the heart cycle, whose dynamics are rhythmic, quasi-periodic, and naturally described through transitions between latent or observable modes. The chapter proposes a fully hybrid, model-based deep learning approach in which a statistical prior is jointly optimised with the predictive model for heart sound segmentation. Its research question is:

RQ3: What is the impact of encoding domain knowledge as a statistical sequential prior, expressed simultaneously in the model structure and in the loss function, for learning to segment rhythmic, quasi-periodic biomedical signals?

Unlike the previous chapters, where the dominant intervention concerns either the input representation or the geometry of the signal, Chapter 4 constrains learning through the temporal organisation of the phenomenon itself. The heart cycle is not treated as an arbitrary sequence of samples, but as a structured process whose transitions obey regularities that can be made explicit. This allows the model to learn not only from labeled outcomes, but also from assumptions about the expected progression of physiological states.

A final research direction is pursued in Chapter 5 related to self-supervised learning. Despite being a slightly disconnected from the previous application driven research question, it is nonetheless related in a more general sense to the properties of $\mathcal{C}$ with respect to an *arbitrary* downstream task. Specifically, this chapter examines the assumption that self-supervised learning should produce representations that capture general properties of the data distribution and can therefore be transferred across any downstream task imposed therein. However, we show that representations are not truly general in an unconstrained sense. Rather, they are bound by the degrees of variation they capture during pretraining and by the invariances implicitly induced by the pretext objective. The research question is:

RQ4: Are self-supervised representations truly task-agnostic, or does their apparent generality conceal instability when transferred to unseen in-domain datasets?

The chapter shows that representations which appear to simultaneously explain and discriminate the data may fail to generalise robustly. In particular, the more degrees of variation they encode, the more unstable they may become under small perturbations or distributional shifts within the same domain. This result reframes self-supervised learning within the thesis' central argument: even when prior knowledge is not explicitly imposed, the learning configuration still induces constraints. These constraints shape what the representation captures, what it discards, and how stable it remains downstream.

1.1 Thesis Overview

Taken together, each chapter answers a common question from complementary perspectives: which dimensions of a learning task can be manipulated to incorporate domain knowledge, and what are the consequences of doing so? Chapter 2 studies this question

through the input representation Φ ; Chapter 3 through geometrical inductive bias in $\mathcal{F}$; Chapter 4 through statistical sequential priors expressed in both $\mathcal{F}$ and $\mathcal{L}$; and Chapter 5 by examining what happens when models are assumed to learn general-purpose representations without explicit downstream supervision. The thesis therefore progresses from explicit, domain-driven interventions to a critical analysis of models that claim generality, showing that all learning systems are constrained, whether those constraints are deliberately designed or implicitly inherited from the training task or objective.

Author's note

When we set out to do a PhD. degree, not only was the deep learning field very different, but so were the expectations for our work.

This thesis was originally defined within a broader research project with a very clear application in upper gastrointestinal endoscopy. As is common in healthcare projects, from a methodological perspective, one is always conditioned to obtain the data first in order to develop appropriate methods. As the project unfolded, it became more apparent that the data was significantly harder to collect and annotate.

Given that this is a Computer Science thesis through and through, one cannot rely solely on the medical side of the problem or on data collection, no matter how challenging or impactful it may be. That said, with the exception of Chapter 5, our methods were guided by practical problems encountered when solving biomedical tasks. The common theme is how much and to what extent one can express domain knowledge in the deep learning era. To that hand, modalities with regularities in time, such as heart sounds, or in texture, such as images of medical lesions, presented the perfect playground to test our hypothesis. The specific choices of applications were mainly opportunistic and dictated by availability of in-house data or how well they aligned with ongoing projects within the research group.

Nevertheless, we note that our primary focus was to develop new methods or insights that generalize beyond the domains empirically verified in this thesis.

1.2 Notation

We aim to follow the overall notation of the famous Deep Learning book [1]. However, it is used as a reference and not a rigid template. Due to the plurality of concepts explored

in this thesis, each chapter is self-contained, and deviations from the notation set forth herein are always explicated in the introduction of new concepts.

Numbers and Arrays

a	A scalar (integer or real)
$\mathbf{a}$	A vector
$\mathbf{A}$	A matrix
$\mathbf{A}$	A tensor
$\mathbf{I}_n$	Identity matrix with n rows and n columns
$\mathbf{I}$	Identity matrix with dimensionality implied by context
$\mathbf{e}^{(i)}$	Standard basis vector $[0, \dots, 0, 1, 0, \dots, 0]$ with a 1 at position i
$\text{diag}(\mathbf{a})$	A square, diagonal matrix with diagonal entries given by $\mathbf{a}$
a	A scalar random variable
$\mathbf{a}$	A vector-valued random variable
$\mathbf{A}$	A matrix-valued random variable

Sets and Graphs

$\mathbb{A}$	A set
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	The set containing 0 and 1
$\{0, 1, \dots, n\}$	The set of all integers between 0 and n
$[a, b]$	The real interval including a and b
$(a, b]$	The real interval excluding a but including b
$\mathbb{A} \setminus \mathbb{B}$	Set subtraction, i.e., the set containing the elements of $\mathbb{A}$ that are not in $\mathbb{B}$
$\mathcal{G}$	A graph
$Pa_{\mathcal{G}}(\mathbf{x}_i)$	The parents of $\mathbf{x}_i$ in $\mathcal{G}$

Indexing

a_i	Element i of vector $\mathbf{a}$, with indexing starting at 1
$\mathbf{a}_{-i}$	All elements of vector $\mathbf{a}$ except for element i
$A_{i,j}$	Element i, j of matrix $\mathbf{A}$
$\mathbf{A}_{i,:}$	Row i of matrix $\mathbf{A}$
$\mathbf{A}_{:,i}$	Column i of matrix $\mathbf{A}$
$A_{i,j,k}$	Element (i, j, k) of a 3-D tensor $\mathbf{A}$
$\mathbf{A}_{:, :, i}$	2-D slice of a 3-D tensor
$\mathbf{a}_i$	Element i of the random vector $\mathbf{a}$

Calculus

$\frac{dy}{dx}$	Derivative of y with respect to x
$\frac{\partial y}{\partial x}$	Partial derivative of y with respect to x
$\nabla_x y$	Gradient of y with respect to x
$\nabla_X y$	Matrix derivatives of y with respect to X
$\nabla_{\mathbf{x}} y$	Tensor containing derivatives of y with respect to $\mathbf{x}$
$\frac{\partial f}{\partial \mathbf{x}}$	Jacobian matrix $J \in \mathbb{R}^{m \times n}$ of $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$
$\nabla_{\mathbf{x}}^2 f(\mathbf{x})$ or $\mathbf{H}(f)(\mathbf{x})$	The Hessian matrix of f at input point $\mathbf{x}$
$\int f(\mathbf{x}) d\mathbf{x}$	Definite integral over the entire domain of $\mathbf{x}$
$\int_S f(\mathbf{x}) d\mathbf{x}$	Definite integral with respect to $\mathbf{x}$ over the set S
Probability and Information Theory	
$P(a)$	A probability distribution over a discrete variable
$p(a)$	A probability distribution over a continuous variable, or over a variable whose type has not been specified
$a \sim P$	Random variable a has distribution P
$\mathbb{E}_{\mathbf{x} \sim P}[f(\mathbf{x})]$ or $\mathbb{E}f(\mathbf{x})$	Expectation of $f(\mathbf{x})$ with respect to $P(\mathbf{x})$
$\text{Var}(f(\mathbf{x}))$	Variance of $f(\mathbf{x})$ under $P(\mathbf{x})$
$\text{Cov}(f(\mathbf{x}), g(\mathbf{x}))$	Covariance of $f(\mathbf{x})$ and $g(\mathbf{x})$ under $P(\mathbf{x})$
$H(\mathbf{x})$	Shannon entropy of the random variable $\mathbf{x}$
$D_{\text{KL}}(P \ Q)$	Kullback-Leibler divergence of P and Q
$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution over $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

Functions

$f : \mathbb{A} \rightarrow \mathbb{B}$	The function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f \circ g$	Composition of the functions f and g
$f(\mathbf{x}; \boldsymbol{\theta})$	A function of $\mathbf{x}$ parametrized by $\boldsymbol{\theta}$. (Sometimes we write $f(\mathbf{x})$ and omit the argument $\boldsymbol{\theta}$ to lighten notation)
$\log x$	Natural logarithm of x
$\sigma(x)$	Logistic sigmoid, $\frac{1}{1 + \exp(-x)}$
$\zeta(x)$	Softplus, $\log(1 + \exp(x))$
$\ \mathbf{x}\ _p$	L^p norm of $\mathbf{x}$
$\ \mathbf{x}\ $	L^2 norm of $\mathbf{x}$
x^+	Positive part of x , i.e., $\max(0, x)$
$\mathbf{1}_{\text{condition}}$	is 1 if the condition is true, 0 otherwise

Chapter 2

Deep Learning with prior knowledge

This Thesis was designed to be self-contained for a contemporary deep learning practitioner or researcher. This means that we will not detail how and why mini-batch stochastic gradient descent works, articulate explicitly the layers of each deep learning architecture, nor the intricacies of the interaction of all these components when parallelised across several diverse devices.

This is not to say that background is not provided. Each Chapter has its own background section, followed by the development of novel methodological concepts expressing the proposed statistical or geometrical prior with as much generality as possible, concluding with a select biomedical application where domain knowledge expectations align with the conditions of the proposed (statistical or geometrical) priors.

2.1 Background

In this Section, we will articulate from a fundamental way what we mean when we conduct (deep) learning models with *prior* knowledge. We will start from the core theoretical principles that guide this dissertation.

We close with an application in which these introductory concepts can be leveraged to solve a problem that remains relevant from a medical perspective. We use a relatively straightforward set of methods in the context of contemporary Deep Learning. However, since our methodological decisions are guided by strong domain knowledge, this first application serves as the perfect companion to this introductory Chapter.

2.1.1 Maximum Likelihood Estimation

Suppose we have a collection of observations $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$, where $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$. We assume these data are independent and identically distributed (i.i.d.) and generated by an unknown underlying distribution, $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$.

Since p_{data} is unknown, we attempt to approximate it using a parametric family of probability density functions $\mathcal{P} = \{p(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$. The standard approach to selecting the optimal parameters is Maximum Likelihood Estimation (MLE):

$$\begin{aligned} \boldsymbol{\theta}_{\text{ML}} &= \arg \max_{\boldsymbol{\theta} \in \Theta} p(\mathbf{x}_1, \dots, \mathbf{x}_M; \boldsymbol{\theta}) \\ &\stackrel{\text{i.i.d.}}{=} \arg \max_{\boldsymbol{\theta} \in \Theta} \prod_{i=1}^M p(\mathbf{x}_i; \boldsymbol{\theta}) \\ &= \arg \max_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^M \log p(\mathbf{x}_i; \boldsymbol{\theta}). \end{aligned} \quad (2.1)$$

Maximising the log-likelihood is equivalent to minimising the negative log-likelihood. By normalizing by the dataset size M , we can view this as an expectation over the empirical distribution $p_{\mathcal{D}}(\mathbf{x}) = \frac{1}{M} \sum_{i=1}^M \delta(\mathbf{x} - \mathbf{x}_i)$:

$$\boldsymbol{\theta}_{\text{ML}} = \arg \min_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}} [-\log p(\mathbf{x}; \boldsymbol{\theta})]. \quad (2.2)$$

This formulation provides a direct bridge to information theory. Minimising the negative log-likelihood is equivalent to minimising the Kullback-Leibler (KL) divergence between the empirical distribution and the model distribution:

$$\begin{aligned} D_{\text{KL}}(p_{\mathcal{D}} \| p_{\boldsymbol{\theta}}) &= \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}} [\log p_{\mathcal{D}}(\mathbf{x}) - \log p(\mathbf{x}; \boldsymbol{\theta})] \\ &= \underbrace{\mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}} [\log p_{\mathcal{D}}(\mathbf{x})]}_{\text{Entropy of data (constant)}} + \underbrace{\mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}} [-\log p(\mathbf{x}; \boldsymbol{\theta})]}_{\text{Cross-entropy}}. \end{aligned} \quad (2.3)$$

Since the entropy of the empirical data is constant with respect to $\boldsymbol{\theta}$, MLE is asymptotically equivalent to finding the distribution in our model family $\mathcal{P}$ that is “closest” to nature in terms of KL divergence.

2.1.2 Empirical Risk Minimisation (ERM)

While MLE focuses on density estimation, modern machine learning is often concerned with predictive tasks where we wish to learn a mapping $f : \mathcal{X} \rightarrow \mathcal{Y}$ from inputs to targets.

Let us extend our dataset to pairs $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^M \in (\mathcal{X} \times \mathcal{Y})^M$.

We can unify the notation by defining a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ that measures the discrepancy between the prediction and the target. The goal is to minimize the *true risk* $R(f)$, defined as the expected loss over the true data distribution:

$$f^* = \arg \min_{f \in \mathcal{F}} R(f) = \arg \min_{f \in \mathcal{F}} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p_{\text{data}}} [\mathcal{L}(f(\mathbf{x}), \mathbf{y})]. \quad (2.4)$$

Since p_{data} is unknown, we approximate $R(f)$ using the *empirical risk* $\hat{R}(f)$ computed over the training set:

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \hat{R}(f) = \arg \min_{f \in \mathcal{F}} \frac{1}{M} \sum_{i=1}^M \mathcal{L}(f(\mathbf{x}_i), \mathbf{y}_i). \quad (2.5)$$

Note that MLE is simply a specific instance of ERM where the loss function is the negative log-likelihood, $\mathcal{L}(f(\mathbf{x}), \mathbf{y}) = -\log p(\mathbf{y}|\mathbf{x}; f)$.

2.2 The Necessity of Inductive Bias

Minimizing the empirical risk $\hat{R}(f)$ does not guarantee a low true risk $R(f)$. This discrepancy is the core problem of generalization. According to the “No Free Lunch” theorems [2], averaged over all possible data generating distributions, every classification algorithm has the same error rate as random guessing. Furthermore, finite sample bounds [3, 4] indicate that without assumptions on the hypothesis space $\mathcal{F}$, it is impossible to bound the difference between empirical and true risk.

Therefore, to learn efficiently from a finite number of samples M , we must inject prior knowledge into the learning process. This restriction of the solution space is known as **inductive bias**. We categorize the sources of this bias into four pillars: the choice of parametric model, the structural design (architecture), the loss function, and data processing.

2.2.1 Parametric Models (Hypothesis Space)

The first level of inductive bias is the choice of the hypothesis space $\mathcal{F}$. Instead of searching over all measurable functions, we restrict our search to a parametric family $\{f(\cdot; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$.

The capacity of this family determines the bias-variance trade-off. If $\mathcal{F}$ is too small (e.g., linear models), it may fail to capture the complexity of p_{data} (underfitting). If $\mathcal{F}$ is too large, minimising the empirical risk may lead to memorisation of the training set

rather than generalisation (overfitting). Modern deep learning often operates in the “over-parameterised” regime, where explicit capacity control is replaced by implicit biases in the optimisation dynamics.

2.2.2 Structural Inductive Bias (Architecture)

Beyond the raw capacity of the model, the geometric structure of the function approximator is critical. A good structural inductive bias aligns the model’s internal constraints with the underlying structure of the data domain [5].

For example, while a standard Multi-Layer Perceptron (MLP) is a universal approximator, it ignores the spatial structure of images. In contrast, Convolutional Neural Networks (CNNs) enforce *translation equivariance* [6]. This constraint effectively reduces the search space to functions that respect the grid topology of images, vastly improving sample efficiency. Similarly, Graph Neural Networks enforce permutation invariance [7], and Transformers model global relational dependencies [8]. These architectural choices are not merely engineering decisions; they are hard-coded assumptions about the physics of the data.

2.2.3 Loss Function

The choice of loss function $\mathcal{L}$ dictates *what* properties of the posterior distribution $p(\mathbf{y}|\mathbf{x})$ we wish to estimate with respect to some downstream task.

- **Mean Squared Error (MSE) – Regression:**

Minimizing the squared error $\sum(\mathbf{y} - f(\mathbf{x}))^2$ is equivalent to maximizing the likelihood under the assumption that the target $\mathbf{y}$ is generated by a deterministic function with added **Gaussian noise** ($\epsilon \sim \mathcal{N}(0, \sigma^2)$). It drives the model to estimate the **conditional mean** $\mathbb{E}[\mathbf{y}|\mathbf{x}]$.

- **Cross-Entropy (Log Loss) – Classification:**

Used for categorical/discrete targets. Minimizing Cross-Entropy is equivalent to maximizing the likelihood under a **Bernoulli** (binary) or **Categorical** (multi-class) distribution assumption. The model should estimate the **conditional probabilities** $P(\mathbf{y} = k|\mathbf{x})$ at the end of training.

Naturally, imposing a loss function implies some sort of *a priori* knowledge of the downstream task, reinforcing that learning does not simply emerge, but follows the constraints we imposed before even training the model.

2.2.4 Data Processing

Finally, we can inject bias by manipulating the input manifold itself. Raw sensory data often reside on a high-dimensional manifold with nuisance variables that hinder optimisation.

Preprocessing acts as a relaxation of the complexity requirements on $\mathcal{F}$. For instance, applying a denoising filter to audio or normalising pixel intensities in images simplifies the mapping that f must learn [9]. Data augmentation (e.g., rotating images) can also be viewed as a way to explicitly teach the model invariance to specific transformations, thereby smoothing the optimisation landscape and regularising empirical risk minimization [10].

2.3 Prior knowledge as an intervention on the Learning Triplet

To synthesize the components discussed above—data processing, inductive bias, and loss functions—we define the learning configuration triplet $\mathfrak{C} = (\Phi, \mathcal{F}, \mathcal{L})$, where:

- $\Phi : \mathcal{X}_{\text{raw}} \rightarrow \mathcal{X}$ represents the **data processing** function. This maps raw input from the acquisition domain (e.g., raw waveforms, noisy pixels) to the effective input domain of the model.
- $\mathcal{F} = \{f(\cdot; \theta) : \theta \in \Theta\}$ represents the **hypothesis space**. This encapsulates both the *parametric model* choice and the *structural inductive bias* (e.g., convolution, attention) that constrains the optimization landscape.
- $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ is the **loss function**, representing our assumptions about the target distribution and the metric of success.

Under this framework, the empirical risk minimisation objective becomes a search for the optimal parameters θ^* given a fixed configuration $\mathfrak{C}$:

$$\theta^* = \arg \min_{\theta \in \Theta} \frac{1}{M} \sum_{i=1}^M \mathcal{L} \left(f \left(\Phi(\mathbf{x}_i); \theta \right), \mathbf{y}_i \right). \quad (2.6)$$

In a general setting, the design of a machine learning system involves making simultaneous choices for all elements of $\mathfrak{C}$. One must select an architecture $\mathcal{F}$ that suits the data geometry, a loss $\mathcal{L}$ that reflects the task utility, and a preprocessing strategy Φ that makes the input suitable to be captured by $\mathcal{F}$. These choices are always informed, to various degrees, by a prior knowledge of the domain and downstream task.

2.3.1 Example: ResNet on ImageNet

To illustrate the configuration triplet $\mathfrak{C} = (\Phi, \mathcal{F}, \mathcal{L})$, consider the standard setup of training a ResNet-50 for image classification on ImageNet.

- **Data Processing (Φ):** The input $\mathbf{x}$ is typically assumed to be invariant to minor spatial perturbations. Φ includes resizing to 224×224 , channel-wise normalisation, and stochastic augmentations such as random crops and horizontal flips*.
- **Hypothesis Space ($\mathcal{F}$):** We restrict our search to the family of functions parameterizable by a ResNet-50. This imposes a strong **structural inductive bias** of *translation equivariance* (via convolutions). Crucially, this is a **parametric** assumption: we assume the complexity of the mapping from pixel space to semantic labels can be captured by a fixed finite parameter vector θ characterising some $f \in \mathcal{F}$.
- **Loss Function ($\mathcal{L}$):** By minimising the Cross-Entropy loss, we implicitly model the target $\mathbf{y}$ as a random variable drawn from a **Categorical distribution** parametrised by the network outputs: $\mathbf{y} \sim \text{Cat}(\text{Softmax}(f(\mathbf{x}; \theta)))$.

This Thesis’s central question is to study the interplay between these components. Clearly, conducting a joint optimisation over the entire triplet space is intractable, and even hard to imagine executing in practice. In other chapters, we will design interventions on $\mathcal{L}$ or $\mathcal{F}$ based on some domain knowledge on the downstream task and regularities on the data manifold that are particularly relevant to it.

For the introductory purposes of this chapter, and due to the more application-oriented nature of its contributions, we assume that the inductive bias provided by modern deep architectures ($\mathcal{F}$) and standard loss functions ($\mathcal{L}$) is sufficient, provided the data is presented correctly. Consequently, we focus on the role of designing Φ in order to make

*The role of data-augmentation is a bit ambiguous in general, as it can affect the three components of the $\mathfrak{C}$. However, its most natural instantiation is in Φ .

a downstream task feasible. In this context, Φ should “simplify” the data manifold, effectively “relaxing” the complexity of the input distribution before it reaches the model. This approach shifts the burden of complexity from the learning algorithm to the signal processing stage, validating the hypothesis that a well-conditioned input manifold is a prerequisite for effective empirical risk minimization.

Naturally, this “lunch” is not free. For the application we present in this chapter, we will solve a much easier problem than the one ultimately desired, using better configurations that do not require this heavy processing step.

2.4 Application: Upper gastrointestinal endoscopy

Esophagogastroduodenoscopy (EGD), also known as upper gastrointestinal (UGI) endoscopy or gastroscopy, is currently held as the principal diagnostic tool for the screening of diseases pertaining to the UGI tract such as *helicobacter pylori* infection, gastritis, ulcers, esophageal and gastric cancer [11–16]. The procedure requires the use of an endoscope, which is constituted by a flexible tube with a set of camera lenses at the tip. The video signal is streamed to a display, relaying real-time information to the clinicians during the procedure. The apparatus is manually controlled, allowing regulation of the frontage of the tip, lighting modality and magnification. The endoscope is also equipped with irrigation and insufflation components that enable direct manipulation of the environment to attain an obstructed view of the *lumen* (the interior of the UGI tract) [17]. Therapeutic intervention tools such as a forceps may also be installed for polyp removal and biopsy sampling during standard clinical procedure [18].

Modern endoscopes are typically equipped with two lighting modalities: conventional white light (WLI) and narrow band imaging (NBI). The latter is a special type of imaging modality that filters the response of the endoscope in two narrow bands that match the optical absorption region of oxidized hemoglobin [19, 20], in order to enhance contrast of the mucosal capillaries and glandular structures [19]. In combination with magnification (ME-NBI), it has become the principal tool for non-invasive characterisation of gastric cancer from atrophic gastritis (AG) or gastric intestinal metaplasia (when gastric mucosa resembles the intestinal phenotype), diagnosis of lesion extent, inference of histopathological type and diagnosis of malignancy invasion depth [20]. It also does not suffer from the risk of allergy or physical burden associated with the previous state of

practice modality; chromoendoscopy (CH), where iodine staining is used for a visual assessment of the morphology of the UGI mucosa [21, 22]. When colour or morphological changes on the UGI gastric mucosa are observed, one should switch to the NBI modality for the best possible *in vivo* characterisation and delineation of aberrant tissue [23–26]. Standardised NBI screening has allowed *in vivo* histological diagnosis inference of gastric cancer (GC) [20], with a reported accuracy rate of 88.5% in the differential diagnosis of gastric adenomas and well-differentiated adenocarcinomas [27].

Both (i) photodocumentation completion rates and (ii) characterization of epithelial tissue are amenable to be improved by means of computer assisted decision (CAD) tools. As above-mentioned, this implies mitigation of a net upper bound of 40% less cancer-related deaths due to operator error or misdiagnosis [28, 29], which is a crucial statistic for motivating high impact research.

2.4.1 Deep Learning in Gastric Lesion Characterization

Deep learning in the context of UGI lesion characterization remounts back to 2018, where Kanesaka et al. [14] employed combined features from pre-trained AlexNet [30] and SqueezeNet [31] architectures to classify gastric precancerous conditions using ME-NBI. While the clinical task was elementary (identifying polyps/ulcers vs. neoplastic tissue), it established the baseline for CNN-based feature extraction in the gastric antrum.

Subsequent research rapidly targeted Atrophic Gastritis (AG) and Gastric Intestinal Metaplasia (GIM), the primary precursors to gastric cancer. Guimarães et al. demonstrated that a fine-tuned VGG-16 [32] could outperform endoscopists in detecting AG using WLI, though the study was limited to a single center [33]. Scaling this approach, Zhang et al. [34] utilized a DenseNet architecture [35] to not only detect AG but grade its severity (mild/moderate/severe) with high accuracy ($> 93\%$), validating the model against histopathology.

The scale of data ingestion increased significantly with Lin et al. [36], who conducted a large-scale multi-center study ($n = 7037$ frames from 14 institutions). Using a T-ResNet for ternary classification (Non-AG, AG, GIM), they reported sensitivities above 96%. However, this study highlighted a recurring methodological issue in the field: the use of manually calibrated thresholds on sigmoid outputs to maximize test-set performance, a practice that lacks theoretical guarantees and risks overfitting.

Concerning the detection of Early Gastric Cancer (EGC), initial approaches by Sakai et al. [37] applied sliding-window strategies, using a CNN to process WLI EGD frames patch by patch. Recognizing the limitations of WLI, Li et al. [38] fine-tuned Inception-v3 [39] on ME-NBI data, demonstrating that DL models could achieve significantly higher sensitivity (91.18%) than non-expert endoscopists. However, low inter-observer agreement between the CNN and experts suggested that the model might be learning features distinct from standard clinical criteria (e.g., MESDA-G).

Beyond simple detection, Zhu *et al.* [40] addressed the question of invasion depth of the lesions, which dictates the viability of Endoscopic Submucosal Dissection (ESD). Using a ResNet-50 [41], they achieved high specificity (96.09%), though the reliance on manual thresholding to balance sensitivity remained a limitation.

Horiuchi et al. [42] moved beyond static frames, validating a CNN on video clips using temporal thresholding. Their system showed robustness even against challenging undifferentiated-type cancers, a notoriously difficult diagnosis for humans.

Hu et al. [43] Computer-Aided Diagnosis (CAD) tool based on the predictions of a CNN. In a multi-center trial, they demonstrated that junior endoscopists assisted by the CAD achieved diagnostic accuracy comparable to unassisted experts. Notably, when experts used the CAD system, their performance exceeded their own baseline, highlighting the potential of “human-in-the-loop” AI to standardize and elevate diagnostic quality across experience levels. Another CAD system, currently named ENDOANGEL, was developed using Deep Reinforcement Learning to track the endoscope’s position relative to a virtual stomach model [44]. The efficacy of this system to significantly reduce the operators’ blind spot rate was verified in a randomized controlled trial.

Finally, the recent trend has been to shift toward using Vision Transformers (ViTs) and ensemble fusion models to capture long-range dependencies in endoscopic video, thereby explicitly addressing the receptive-field limitations of relying solely on CNN architectures [45–47].

2.4.2 Characterisation of metaplasia for patient risk stratification

One of the main causes of GC stems from chronic *Helicobacter pylori* (HP) infections [48], where the affected mucosa may display physiological alterations [48] correlated with atrophy and gastric intestinal metaplasia (GIM) [48–50]. Under the Correa hypothesis [29],

carcinogenesis in the UGI occurs in a cascade-like fashion, with multifocal atrophic gastritis (MAG) being considered the first true precancerous condition, followed by gastric intestinal metaplasia (GIM), dysplasia and finally carcinoma [51]. Indeed, there is a link between HP infection and GIM [52, 53]. Furthermore, patients where GIM is identified are 10 times more likely to be later diagnosed with gastric cancer [51, 54]. Although controlled clinical trials of characterization protocols show high diagnostic accuracy [20, 55] when characterizing UGI epithelial issue, reliable optical diagnostic has not yet transferred favourably in a real clinical setting due intra observer variability (even within experts) [56, 57] and the high amount of training required to identify subtle visual cues, specially in the early stages of the carcinogenesis cascade [53, 56, 58]. Identification of gastric intestinal metaplasia (GIM), i.e., a change in the upper GI mucosal pattern from the gastric to the intestinal phenotype. Patients with extensive GIM are 10 times more likely to develop gastric cancer [59].

The structural changes associated with GIM are visually apparent during the endoscopic examination, specially through advanced imaging techniques, such as narrow-band imaging (NBI). However, as it stands, optical diagnosis needs to be validated against a gold standard by means of a biopsy of the suspicious tissue. Naturally, the reliance on histopathological analysis leads to increased costs, slower follow-up, and increased patient burden.

In an effort to minimize the need of a histological gold standard, the endoscopic grading of gastric intestinal metaplasia (EGGIM) was proposed as a standardized optical diagnosis protocol of GIM [60]. However, in order to implement EGGIM in real clinical practice, one has to first address the intra/inter observer variability that is associated with high missing rate of cancerous and precancerous conditions during screening [61]. Naturally, there is value in leveraging the advancements in deep learning for biomedical tasks. However, for GIM characterisation, it has been shown that, although out-of-the-box deep learning solutions work well on average, they display high variance in cross-validation analysis [?], i.e., they are highly sensitive to the training set. This is unsurprising as EGD still frames display a lot of variability, with additional limitations such as noisy labels due to the aforementioned operator limitations, and lack of sufficiently large amounts of readily available, high-quality datasets [62].

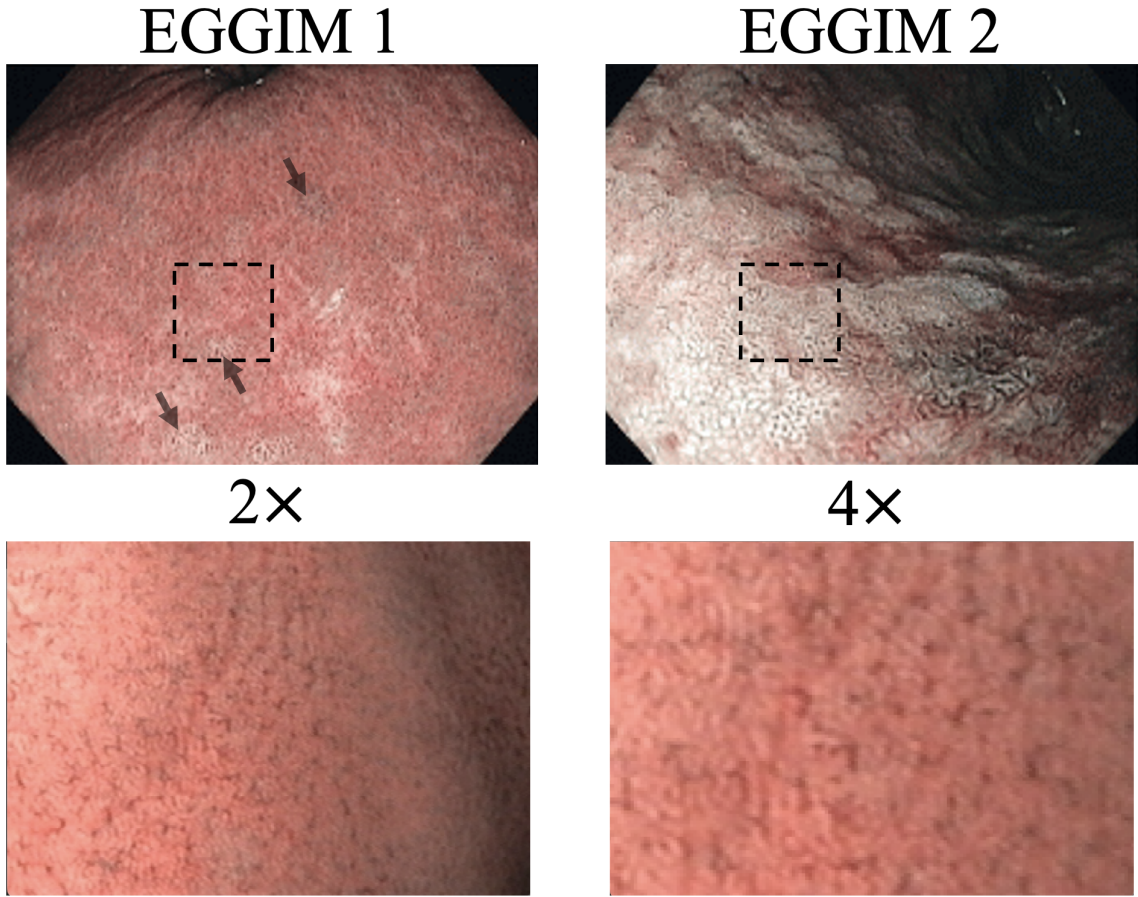


FIGURE 2.1: (Top) Two EGD frames of the antrum with EGGIM scores of 1 ($\leq 30\%$ GIM), and 2 ($\geq 30\%$ GIM). The dashed square indicates a illustrative example of a patch. Translucent arrows in the middle EGD frame indicate focal GIM. This example of EGGIM 1 is specially illustrative of the periodic and self-similar character of GIM distribution. Data adapted from [60]. (Bottom) A section of an EGD frame observed at different scales. Note how the textural attributes remain identical throughout.

2.4.3 The GIM self-similarity hypothesis

We propose to simplify the GIM characterisation task under the pretence of our *GIM self-similarity hypothesis*. Our theory is based on two key postulates: firstly, like many other pathological phenomena, the gastric carcinogenesis cascade [51] categorizes an underpinning biological process that unfolds in a self-seeding way [63]. Complex macroscopic structures emerge from the repeated interaction between the UGI and the intestinal-like epithelial cells at the level. While seemingly intricate at a single scale, these patterns become predictable when one analysis more than one scale. This *self-similar* behaviour is well-understood in the scope of multifractal analysis [64] and is commonly observed in the biological sciences [65]. Secondly, the textural patterns described by such a process may have stationary increments [66]. Given these assumptions, we contend that it should



FIGURE 2.2: Representative examples of EGD frames of healthy anatomical landmarks used in the EGGIM protocol.

be possible to estimate GIM by looking only at a small region of the EGD frame (see Fig. 2.1).

If that is the case, then one can better leverage the capacity of $\mathcal{F}$, e.g., the set of CNNs studied in [62], by looking only at small patches of the EGD frame representing the phenotypical characteristic displayed in that image.

2.4.4 Endoscopic grading of gastric intestinal metaplasia (EGGIM)

The endoscopic grading of intestinal metaplasia (EGGIM) protocol [60] requires grading of five key anatomical locations under NBI per patient: the lesser curvature (LC) and greater curvature (GC) of both the antrum and the body, and finally the incisura (see Fig. 2.2).

In each EGD, frames pertaining to the site (S) will be graded by a human expert based on the visual extent of gastric intestinal metaplasia (GIM) using the following criteria:

$$\mu(S) = \begin{cases} 0 & \text{No GIM (0\%)} \\ 1 & \text{Focal GIM } (\leq 30\%) \\ 2 & \text{Extensive GIM } (> 30\%) \end{cases} \quad (2.7)$$

where the set of sites is defined as:

$$\text{Sites} = \{\text{Antrum LC}, \text{Antrum GC}, \text{Incisura}, \text{Body LC}, \text{Body GC}\}$$

The EGGIM score is calculated by summing scores across these key anatomical sites. The total score is computed as follows:

$$\text{EGGIM score} = \sum_{S \in \text{Sites}} \mu(S) \quad (2.8)$$

The maximum possible score is 10. Patients are stratified into risk categories for gastric cancer as follows:

TABLE 2.1: EGGIM Risk Stratification

EGGIM Score	Risk Level	Clinical Implication
0 – 4	Low Risk	Routine surveillance may not be urgent
5 – 10	High Risk	High risk of gastric cancer; surveillance recommended

2.5 Data collection and annotation

Our experiments make use of two datasets collected at Instituto Português de Oncologia (IPO): a retrospective dataset collected from December 2019 to September 2020 with a total of 414 NBI just frames and a prospective dataset collected in 2024 comprised of 80 patients. The following exclusion criteria for EGD frames were applied: excessive blurriness or halation, post-biopsy bleeding, intra-procedural flushing, and overabundance of mucus. Of a total of 80, only 65 of the prospective patients had EGGIM scores for all required markers defined on [67] (as listed on Fig. 2.2). We only have these patients for the final stage of risk-stratification validation. Note that only 13 out of 65 (0.20%) of the patients have a total EGGIM score ≥ 5 , meaning that the dataset is highly skewed towards low-risk examinations.

2.5.1 Patch extraction and annotation

A trained researcher annotated the EGGIM score of each frame and a patch with a fixed resolution of 224×224 with his own local EGGIM score. Note that the patch EGGIM score could diverge from the frame EGGIM score.

Patches were systematically assigned to stand roughly in the same location depending on the anatomical location (as illustrated in Fig. 2.1). Minimal deviations from this standard placement were made only when instrumental clutter, foam, or lighting artefacts occluded or distorted the mucosal patterns. We empirically set the patch dimensions to be as small as possible while ensuring that a human annotator could consistently distinguish normal from abnormal tissue.

2.5.2 Configuration instantiation for patch GIM score prediction

Restricting the input to a single representative patch reduces the problem complexity enough to increase the capacity of $f \in \mathcal{F}$ to learn from the limited ammount of samples at our disposal. Effectively, the goal is to avoid the limitations observed in [62] to

offer robust predictions across different patients.

Let us instantiate the configuration triplet $\mathfrak{C} = (\Phi, \mathcal{F}, \mathcal{L})$ (see Section 2.3) for the purposes of this task.

- **Data Processing (Φ):** Assuming the GIM self-similarity hypothesis, the first step of Φ is to select a representative patch for each image and discard all other pixels. Φ includes resizing to 224×224 , channel-wise normalisation, and stochastic augmentations via random vertical or horizontal flips.
- **Hypothesis Space ($\mathcal{F}$):** The family of functions parameterizable by a ResNet-50.
- **Loss Function ($\mathcal{L}$):** Cross-Entropy loss between the patch and the categorical score defined in 2.7.

2.6 Addressing probability shifts for cross-validation

As shown in Fig. 2.3, there is a noticeable shift in probability distribution from the retrospective to the prospective datasets. This is surprising, since both were collected at the same oncological centre, illustrating how susceptible the data acquisition process is to sampling bias. Notwithstanding, we found benefit when augmenting our training data with the retrospective dataset, provided that we carefully sub-sampled it. We expect our configuration $\mathfrak{C}$ to be sample efficient. That being said, sample efficiency is but a lower bound on how good the model can be with limited samples, *so augmenting the number of samples is always desired, especially in deep learning*.

We designed a leave-one-patient-out-cross validation that unfolds as follows, for each fold (patient):

1. Remove the validation patient from the prospective dataset
2. Apply a stratified shuffled split of 90/10 to the remainder of the prospective data conditioned on our target variable (patch EGGIM score).
3. Augment the training set via stratified sampling of the retrospective dataset using the empirical class statistics of the prospective data selected for training

Early stopping is applied, and we compute the EGGIM score for the held-out patient.

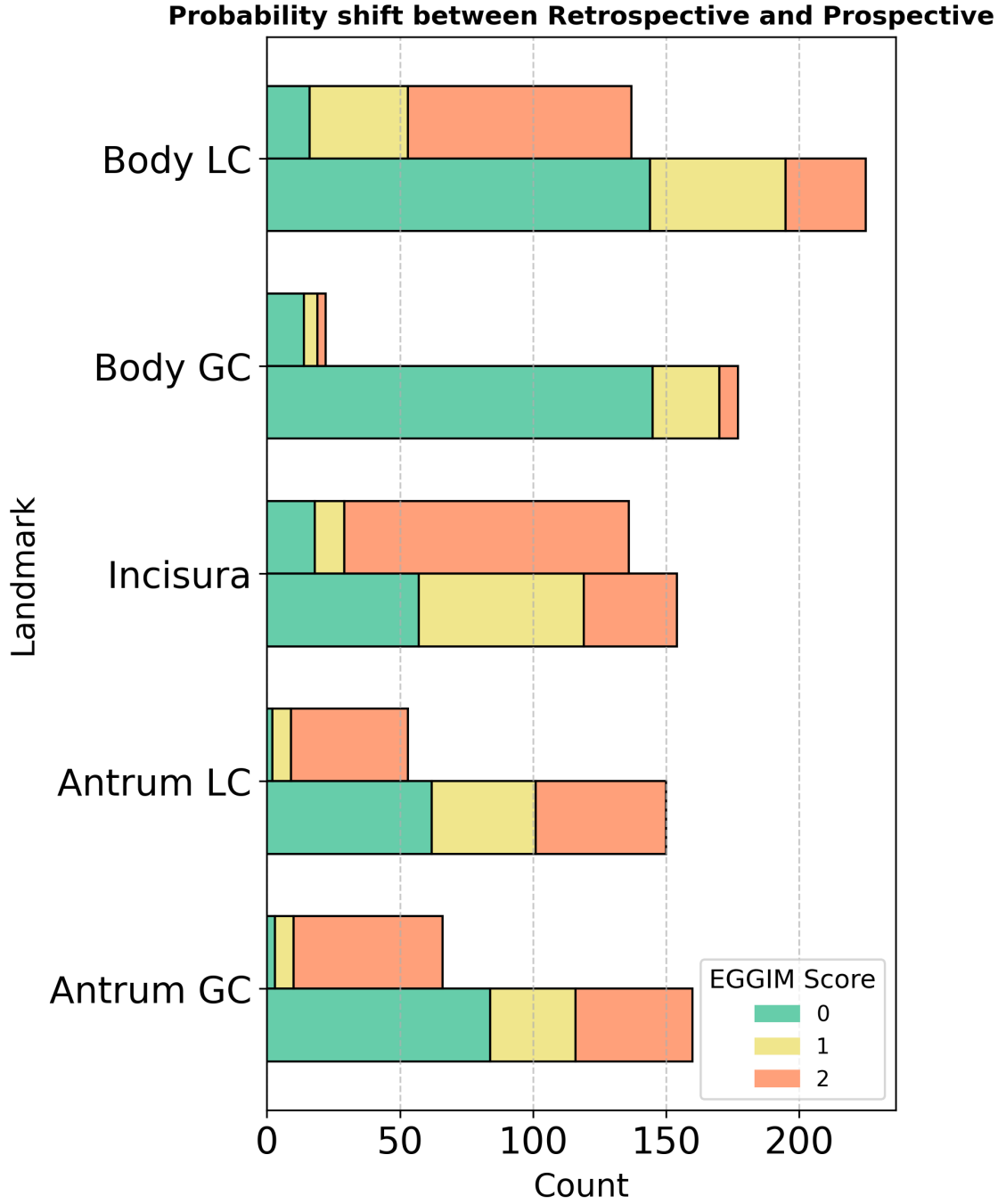


FIGURE 2.3: For each landmark, the probability shift between the retrospective (top) and the prospective (bottom) datasets.

2.7 Experiments

2.7.1 Validation of self-similarity hypothesis

To validate the self-similarity hypothesis, we apply a linearly weighted Cohen kappa test that yields 94.59% agreement (CI 99% with bootstrap $n = 1000$: 91.81% – 96.92%) between

the patch and frame-wise EGGIM scores provided by the annotator in the prospective dataset.

2.7.2 Model evaluation over highly heterogeneous and limited patient data

Given the findings of our previous work [62], we select a ResNet-50 pretrained on ImageNet as the backbone, and fine-tune a linear classifier for the task of patch GIM score prediction. Experiments were carried out using the Adam gradient descent algorithm to minimise the weighted cross-entropy loss, with a starting learning rate of 0.0001, with early-stopping. The training set is further augmented with random flips. A learning rate scheduler monitored the validation set loss and reduced the learning rate by a factor of 0.2 with patience of 3 up to a maximum of 400 epochs. We select accuracy, sensitivity, and specificity as our performance metrics for the patch EGGIM score prediction task 2.2. All experiments were conducted using the leave-one-patient-out cross-validation strategy described in 2.6.

The patch-wise class sensitivity and specificity are collected in Table 2.2.

TABLE 2.2: Leave-one-patient-out cross-validation metrics of the ResNet-50 model.

Metric	EGGIM 0	EGGIM 1	EGGIM 2
Accuracy	0.85 ± 0.16	0.90 ± 0.12	0.86 ± 0.15
Sensitivity	0.78 ± 0.28	0.73 ± 0.33	0.94 ± 0.18
Specificity	0.94 ± 0.14	0.95 ± 0.10	0.84 ± 0.19

2.7.3 Patient risk stratification validation

In order to assess the viability of these models for EGGIM risk stratification, we compute the total EGGIM score of the predictions of the model and compute the total EGGIM score described in Section 2.4.4. Using the empirical cutoff of 5 [60] (see Fig. 2.4), the model predicted the right follow-up for 57 out of the 65 patients. We measured an accuracy of 0.88, a sensitivity of 1.00, and a specificity of 0.85 using this threshold.

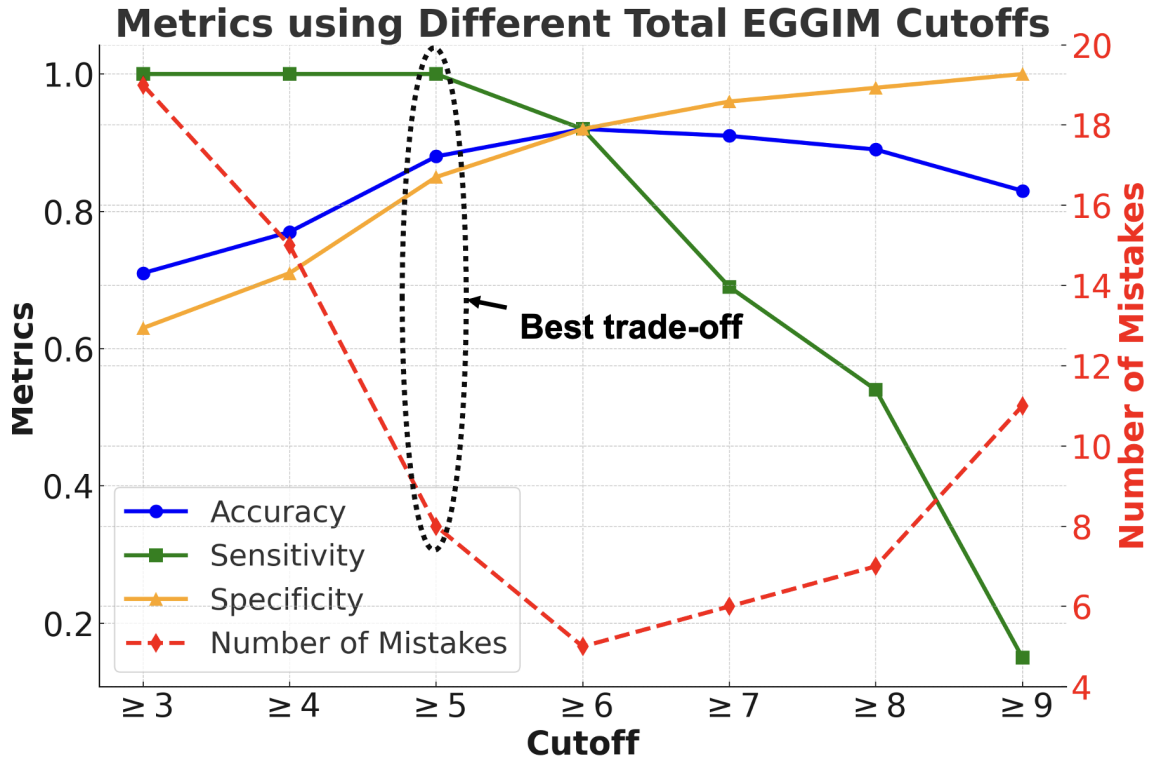


FIGURE 2.4: Performance metrics of total EGGIM score prediction using different cutoff values. Notice that the number of mistakes decreases substantially as one approaches the cutoff of ≥ 5 prescribed by [60]. Note that this trade-off is computed in a way that maximises sensitivity to high-risk patients.

2.7.4 Discussion and conclusion

There is almost perfect agreement between the patch and frame-wise EGGIM annotations

The high-weighted Cohen kappa indicates almost perfect agreement signalling that indeed a single patch can be predictive of the entire frame's score. However, note that this scoring is given by the same annotator who was not blinded to the global and patch views. Further research is needed to support this result.

Our methodology is robust across patients in an extremely unbalanced scenario

Using patch-wise information, coupled with our stratified retrospective augmentation (see Section 2.6) addresses the model variability discussed in previous work [62].

The best threshold agrees with the empirical human threshold of [60]

We inspected the performance metrics using several cutoff values (see Fig. 2.4). We observe the best performance when prioritising sensitivity by setting the empirical cutoff to ≥ 5 as proposed by [60] (this threshold was also used to determine the ground-truth variables). Note that this experiment reveals that one can get up to 0.91 accuracy by increasing the threshold to ≥ 6 , but this would lead to an increasing amount of false negatives in this unbalanced scenario, which is extremely troublesome for the purposes of gastric cancer screening.

The EGGIM risk stratification prediction is promising but is not yet fully automated

Although we were able to predict the right outcome for 57 out of the 65 patients, we should note that a clear limitation is the manual placement of the patch location. Although we avoided adjusting the patch coordinates to frame-specific conditions, there were cases where we had to, inevitably, change its placement significantly to ensure that the patch was representative, e.g., when light reflection artefacts dominated the scene at the default location.

We set forth the GIM self-similarity hypothesis, where we conjecture that the textural patterns [20] that characterise the different stages of the carcinogenesis cascade [51] are governed by some stationary process. Under the pretence of this hypothesis, we reduce EGGIM score classification to a single patch for each frame. Leveraging a pre-trained ResNet-50 is shown to increase EGGIM score prediction using a 224×224 patch of each original frame. Final risk stratification computed from patch predictions was accurate for 57 of 65 patients in a leave-one-patient-out cross-validation setting, and all high-risk patients were correctly identified despite a bias toward low risk in the prospective dataset.

Much can be done to enhance performance, given that we only leverage a pre-canned estimator (ResNet-50). A natural next step would be to replace the patch encoder with a foundational model, especially since these are gaining traction in adjacent modalities such as histopathology [68]. Additional conditional information might be used to enhance performance, such as each frame’s anatomical location, especially since there is also an imbalance with respect to this variable (see Fig. 2.3, e.g., Body GC).

Ultimately, the goal is to automate the patch extraction process, either by regressing the coordinates of the manually placed boxes or by averaging predictions on a random neighbourhood of overlapping patches. Naturally, the Transformer seems to be a suitable

structural inductive bias, due to its global receptive field and explicit patching of the input images [69].

2.8 Chapter conclusion

This chapter introduced the core concepts of this thesis, namely the interplay between the data-manifold $\mathbb{M}$, the objective $\mathcal{L}$ and the inductive bias $\mathcal{F}$, and how our knowledge of the downstream task may be used to leverage them.

We studied an application for upper gastrointestinal endoscopy, where we posited that epithelial tissue of the gastric tract displayed a degree of self-similarity, i.e., it displays regularity across scales. This allowed us to simplify the task of metaplasia characterisation considerably, relegating the input from the whole frame to a single representative patch of the image. As opposed to preliminary work where the models that processed the entire image displayed sensitivity to the train set [62], we demonstrated through an extensive cross-validation procedure that indeed our model was generalising to unseen patients.

The notion of self-similarity ties in specially with the next Chapter, where we articulate it directly as an inductive prior restricting $\mathcal{F}$ for the more general task of medical image segmentation. In Chapter 4, we will impose restrictions on both the $\mathcal{F}$ and the objective or cost function $\mathcal{L}$ for sequential modalities, with applications to heart sound segmentation.

2.8.1 Outcomes

1. 3 publications in peer-reviewed conferences [62, 70, 71].
2. 2 publication in Q1 journals (second author) [72, 73].
3. 1 unindexed national proceeding [74].
4. Design of data collection and annotation protocol for EGD frames and video together with Superannotate [75] and the IPO Porto medical research team.

Extended state of the art

In order for the main document to be focused on the new methodological concepts developed in the period of this Thesis, we provide an extended state-of-the-art with more clinical detail in the Supplementary material. The interested reader is invited to consult Appendix A.

Chapter 3

Multifractal Recalibration

In Computer Vision research, there is often a bias towards tasks related to urban settings populated by man-made objects, such as CityScapes [76] or ADE20K [77]. Other datasets such as COCO [78] may be inserted in the broader category of *natural* image segmentation *in the wild*. These so-called *natural* objects often exhibit consistent geometric features that are distinctive and recognizable according to our *Euclidean* understanding of the world (e.g., the *pointy* ears of a cat or the *curvy* horns of a cow; see Figure 3.3). However, most domain-specific tasks modalities such as those related to biology, material science, or even in general physics, often deal with far more *irregular* objects, be that in shape, color, or texture. That is to say, we can no longer describe them using our familiar Euclidean language in a satisfactory, concise way.

This chapter is concerned with the semantic segmentation within domain-specific contexts, such as those pertaining to biology, physics, or material science. Under these circumstances, the objects of interest are often irregular and have *fine structure*, i.e., detail at arbitrarily small scales. Empirically, they are often understood as self-similar processes, a concept grounded in Multifractal Analysis. We find that this multifractal behavior is carried out through a convolutional neural network (CNN), if we view its channel-wise responses as self-similar measures. Equipped with these knowledge we will propose two new channel attention functions that shall recalibrate the responses of a convolutional neural network

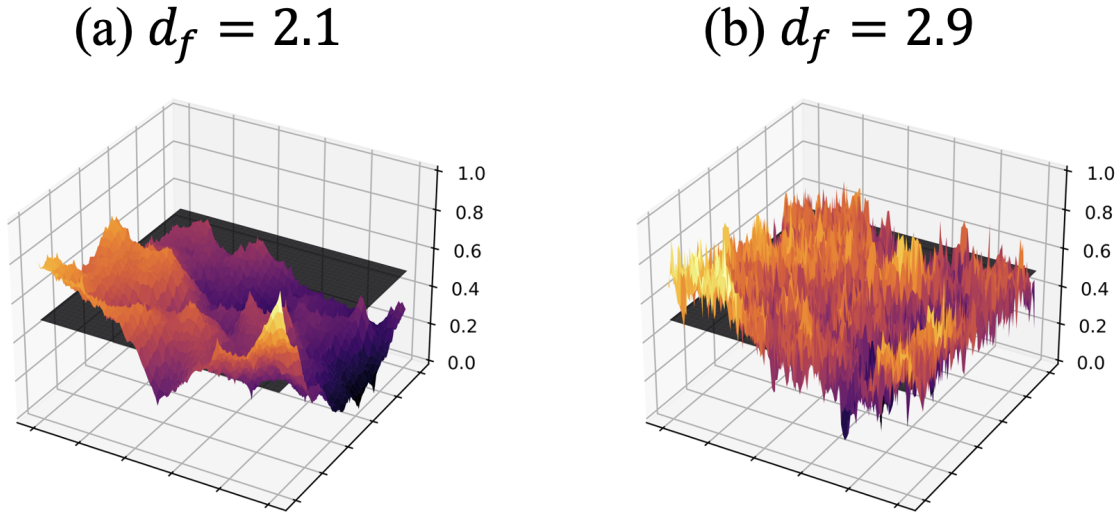


FIGURE 3.1: Two fractional Brownian motion surfaces embedded in 3-dimensional space with known theoretical values of d_f . For reference, a plane with $d_f = 2$ was plotted at $z = 0.5$ for each of the graphs.

3.1 Informal introduction

For instructive purposes consider the example given by [79]. Imagine that a surface with dimension d_f was digitized and is now stored in a computer. Assign unit mass to each pixel contained within this surface, resulting on the total volume $M \propto \epsilon^{d_f}$. If this surface is embedded in a d -dimensional space (note that $d_f \leq d$), then one can compute the fraction of space taken by the surface as $\rho = M\epsilon^{-d}$, where ϵ of the ϵ -covers of the d -dimensional space. If we now change the scale to 2ϵ , then the number of pixels in the surface will be $2^{d_f} M$, and the density will become $2^{d_f-d}\rho$. It follows that, as the surface gets more irregular, the more efficiently it can fill space, and so the closer its scaling exponent gets to d (see Figure 3.1). When quantity d_f is described as a non-integer number, we say that this surface is a *fractal*. It is not hard to imagine that each pixel might be related to a different scaling exponent, and hence we would say that in such a case, the surface would be distributed in a *multifractal* fashion.

Central to Fractal Geometry is notion is of *self-similarity*: a self-similar set $\mathbb{I} \subseteq \mathbb{R}$ can be partitioned into N equal parts scaled by a factor $r = \frac{1}{N^{-D}}$ so that:

$$N \cdot \left(\frac{1}{N^{-D}} \right)^D = N \cdot r^D = 1,$$

where $N > 0, D \in \mathbb{R}_0^+$. This in turn implies that:

$$N = r^{-D} \iff D = -\frac{\log N}{\log r}. \quad (3.1)$$

Note that self-similarity is not necessarily unique to study of fractal objects. It is easy to see that even the Platonic solids such as cube or a tetrahedron obey this property. *Fractal Geometry* is concerned with sets that have fine-structure, i.e., detail at all scales [80]. Often these sets are also (either exactly or statically) *self-similar*, which is to say that small portions of the set may be magnified and distorted in such a way as to resemble the set as a whole. Typically, one can generate such sets by means of simple, iterative procedures [80].

Useful Identities

Increasing set sequence A sequence $\{\mathbb{B}_i\}$ such that $\mathbb{B}_i \subseteq \mathbb{B}_{i+1}$.

Set diameter $\text{diam}(\mathbb{I}) = \sup\{|x - y| : x, y \in \mathbb{I}\}$.

σ -algebra on $\mathbb{I}$ A collection $\mathcal{B}$ of subsets of $\mathbb{I}$ satisfying:

1. $\emptyset \in \mathcal{B}$ (and $\mathbb{I} \in \mathcal{B}$)
2. $\mathbb{A} \in \mathcal{B} \implies \mathbb{A}^c \in \mathcal{B}$ (Closed under complements)
3. $\mathbb{A}_i \in \mathcal{B} \implies \bigcup_{i=1}^{\infty} \mathbb{A}_i \in \mathcal{B}$ (Closed under countable unions)

Measurable space $(\mathbb{I}, \mathcal{B}, \mu)$ A measure $\mu : \mathcal{B} \rightarrow [0, \infty]$ satisfying:

1. $\mu(\emptyset) = 0$
2. $\mathbb{A} \subseteq \mathbb{B} \implies \mu(\mathbb{A}) \leq \mu(\mathbb{B})$ (Monotonicity)
3. **Countable Sub-additivity:**

$$\mu \left(\bigcup_{i=1}^{\infty} \mathbb{B}_i \right) \leq \sum_{i=1}^{\infty} \mu(\mathbb{B}_i)$$

with equality (σ -additivity) if the sets $\mathbb{B}_i$ are disjoint.

3.2 Background

This chapter will closely follow the notation used in the books by Falconer [80, 81]. A more hands-on overview of this concept with an emphasis on images may be found in [82]. For

our purposes, the early multifractal formalism established by Evertz and Mandelbrot [64] will be sufficient, but the reader is referred to [83–85] for more sophisticated contemporary formalisms.

3.2.1 Lebesgue Measure

The *Lebesgue measure*, λ , of a set in $\mathbb{R}$ is the diameter of the smallest possible countable covering $\mathbb{I} = \cup_i [a_i, b_i]$. Concretely:

$$\lambda(\mathbb{I}) = \inf_{\mathbb{U}} \left\{ \sum_{i=1}^{\infty} (b_i - a_i) : \mathbb{U} \subset \cup_{i=1}^{\infty} [a_i, b_i] \right\}. \quad (3.2)$$

One can extend this notion to the n -dimensional spaces.

Definition 3.1 (Lebesgue measure on $\mathbb{R}^n$). Let $\mathbb{I} = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n : a_i \leq x_i \leq b_i\}$ then the volume is defined as

$$\text{vol}^n(\mathbb{I}) = (b_1 - a_1)(b_2 - a_2) \dots (b_n - a_n).$$

The n -dimensional Lebesgue measure λ^n is given by:

$$\lambda^n(\mathbb{I}) = \inf \left\{ \sum_{i=1}^{\infty} \text{vol}^n(\mathbb{I}_i) : \mathbb{I} \subset \cup_{i=1}^{\infty} \mathbb{I}_i \right\}$$

The Lebesgue measure is implicitly used in most of the practical applications of calculus: from probability density functions to Riemann integrals, we usually assume a Lebesgue measurable space.

More concretely, the Lebesgue measures capture our notion of length, area, and, more generally, n -dimensional volume. If $\mathbb{I}$ defines a line on some metric space, we expect $\lambda^i > 0$ iff $i = 1$. Loosely speaking, typical objects we think of have n -dimensional will have only have Lebesgue positive measure up to λ^n : a point has 0 measure, a line has 0 area, a square has 0 volume, and so on.

3.3 Objects with Lebesgue measure 0

We approach these more complicated objects conceptually. Let us build the famous Cantor set and observe its Lebesgue measure λ . Consider that we build $\mathbb{I}$ using a recursive, iterated procedure (see Supplement B.4 for a more complete discussion) as:

1. $\mathbb{B}_0 = [0, 1]$
2. At iteration 1, discard the middle third of $\mathbb{B}_0$, yielding: $\mathbb{B}_1 = [0, 1/3] \cup [2/3, 1]$.
3. By induction: $\mathbb{B}_k = \cup_i [3^{-k}i, (i+1)3^{-k}]$.

The Cantor set is given by $\mathbb{F} = \lim_{k \rightarrow \infty} \cup_i \mathbb{B}_k^i$, where $\mathbb{B}_k^i \subset \mathbb{B}_k$.

Note that, at each k , we will have 2^k intervals. Consider the measure space $(\lambda, \mathbb{F}, \mathcal{B})$, by Definition 3.1:

$$\lambda(\mathbb{F}) = \lim_{k \rightarrow \infty} \inf \left\{ \sum_{i=1}^{2^k} \text{vol}(\mathbb{B}_k^i) : \mathbb{B}_k^i \in \mathbb{B}_k, \mathbb{F} \subset \bigcup_{i=1}^{2^k} \mathbb{B}_k^i \right\}.$$

By construction, the covers at level k , $\{\mathbb{B}_k\}$, will all have the same length 3^{-k} , thus the above expression reduces to:

$$\lambda(\mathbb{F}) = \lim_{k \rightarrow \infty} \underbrace{|\{\mathbb{B}_k\}|}_{2^k} \underbrace{\text{vol}(\mathbb{B}_k)}_{3^{-k}} = 0. \quad (3.3)$$

3.4 Measuring sets of Lebesgue measure 0

It is surprising that our traditional notion of length does not apply to the Cantor set. Indeed, by inspecting it on Figure 3.2 it is hard, if not impossible, to describe the shape of the $\mathbb{F}$ with natural language without a circular definition: its shape is “irregular”. The lack of a characteristic shape alludes to an interesting connection with these mathematical objects with the ones typically found in Nature – as famously proclaimed by Mandelbrot [86]: “clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor light travels in a straight line”.

Is there any way to “measure” the Cantor set?

Equation 3.3 goes to 0 because the number of covers grows asymptotically faster than their diameter. It is as if one cannot measure the set globally, even with infinite precision, at least equipped solely with the Lebesgue measure. The question thus becomes: does any measure μ other than λ exist so that $0 < \mu(\mathbb{F}) < \infty$?

This problem has been famously addressed for sets with the same nature as the Cantor set by Felix Hausdorff [87], and later inspired Mandelbrot to develop the Fractal Geometry theory [86].

Definition 3.2 (Hausdorff measure on $\mathbb{R}^n$). Let $\mathbb{F} \subset \mathbb{R}^n$ so that $\mathbb{F} \subset \bigcup_{i=1}^{\infty} \mathbb{B}_i$. For any $\delta > 0$, then denote $\{\mathbb{B}_i\}$ as a countable set of δ -covers of $\mathbb{F}$:

$$\mathcal{H}_{\delta}^d(\mathbb{F}) = \inf \left\{ \sum_{i=1}^{\infty} \text{diam}(\mathbb{B}_i)^d : \text{diam}(\mathbb{B}_i) < \delta, \{\mathbb{B}_i\} \text{ covers } \mathbb{F} \right\}$$

then the d -dimensional Hausdorff measure, $\mathcal{H}^d$, is given by:

$$\mathcal{H}^d = \lim_{\delta \rightarrow 0} \mathcal{H}_{\delta}^d.$$

One motivation for assigning a measure to the Cantor set is the lack of a characteristic shape. The Hausdorff measure does indeed reflect this property.

Proposition 3.3 (The Hausdorff measure is scale-free). Suppose S is scaling operator by a factor of α so that $S(\mathbb{F}) = \{\alpha x : x \in \mathbb{F}\}$, then:

$$\mathcal{H}^s(S(\mathbb{F})) = \alpha^s \mathcal{H}^s(\mathbb{F})$$

Proof. By Definition 3.2, where $\{\mathbb{B}_i\}$ is an at-most- δ -cover of $\mathbb{F}$:

$$\sum_i \text{diam}(S(\mathbb{B}_i))^s = \sum_i |\alpha|^s \text{diam}(\mathbb{B}_i)^s = \alpha^s \sum_i \text{diam}(\mathbb{B}_i)^s.$$

by plugging definition 3.2

$$\mathcal{H}^s(S(\mathbb{F})) = \mathcal{H}_{\alpha\delta}^s$$

The proof follows by taking the limit: $\lim_{\delta \rightarrow 0} \mathcal{H}_{\alpha\delta}^s = \alpha^s \lim_{\delta \rightarrow 0} \mathcal{H}_{\delta}^s$. $\square$

The Hausdorff measure of a set $\mathcal{H}^d$ (e.g., $\mathbb{I} \subset [0, 1]$), either goes to 0 or ∞ as $\delta \rightarrow 0$. The exponent $\cdot^d$ indicates that now one can calculate the limiting behavior of a measure by a particular exponent. The question that one should answer next is whether it is possible to find d^* so that the set $0 < \mathcal{H}^{d^*}(\mathbb{F}) < \infty$?

Remark 3.4 (Hausdorff measure is non-decreasing for small δ). If $\delta < 1$, the value of $\mathcal{H}_{\delta}^d$ is non-increasing with d . Let $t > d > 0$, then:

$$\sum_{i=1}^{\infty} \text{diam}(\mathbb{B}_i)^t \leq \sum_{i=1}^{\infty} \text{diam}(\mathbb{B}_i)^{t-d} \text{diam}(\mathbb{B}_i)^d \leq \delta^{t-d} \sum_{i=1}^{\infty} \text{diam}(\mathbb{B}_i)^d.$$

Note that this is not a strict inequality because we assume $\text{diam}(\mathbb{B}_i) < \delta$.

Suppose there is $0 < \mathcal{H}^d < \infty$, then necessarily the graph $(d, \mathcal{H}^d)$ will change abruptly from ∞ to 0 precisely at $d = d^*$. The Hausdorff dimension essentially is a clever trick to “detect” this discontinuity.

Definition 3.5 (Hausdorff-Besicovich dimension). The Hausdorff dimension $\dim_H$ is given by :

$$\dim_H \mathbb{F} = \inf\{d \geq 0 : \mathcal{H}^d(\mathbb{F}) = 0\} = \sup\{d : \mathcal{H}^d(\mathbb{F}) = \infty\},$$

i.e., the value of $\mathcal{H}^d$ for the critical d where the Hausdorff measure goes from ∞ to 0.

Now the answer to the problem of $\mathbb{F}$ carrying measure on not requires us to show that there exists a finite and positive Hausdorff measure $\mathcal{H}^{d^*}$. The proof can be involved depending on the set $\mathbb{F}$, and it can be consulted specifically for the Cantor in [80, Example 2.7]. Here, we only provide an informal argument that relies on the assumption that such a d^* exists.

Let $\mathbb{B}_1 = \{\mathbb{F}_L, \mathbb{F}_R\}$, where $\mathbb{F}_L = [0, 1/3]$, $\mathbb{F}_R = [2/3, 1]$, so $m = 2$ according to definition B.4. Extend inductively for $k \rightarrow \infty$. Observing that $\text{diam}(\mathbb{F}_k) = 3^{-k}$:

$$\begin{aligned} \mathcal{H}^{d^*}(\mathbb{F}) &= \mathcal{H}^{d^*}(\mathbb{F}_L) + \mathcal{H}^{d^*}(\mathbb{F}_R) \stackrel{3.3}{=} (3^{-d^*})\mathcal{H}^{d^*}(\mathbb{F}) + (3^{-d^*})\mathcal{H}^{d^*}(\mathbb{F}) \\ 0 < \mathcal{H}^{d^*}(\mathbb{F}) < \infty &\implies d^* = \dim_H = \frac{\log 2}{\log 3} \approx 0.63. \end{aligned}$$

Note that the Hausdorff dimension here is a fractional number. This is no coincidence, and inspired Mandelbrot to coin the terms *fractal* set and dimension [86]. Indeed, this quantity is better suited for objects that have no *characteristic* length (as defined by the classical Lebesgue measure).

3.4.1 Fractal sets and dimension

In practice, $\dim_H$ is very hard to compute. However, an easier quantity called the *box-dimension* can serve as a great empirical prior. We are particularly interested in self-similar, potentially fractal sets.

Intuitively, the Hausdorff dimension will depend on $(\mathcal{H}^d, \mathbb{F}, \mathcal{B})$, concretely, depends on the properties of the σ -algebra $\mathcal{B}$. Suppose that we restrict ourselves to defining $\mathcal{B}$ on non-overlapping, *diadic cubes*:

$$\mathbb{B}_\delta \in \mathcal{B}_\delta := [m_1\delta, (m_1 + 1)\delta] \times \dots \times [m_n\delta, (m_n + 1)\delta], \quad (3.4)$$

where $\{m_i\} \subseteq \mathbb{N}$ defines the coordinate of each cube. Intuitively, we can think of this σ -algebra as imposing several discretization levels on the evaluation of $\mathcal{H}^d$ as a function of δ . Let $\{\delta_k : \delta_{k+1} = r\delta_k\}$ be a decreasing sequence. By slight abuse of notation we suppose $\mathbb{B}_k \equiv \mathbb{B}_{\delta_k}$, the number of hyper-cubes covering $\mathbb{F}$ “with resolution δ_k ” is:

$$N_{\delta_k}(\mathbb{F}) = |\{\mathbb{B}_k \in \mathcal{B} : \mathbb{B}_k \cap \mathbb{F} \neq \emptyset\}| \quad (3.5)$$

We are now interested in aggregating the volume of each of this covers at resolution k . As stated in the informal introduction of this chapter, when we have a proper, bounded, subset $\mathbb{F} \subset \mathbb{R}^n$, then the *scaling* of this set is characterized by:

$$N_{\delta}(\mathbb{F})\delta^d \implies d = \frac{N_{\delta}(\mathbb{F})}{-\log \delta}.$$

By taking the limit $\delta \rightarrow 0$, we get the so called *box*-dimension of $\mathbb{F}$.

Definition 3.6 (Box-dimension). Let $\mathcal{B}$ be a Borel σ -algebra of dyadic cubes of the form of equation 3.4. Then the box-dimension $\dim_B$ of $\mathbb{F}$ is given by:

$$\dim_B \mathbb{F} = \lim_{\delta \rightarrow 0} \frac{\log N_{\delta}(\mathbb{F})}{-\log \delta}. \quad (3.6)$$

Remark 3.7. The box-dimension may not provide the “optimal partition” of $\mathbb{F}$, so:

$$\mathcal{H}_{\delta}^d \leq N_{\delta}(\mathbb{F})\delta^d,$$

so it follow that the box-dimension is an upper-bound for the Hausdorff dimension.

We will generally describe principles where one has claim equality in Remark 3.7. Specifically, provided the set scales with k in way that it “fits well” inside boxes of side δ^{-k} .

For the purposes of this chapter we will assume that:

$$\dim_B \mathbb{F} \approx \dim_H \mathbb{F}.$$

This assumption is central to the practical application of Fractal Geometry in various domains, such as physics, biology, and imaging [63, 80, 86].

Definition 3.8 (Empirical fractal dimension). Suppose that $\mathbb{F}$ is self-similar with the same scaling exponent everywhere. Then, for a finite set of measurements $\mathcal{R} = \{r : 0 < r \leq 1\}$:

$$\dim_B \mathbb{F} \approx \hat{D}(\mathbb{F}) = \text{slope of } (\log N_r(\mathbb{F}), -\log r)_{r \in \mathcal{R}},$$

where $\hat{D}$ is the empirical estimate of the *box dimension* of $\mathbb{F}$.

Onward, we will use the terms Hausdorff and box dimension interchangeably and will refer to the theoretical Hausdorff dimension explicitly if there are meaningful theoretical differences.

3.4.2 Multifractal formalism

In essence, Multifractal formalism extends Fractal Geometry from sets with measure that scales solely with one exponent, to each element of the support of the measure scaling according to a possibly unique scaling law. Generally speaking, the measure depends on the spectrum of scaling exponents.

For characterizing the scaling laws, we are concerned specifically with the *Hölder exponent* or *singularity strength* of μ at scale r around x [64, 88]:

$$\alpha(x) = \sup_{|x-x_0| \leq \delta} \{\alpha' : |g(x) - g(x_0)| \leq C|x - x_0|^{\alpha'}\}, \quad (3.7)$$

Hölder regularity is also often expressed as the distance of $f(x)$ to a polynomial $P(x - x_0)$ with degree $k < \alpha$ [84]. Intuitively, having well-defined bounded derivatives up to order α means that the function tends to vary more smoothly, whereas we expect functions with smaller values of α (especially for $\alpha < 1$) to behave more irregularly. It also means that the function will behave in a scale-free fashion around x with exponent $\alpha(x)$.

Consider a measure supported on $\mathbb{I}$. If μ is self-similar at some $x \in \mathbb{I}$, then we expect to derive $\alpha(x)$ by observing $(\alpha_k(x), k)$ for increasingly small discrete k . The so-called *local Hölder exponent* or *singularity strength* [64, 89, 90] of $x \in \mathbb{B}_k \subset \mathbb{I}$:

$$\alpha_k(x) \sim \frac{\log \mu(\mathbb{B}_k(x))}{\log \delta^{-k}}. \quad (3.8)$$

Definition 3.9 (Multifractal set). Suppose $\mathbb{I} = \cup_{\alpha} \mathbb{I}_{\alpha}$, where each $\mathbb{I}_{\alpha}$ is non-empty and scales with distinct α . Consider that a map $\alpha \mapsto \mathbb{I}_{\alpha}$ exists and

$$\mathbb{I}_{\alpha} \cap \mathbb{I}_{\alpha'} = \emptyset \implies \alpha \neq \alpha'.$$

$\mathbb{I}$ is said to be *multifractal*, with multifractal or singularity spectrum $(\alpha, \dim_H(\mathbb{I}_{\alpha}))$.

For our purposes:

$$\forall x \in \mathbb{I}_{\alpha} \subset \mathbb{I} \implies \alpha_k(\mathbb{I}_{\alpha}) = \alpha_k(x).$$

From an intuitive perspective, this means that we will inspect the covers of level-set $\mathbb{I}_\alpha$ at several resolutions imposed by k .

3.4.3 Multiplicative cascades and $f(\alpha)$

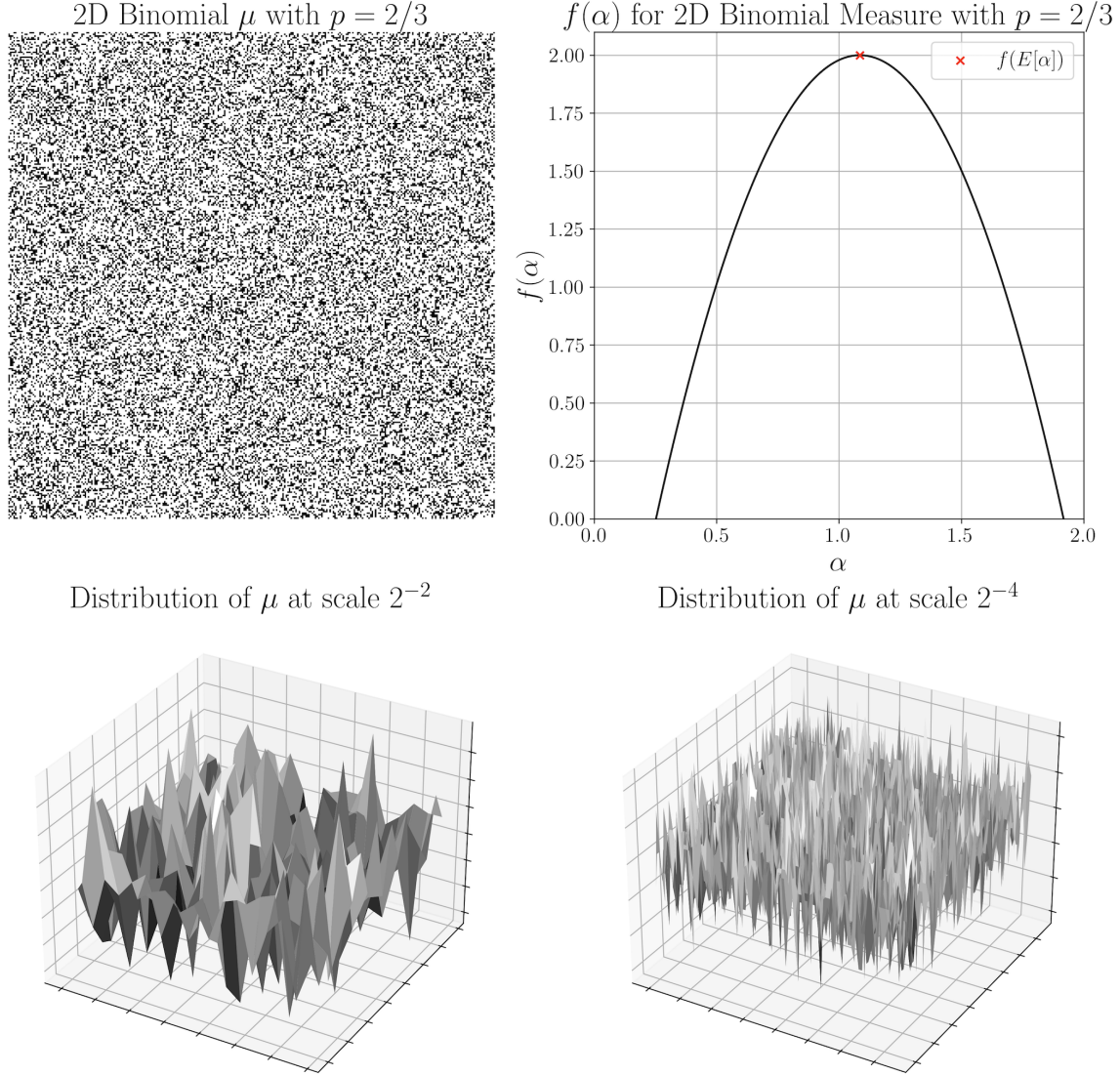


FIGURE 3.2: (Top-left) The realization of binomial measure μ with $p = 2/3$ supported in 2-dimension Euclidean space. (Top-right) The associated multifractal spectrum. Notice that f is a concave parabola and μ is singular, but $f(E_x[\alpha(x)])$ will still be associated with the dimension of the support of μ which is 2. (Bottom-left+right) 3D visualization of distribution of the μ at two different scales 2^{-k} , $k = \{2, 4\}$. Notice how the surface of μ appears to be more irregular as k increases.

It is instructive to construct a self-similar measure and observe the cases where the term *multifractal* becomes necessary. Suppose that $\mathbb{I} = [0, 1]$, but μ will assign mass according to some parameter $p \in [0, 1]$ at $\mathbb{I}_0 := [0, 2^{-1})$ (left) and $(1 - p)$ at $\mathbb{I}_1 := [2^{-1}, 1)$ (right). We will extend this inductively and construct a binomial expansion $\mathbf{w}_k$ for each

dyadic interval by marking the left and right expansions with 0 and 1, respectively. Suppose that

$$\mathcal{B}_k = \cup_i \mathbb{I}_{\mathbf{w}_k} = \cup_i [i2^{-k}, (i+1)2^{-k}),$$

where $\mathbf{w}_k = (w_1, w_2, \dots, w_k)$, $w_i \in \{0, 1\}$. Note that we choose by convention the expansion with infinitely many zeroes to account for dyadic points (i.e., all $x = \frac{m}{2^{-k}}$, where m and k are integers). Then, by construction $\mathbb{I}_{\mathbf{w}_k} = \mathbb{I}_{\mathbf{w}_{k0}} \cup \mathbb{I}_{\mathbf{w}_{k1}}$, and so:

$$\mu(\mathbb{I}_{\mathbf{w}_k}) = p^{n_0(\mathbb{I}_{\mathbf{w}_k})} (1-p)^{k-n_0(\mathbb{I}_{\mathbf{w}_k})}, \quad (3.9)$$

where $n_0(\mathbb{I}_{\mathbf{w}_k})$ denotes the number of zeroes in $\mathbf{w}_k$.

The value of $\mu(\mathbb{I}_{\mathbf{w}_k})$ depends solely on p and $n_0(\mathbb{I}_{\mathbf{w}_k})$. In practice, the coarse local Hölder exponent at x is computed by evaluating equation 3.8 at several *scales* k and an estimate estimate is produced analogously to equation 3.8.

The concept of a multifractal measure becomes clear when one observes that μ is exactly the *Binomial* measure. Specifically, that the number of points that scale with α , $N_k(\alpha)$, can be attained by computing the binomial expansion, and thus:

$$N_k(\alpha) = \binom{k}{n_0(\alpha)k} \sim (2^{-k})^{-f(\alpha_k)}, \quad (3.10)$$

so that $f(\alpha) = -\log_2(n_0(\alpha)^{n_0(\alpha)}(1-n_0(\alpha))^{1-n_0(\alpha)})$, by leveraging Stirling's approximation [64]. This relation $N_k(\alpha) \sim r^{-f(\alpha_k)}$, is homologous to equation 3.1. For measures where such a relation holds, the value of $f(\alpha)$ will exactly be the Hausdorff dimension of the following level set [64, 66, 90]:

$$\mathbb{I}^\alpha = \{x \in \mathbb{I} : \alpha(x) = \alpha\}. \quad (3.11)$$

Since f maps onto a spectrum of fractal dimensions of each $\mathbb{I}^\alpha$, it is henceforth called the *multifractal spectrum* (as postulated in Definition 3.9).

These relations can be derived for the more general Multinomial measures parameterized by m parameters so that $\mu(\mathbb{I}_{\mathbf{w}_k}) = \prod_{i=1}^m p_i^{n_i(\mathbb{I}_{\mathbf{w}_k})}$ [64].

3.5 Channel attention functions as inductive prior

Assuming we can extract either a quantity related to the box-dimension(s) in the forward step in a deep neural network, understanding how to leverage this quantity from an architectural perspective is not trivial. This is especially true for the case of dense-prediction

tasks, where the scaling exponent(s) are low resolution global descriptor(s).

Inspired by early work by Vehel et al. [91], we assume that the singularity spectrum f induces a quotient space that should correlate with fundamental visual primitives. That being said, we observe the phenomena as a sort of hierarchical decomposition g_{lc} , and our pretense is that, at layer l , the semantic level will be the same, but each learned filter g_{lc} will potentially independent (in the semantic, not spatial sense) measures on the input image. This assumption is also in line with classical computer vision methods [90], where several measures on input image $\mathbb{I}$ are induced, each responding to potentially different (multi)fractal primitives.

On the other hand, channel attention functions, namely those based on squeeze-and-excitation (SE) [92] are designed to recalibrate the outputs of a convolutional layer of filters as a function of global, statistical quantities. Remarkably, they have been shown to increase semantical segmentation performance, especially in medical imaging [93].

We thus propose two new attention functions (see Section XXX), where filters that share the same layer l are re-calibrated in a non-linear way as a function of their scaling exponents.

In the remainder of this Section, we briefly describe the original SE layer, alongside other derivative works that include increasingly more complicated statistical descriptors of each channel (filter).

In the remainder of this work, we will denote the input of a deep neural network (DNN) encoder, g , with an input tensor $\mathbf{X} \in \mathbb{I}_0$, so that $g_l(\mathbf{X}) = g_l \circ \dots \circ g_2 \circ g_1(\mathbf{X})$ where $g_l : \mathbb{I}_{l-1} \rightarrow \mathbb{I}_l$ for $\mathbb{I}_l \equiv \mathbb{R}^{H_l \times W_l \times C_l}$. Thus, $l \in \{1, \dots, L\}$ indexes the layers of g . Furthermore, $g_{lhw c}$ signifies the c -th channel, at position (h, w) of the l -th layer. Assume henceforth that this indexing is consistent throughout, and that index omission implies tensor slicing, e.g. $g_{lc} \equiv g_{l::c}$.

3.5.1 Squeeze-and-Excite (SE) [92]

Squeeze-and-excitation networks were originally studied by [93], and we will onward use the acronym cSE (channel SE) for consistency with medical image segmentation literature, even though cSE and SE are functionally equivalent.

The output of each encoder layer g_l is characterized by a “squeeze” function $g : \mathbb{I}_l \rightarrow \mathbb{R}^{C_l}$ such that:

$$h(g_l(\mathbf{X})) = \sigma(\mathbf{W}_2 \text{ReLU}(\mathbf{W}_1 \text{GAP}(g_l(\mathbf{X})))), \quad (3.12)$$

where GAP stands for spatial global average pooling, $W_1 \in \mathbb{R}^{\lfloor \frac{C_l}{s^*} \rfloor \times C_l}$, $W_2 \in \mathbb{R}^{C_l \times \lfloor \frac{C_l}{s^*} \rfloor}$, for some $1 \leq s^* < C_l$, and σ is the sigmoid activation function. The output of the encoder at level l for $l < L - 1$ thus becomes:

$$g_l^{\text{Excited}}(\mathbf{X}) = g_l(\mathbf{X}) \odot h(g_l(\mathbf{X})), \quad (3.13)$$

where $\odot$ is the (broadcastable) element-wise product.

3.5.2 Spatial and Channel SE[93]

For the purposes of semantic segmentation of medical images, it is proposed to perform the squeeze operator along the spatial dimensions, in order to have an attention function with the original feature map resolution.

The authors propose an additional branch, where the pooling operator of equation 3.12 is replaced by a learnable 1×1 convolution layer, producing $h^{\text{Spatial}}(g_l(\mathbf{X}))$. The authors propose the final activations to be

$$g_l(\mathbf{X}) = \max(g_l(\mathbf{X}) \odot h(g_l(\mathbf{X})), g_l(\mathbf{X}) \odot h^{\text{Spatial}}(g_l(\mathbf{X}))), \quad (3.14)$$

where $\max$ is performed element-wise, i.e., the function is a maxout layer.

3.5.3 Style-based Recalibration (SRM) [94]

Style-based Recalibration (SRM) integrates the standard deviation as a proxy for style in the context of style-transfer. Denoting a global standard-deviation pooling layer as GSP, the squeeze function is defined as:

$$h(g_l(\mathbf{X})) = \phi([\text{GAP}(g_l(\mathbf{X})), \text{GSP}(g_l(\mathbf{X}))]), \quad (3.15)$$

where $\phi : \mathbb{R}^{C_l \times 2} \rightarrow \mathbb{R}^{C_l}$ is a learnable linear map. An additional batch-normalization layer is applied to h .

3.5.4 Frequency Channel Attention (FCA) [95]

Frequency channel attention splits the C_l channels of each layer l in k groups such that $[g_l(\mathbf{X})_1, \dots, g_l(\mathbf{X})_k] = g_l(\mathbf{X})$. Then the two-dimensional discrete cosine transform (DCT)

filters are pre-computed and stored on a tensor $\mathbf{B}$ such that

$$B_{l,x,y}^{i,j} = \cos\left(\frac{\pi x}{H_l}\left(i + \frac{1}{2}\right)\right) \cos\left(\frac{\pi y}{W_l}\left(j + \frac{1}{2}\right)\right). \quad (3.16)$$

The squeeze operation is performed for each of the k groups

$$h(g_l(\mathbf{X}))_k = \sum_x \sum_w g_l(\mathbf{X})_k B_{l,x,y'}^{i_k,j_k} \quad (3.17)$$

i_k, j_k depend on k since there is a distinct DCT basis for each of the k groups

$$h(g_l(\mathbf{X})) := \sigma(\mathbf{W}[h(g_l(\mathbf{X}))_1, \dots, h(g_l(\mathbf{X}))_k]),$$

so that $\mathbf{W} \in \mathbb{R}^{C_l \times C_l}$.

3.6 Literature review

3.6.1 Fractals in classical computer vision

The first applications of Fractal Geometry in Computer Vision date to back the 1980s [82]. In this era, Pentland seminally argued that since the intensity of an image is a function of the angle between the illuminant and the surface normal, then if the normals follow a Fractal Brownian distribution and the perspective is kept constant, then the fractal dimension is carried to a digital photograph [96]. The introduction of the differential box counting (DBC) algorithm [97] was pivotal to increase adoption of fractal geometry in the computer vision community, since it allowed for efficient estimation of the number of boxes covering a 3D surface (e.g., a gray-scale image).

The fractal dimension has been used extensively as a descriptor for texture [90, 98, 99]. Quantities related to the more general *set of generalized dimensions* [100] embedded within morphological models and clustering approaches were also demonstrated to be effective in the semantic segmentation of textural attributes, such as mosaics of textures from the Brodatz album [101]. Other classical approaches for multifractal segmentation can be found in the references of [99, 100].

In [91], Vehel et al. propose multifractal formalism as an alternative to the traditional theory of filtering. They noted that the differentiation operation commutes with convolution, and so edges and regions of interest were found after the image was smoothed with by means of convolution with some function g . Thus, by leveraging multifractal formalism, one could potentially assign each point to a specific primitive (e.g., an edge)

depending on the scaling exponent around its (local) neighborhood, and each resulting on a *level set partition* that could be associated with the output of the convolution of the image with said filter g .

Vehel et al., also introduced the important idea of capturing this scaling behavior as a function of several different *measure functions** (such has the maximum, minimum, sum, etc.) in order to mitigate the noise sensitivity these methods display in practice. This inspired several contributions, where quantities related either to a set of fractal dimensions [99], or even the so called multifractal spectrum (MFS) [64, 90], $f(\alpha)$, were computed for a set of general filters, each defining a *measure* of the original (gray-scale) image.

Most of these early approaches were tailored towards gray-scale images, although there is some literature inspired on early applied multifractal theory [82] to RGB images have also been developed [102, 103] for the monofractal case, i.e., when the image is governed by a single scaling exponent.

3.6.2 Fractals in Modern Computer Vision

Fractal geometry has received minimal attention as a design component for end-to-end deep learning architectures until relatively recently. One of the main limiting factors is the expensive and multi-scale analysis required to compute the local scaling exponents. There is an aggravating factor the most efficient methods to compute this quantities rely on DBC, and are thus non-differentiable.

It is also not entirely clear how fractal geometry fits in architectures such as convolutional neural network (CNN)s, since they can learn arbitrary functions, and are thus not constrained to behave strictly as a set of Wavelet basis functions, as typically required in multifractal formalism. In the context of contemporary medical imaging (multi)fractal techniques have only been leveraged for static feature extraction for the purposes of classification using pre-trained DNNs [70, 104–106].

Fully end-to-end approaches have only been recently developed in the computer vision community. For texture classification, Yong Xu et al. [66] adapted their hand-crafted approach [90] to be supported end-to-end. Their work was motivated to complement global average pooling (GAP) with relaxed soft version of the MFS. After computing a soft-histogram over Q level sets, they compute a quantity proportional to the fractal dimension by max-pooling operations of increasing size for each level set Q . They fuse the

*They were actually described as capacity functions, but this claim is not correct [80].

soft MFS with GAP using bilinear pooling. Chen et al. [107] assume that a CNN is roughly a counterpart of the wavelet transform and consider that the fractal dimension is encoded throughout the layers and channels of a CNN, i.e., it as a function of a 4-dimensional tensor. They relax the DBC algorithm [97] to be used during training of the CNN. The surface tensor is constructed by imposing size-normalization constraints over spatial and channel dimensions of each layer. The proposed DBC algorithm is computed in overlapping windows and they learn a set of soft histograms for each l . Like [66] they also use a GAP layer on the embedding, but instead of bilinear pooling, they simply concatenate both feature maps.

For the task of few-shot learning, [108] introduced the fractal dimension as a prior for intra (inter) class similarity (dissimilarity), under the pretense that image regularity will differ across semantic categories. The DBC is also adapted, this time solely by removing the non-differentiable rounding functions. Since they assume a perfect linear relationship, they skip the ordinary least-squares (OLS) step required to estimate the scaling exponents [90] and collect the point-wise estimate of the DBC in order to eventually compute the so-called *fractal embedding*.

Recently, [109] proposed the Multi-layer Fractal Encoding Network (MFEN). The box-dimension is computed by exhaustive thresholding of the ℓ_2 -norm distance of all points. The fractal dimension of each layer is derived using the ridged regression method. These fractal features are concatenated with the per-layer global average pooling output and a final dense multilayer perceptron (MLP) outputs the final predictions.

3.6.3 Statistical Channel Attention Functions

Squeeze-and-Excitation (SE) Networks [92] had a profound impact in modern deep learning architectures. SE is a simple *plug in* component that allowed each encoder in a CNN to have global receptive field. This was made possible since the input at depth l is recalibrated according to some *channel attention function* g . In this case, g is the output of an MLP over the GAP of some encoder at depth $l - 1$. One of the main limitations of the original SE module is the fact that GAP is a very limited feature descriptor by virtue of (a) being a global descriptor, by (b) being the simplest statistic one can compute to describe each filter, and (c) by the representation power and compute requirements of g . The reader is referred to [110] where a very comprehensive survey of extensions to SE is presented.

In [111], the second-order statistics are modeled through the covariance matrix across channels. However, they had to reduce the total number of channels in layer l , C_l , by means of 1×1 convolution to mitigate the introduced computational burden. The authors of [94] propose the style-based recalibration (SRM) module that uses both global average and standard deviation pooling to expand the receptive field of f_l . They learn a non-linear map of the mean and variance of each channel using channel-wise dense layers. In [95], the frequency channel attention (FCA) module was introduced. The authors illustrated how GAP is proportional to the lowest frequency component of the 2-dimensional Discrete Cosine Transform (DCT).

Specifically to Medical Imaging Segmentation, [93] propose recalibration using two branches: the *channel squeeze spatial excite* branch (cSE), a direct implementation of the original SE module, and the *spatial squeeze channel excite* (sSE) branch, that acts as spatial attention [112]. The output layer is an element-wise max-out between these two branches (scSE). The pursuit of channel-wise descriptions was further explored in [113, 114], but explicitly tailored to 3D segmentation and without emphasis in points (a) and (b), but in terms of (c). Related to the latter, other approaches demand an increasing amount of compute and/or architectural changes, such as introducing Transformers [115–117].

3.7 Methods

Our contribution relates more closely with the *Histogram Method* given that we are focusing on imaging modalities. Empirical results appear to favor the method in this context of Computer Vision, especially if one computes several different measures (i.e., filters) over the same *static* phenomena (i.e., image) [90, 91, 118].

The *Method of Moments* is arguably more adopted in general. It uses the *Legendre* transform to write f and α as functions of the normalized moments of μ , $\tau(q)$:

$$f(\alpha(q)) = q\alpha(q) - \tau(q) \text{ so that } \alpha(q) = \frac{\partial \tau(q)}{\partial q}, q \in \mathbb{R}. \quad (3.18)$$

It is also pivotal for defining the *set of generalized dimensions* D_q , and the interested reader is referred to [63, 119] for additional details. For this paper, we mainly use the link between D_q and $f(\alpha)$ as a convenient formal construct for our proofs.

There are also other ways of computing f by using Wavelet transforms or by detrended fluctuation analysis (DFA), and we recommend [120] for an in-depth overview of these methods.

We look at the l -th output of encoder head $g_l(\cdot) = [\mu_{l_{c_1}}, \dots, \mu_{l_{c_l}}]$, $\mu_{lc} \geq 0 \in \mathbb{R}^{H_l \times W_l}$, i.e., a tensor of unnormalized measures (filters) that we assume are meaningful for the downstream task.

We adjust the responses of the encoder using *learnable functions* of the local (multifractal) or global (monofractal) interactions between the singularity strengths derived across each channel/measure of f_l . We start by demonstrating that the proposed recalibration functions indeed relate to quantities that relate to $f(\alpha)$.

Our proofs will consider measures multinomial measures ν of the form defined in 3.10.

Definition 3.10 (Multinomial measures). Let $\nu = (\mu_q)_{q=1}^Q$ be a multinomial measure. By construction, each μ_q is a binomial measure with potentially unique support. Let $p^{(q)}(\alpha^{(q)})$ be the normalized density of each exponent in ν so that:

$$p(\alpha_\nu) := [p^{(1)}(\alpha^{(1)})/Z, \dots, p^{(Q)}(\alpha^{(Q)})/Z], \quad (3.19)$$

where $Z = \sum_q p^{(q)}(\alpha^{(q)})$.

Note that these multifractal measures have enjoyed great success in practice [63, 118]. We also introduce learnable parameters in their implementation, thus allowing to characterize the scaling nature of other multifractal phenomena beyond multinomial cascades.

3.7.1 Differentiable scaling exponent computation

As mentioned in Section 3.4.2, the Hölder exponent can be computed in an analogous fashion to the box dimension equation 3.8. Defining a finite scale set $\mathcal{R} = \{k_1, \dots, k_r : i < j \implies k_i < k_j\}$, we can derive the following dyadic cubes $\mathbb{I}_{\mathbf{w}_k} = [i_h - 2^{-k}, i_h + 2^{-k}) \times [i_w - 2^{-k}, i_w + 2^{-k})$. For a cube $\mathbb{I}_{\mathbf{w}_k}$ centered on x we can approximate the local singularity strength for the c -th channel of the l -th layer as (power) the power-law relationship manifested at finitely many resolutions k . Onwards, we slightly abuse of notation by letting $\mu(B_k(x)) \in \mathcal{B}(\mathbb{I}) \equiv \mu(\mathbb{I}_{\mathbf{w}_k})$, and we approximate the local Hölder equation 3.8 exponents by solving:

$$\alpha(x) \approx \text{slope of } (\log \mu(B_k(x)), \log k)_{k \in \mathcal{R}},$$

for all x . Concretely, we attain this value through the ordinary least squares solution:

$$\alpha(x) = \frac{\sum_k \left(\log \mu(B_k(x)) - \frac{1}{|\mathcal{R}|} \sum_{k'} \log(\mu(B_{k'}(x))) \right)}{\sum_k \left(\log k - \frac{1}{|\mathcal{R}|} \sum_{k'} \log k' \right)}. \quad (3.20)$$

In our implementation, $\mu(B_k(x))$ is computed by convolving each g_{lc} with a kernel of ones of size k and stride 1 (a static depth-wise convolution layer in practice). This operation extends trivially to tensor operations, so $\alpha_{lc}(x)$ can be efficiently computed by performing this operation in a depth-wise convolution layer, readily available in modern deep learning frameworks^{*}.

Since we are concerned with digital images, onward we assume

$$\text{supp } \mu_l \subseteq \{x : x \in [1, \dots, H_l] \times [1, \dots, W_l]\}.$$

3.7.2 Monofractal Recalibration

With a slight abuse of notation, suppose we collect each Hölder exponent in:

$$\boldsymbol{\alpha} := (\alpha_k(x))_{x \in \text{supp } \mu} \quad (3.21)$$

and let $\alpha_k(\cdot)$ be an approximation of equation 3.20 using resolutions $\mathcal{R} = \{1, \dots, 2^{-k}\}$.

Theorem 3.11 (Fractal recalibration). *Suppose that μ is a multinomial measure as in equation 3.10 with multifractal spectrum f , then almost surely (a.s.) for $\mathcal{R} = \{1, \dots, k\}$ for sufficiently large k :*

$$\mathbb{E}_x[\boldsymbol{\alpha}] \propto D(\text{supp } \mu),$$

where D is the box-dimension of the support.

Proof. We prove the non-trivial case for non-singular multinomial measures using an argument similar to Evertsz and Mandelbrot [64, Section 2.4]. By definition $\mu(\mathbb{I}) = 1$ and so $\mu(\mathbb{I}^\alpha) \leq \mu(\mathbb{I})$. Note also that $\mu(\mathbb{I}^\alpha) = r^{-f(\alpha)} r^\alpha \leq 1$ (refer to equation 3.10), and thus it must be that $f(\alpha) \leq \alpha$. For multinomial measures, $f(\alpha)$ is concave. Also for this case, and referring back to the Legendre transform in equation 3.18, we can assume $q \geq 0$, so the exponent α that maximizes f needs to verify $\frac{\partial f(\alpha(q))}{\partial \alpha} = 0$ — this extremum is a maximum if and only if $q = 0$, which is equivalent box dimension of the support of the measure [63]. Referring back to the shape of f (see Eq. equation 3.10 or Figure 3.2), clearly $\mathbb{E}[\alpha]$ maximizes $f(\alpha)$.

^{*}E.g., PyTorch through `torch.nn.Conv2d` (with `groups=in_channels`) or TensorFlow's `DepthWiseConv2D`.

In order to show that the expectation over x relates to this quantity, we again extend the arguments of Evertsz and Mandelbrot in [64, Section 4.2]: leveraging the law of large numbers, for any $x \in \text{supp}(\mu)$:

$$\alpha(x) \stackrel{\text{a.s.}}{=} \mathbb{E}(\alpha), \text{ where } \alpha \text{ in the r.h.s. is in the domain of the MFS of } \mu: (\alpha, f(\alpha))$$

Denoting $|\text{supp } \mu| = n \in \mathbb{R}^+$, it follows that $\mathbb{E}_x[\alpha] \stackrel{\text{a.s.}}{=} n^{-1} D(\text{supp } \mu)$ $\square$

This theorem provides a fundamental explication to empirical observations: estimations of α tend not to vary meaningfully in the support of the measure [63, 120].

Lemma 3.12 (Monofractal recalibration). *If μ is uniform and supported on fractal set $\mathbb{F}$, μ is said to be monofractal, i.e., it is governed by a single scaling exponent. Under these conditions, the multinomial measure will reduce to the Lebesgue measure λ . However, $\mathbb{E}[\alpha(x)] \propto D(\mathbb{F})$ as in equation 3.1, even though $\lambda(\mathbb{F}) = 0$.*

Proof. The probability mass is uniform so $\mu(\mathbb{I}^{\alpha_i}) = \mu(\mathbb{I}^{\alpha_j})$ for any i, j , i.e., the mass is evenly distributed with respect to the terms of the m -ary expansion of any x , and so $\mathbb{I} = \{\alpha : x \in \text{supp } \mu\}$. The MFS thus collapses to a single value α , which can be recovered using the expectation operator in Theorem 3.11. $\square$

It is not uncommon to assume that α will not vary meaningfully in the support of the measure in many practical applications [63, 120], which is to say that this inductive prior holds even if the MFS is not as well-behaved as our ideated multinomial spectra.

3.7.3 Multifractal recalibration

The previous recalibration strategy is suitable for when each (learned) measure μ is well characterized by a monofractal. Notwithstanding, it does not characterize the entirety of the MFS, it can only capture a value associated with the dimension of its support.

Remember from Section 3.4.2 that $f(\alpha)$ is the *Haurdorff-Besicovich* dimension of $\mathbb{I}^\alpha$. We know that $\mu(\mathbb{I}^\alpha) = N_k(\alpha)(2^{-k}) \sim 2^{-f(\alpha)+\alpha}$. If one treats the Hölder exponent of a binomial cascade at the k -th scale α as a random variable, then by the Gaussian Central Limit theorem, for sufficiently large k , the following two statements hold [64]:

$$p_k(\alpha) \propto \exp \left\{ \frac{1}{2} \left(\frac{\alpha - \alpha_0}{\sigma/\sqrt{k}} \right)^2 \right\} \quad (3.22)$$

around most probable scaling exponent α_0 . Recall that $\mathbb{E}_x[\alpha] \rightarrow \alpha_0$ as $k \rightarrow \infty$. Should μ be supported on a set with Hausdorff-Besicovich dimension D [64]:

$$f_k(\alpha) \approx D + \frac{1}{k} p_k(\alpha). \quad (3.23)$$

Hence, for our purposes, it is fair to assume that the support has fixed dimension, so our inductive bias can be on p_k , since $p_k \propto f_k$.

It is also very possible that some phenomena may not be aptly described solely by this quantity. Indeed the conditional likelihood of the singularities may vary depending on the context. We set forth *Multifractal Recalibration* to address this issue, assuming there the quantity we want to model may also depend on the context (i.e., the desirable mapping $X \mapsto Y$).

Definition 3.13 (Data-dependent multinomial measure). Suppose we build μ^* by weighing multinomial μ as a function of a labeling function $l : \mathbb{I} \rightarrow \{0, 1\}$ defined as the restriction $\mathbb{Y} = \{x \in \mathbb{I} : l(x) = 1\}$. Then, the data-dependent multinomial measure is given by:

$$\mu^*(\mathbb{I}) = \frac{\mu(\mathbb{I} \cap \mathbb{Y})}{\mu(\mathbb{Y})}$$

Note that μ^* has the same scaling behavior as μ , but it assigns all mass to restriction $\mathbb{Y}$.

Lemma 3.14 (Multifractal recalibration). Consider all points x in some level set indexed by q , i.e., $\mathbb{I}^{\alpha(q)} \subseteq \mathbb{I} = \{x : \alpha(x) = \alpha(q)\}^*$. Consider measures μ, μ^* supported on $\mathbb{I}$, the latter being constructed from μ and some labeling function $l(x)$. Then any measure $\hat{\mu}$ that verifies:

$$\hat{\alpha} = \arg \max_{\alpha'} \frac{1}{Z} \exp(-(\alpha' - \alpha^*)^2),$$

has to have perfect level-set likelihood ratio $\frac{\hat{\mu}(\mathbb{I}^{\alpha(q)})}{\mu^*(\mathbb{I}^{\alpha(q)})} = 1 \forall q$, where $Z = \int \exp\{-(\alpha' - \alpha^2)\} d\alpha$.

Proof. According to the definition of coarse Hölder exponent equation 3.8, for any x, q such that $x \in \mathbb{I}^{\alpha(q)}$:

$$\alpha'_q(x) := \frac{\log \mu'(\mathbb{I}^{\alpha(q)})}{\log 2^{-k}} \text{ and } \alpha_q^*(x) := \frac{\log \mu^*(\mathbb{I}^{\alpha(q)})}{\log 2^{-k}}.$$

*Using index q is merely evocative, but not equivalent to, using the Legendre transform.

For sufficiently large fixed k , we can disregard the denominators. Moreover, since Z does not depend on q , pick arbitrary $\alpha(q)$:

$$\begin{aligned} \alpha'_q(x) - \alpha_q^*(x) &\approx \log \mu'(\mathbb{I}^{\alpha(q)}) - \log \mu^*(\mathbb{I}^{\alpha(q)}) \\ \implies \exp \left\{ - \left(\log \mu(\mathbb{I}^{\alpha(q)}) - \log \mu^*(\mathbb{I}^{\alpha(q)}) \right)^2 \right\} &= \exp \left\{ - \left(\log \frac{\mu'(\mathbb{I}^{\alpha(q)})}{\mu^*(\mathbb{I}^{\alpha(q)})} \right)^2 \right\}. \end{aligned}$$

Thus the maximizer α' of the above stated Gaussian function must have level-set likelihood ratio $\frac{\mu'(\mathbb{I}^{\alpha(q)})}{\mu^*(\mathbb{I}^{\alpha(q)})} = 1$ for all q . $\square$

3.7.4 The Learning Triplet for Fractal Recalibration

We now instantiate configurations $\mathfrak{C} = (\Phi, \mathcal{F}, \mathcal{L})$ for the domain of irregular, fine-structured segmentation. In this framework, the critical shift lies in the Hypothesis Space: we move away from learning mappings based solely on Euclidean geometry (edges, shapes) and instead constrain the model to value features based on their complexity (scaling behavior).

- **Data Processing (Φ):** Here, Φ represents the acquisition and representation of domain-specific data (e.g., biological tissues, material surfaces) where objects exhibit fine structure at arbitrary scales.
- **Hypothesis Space ($\mathcal{F}$):** The hypothesis space is defined by U-Net architectures with convolutional encoders that are recalibrated using tFractal Channel Attention. By incorporating the differentiable Hölder exponent computation (Eq. 3.20) directly into the architecture, we enforce a specific constraint: the utility of a learned filter g_{lc} is determined not by its activation magnitude, but by its scaling behavior over the resolution set $\mathcal{R}$.

We distinguish two specific complexity-based hypotheses:

1. **Monofractal Hypothesis ($\mathcal{F}_{\text{mono}}$):** We assume that for a given semantic class, complexity is a collection of global invariant. The inductive bias, supported by Theorem 3.11, is that the "correct" features are those that exhibit a stable **global fractal dimension**. The attention mechanism recalibrates the network to favor channels where the expected singularity strength $\mathbb{E}[\alpha]$ aligns with the constituents of the region of interest.

2. **Multifractal Hypothesis ($\mathcal{F}_{\text{multi}}$):** This hypothesis constrains the recalibration at each encoder level to share a *distribution* of singularity exponents. The inductive bias here is that the feature maps are weighted according to an explicit quantity that relates directly to their **multifractal Spectrum**. As suggested by Lemma 3.14, this also allows the model to distinguish between regions that may share the same support but differ in response with respect to texture regularity.

- **Loss Function ($\mathcal{L}$):** While the extrinsic optimization objective remains a standard segmentation loss (e.g., Dice/Cross-Entropy), the Fractal Recalibration modules introduce an **intrinsic regularization**, which depends on scaling laws captured at each layer of the encoder.

3.8 Application: medical image segmentation

Remembering Definition 3.13, our goal is to given that we observe a irregular surface (μ) but only certain, usually connected subsets are of interested for a particular task (μ^*). For instance, in medical image segmentation tasks (see Figure 3.3), it is often true that all the scene is highly irregular. Notwithstanding, the downstream task usually pertains to identifying a peculiar type of irregularity (likely characterized by more than one scaling exponent) typical of a certain pathology.

In Medical Imaging, both structural and surface-level regularities may be leveraged [63]. These complex patterns often emerge in systems where agents interact, and thus organize, under a set of simple rules that are iterated throughout time [121] and is commonly observed in biological systems. Consequently, entire new fields have emerged in recent history, such as *Fractalomics* [65], where statistical quantities based in Fractal Geometry are leveraged to quantitatively describe morphological and structural complexity of cellular or biological phenomena [122]. Pathologies such as tumors have also described as behaving *statistically* as fractals, in both space and time [63, 123]. Examples range from histology [105, 124, 125], endoscopy [70, 118], dermoscopy [126, 127], to magnetic resonance imaging (MRI) [128–130] and many others (see [63, 131] and references therein). On the other hand, Fractal Geometry has received limited attention in the deep learning era, especially in the design of end-to-end architectures. This trend recently changed, and some successful approaches have been proposed for texture classification [66, 107, 109]

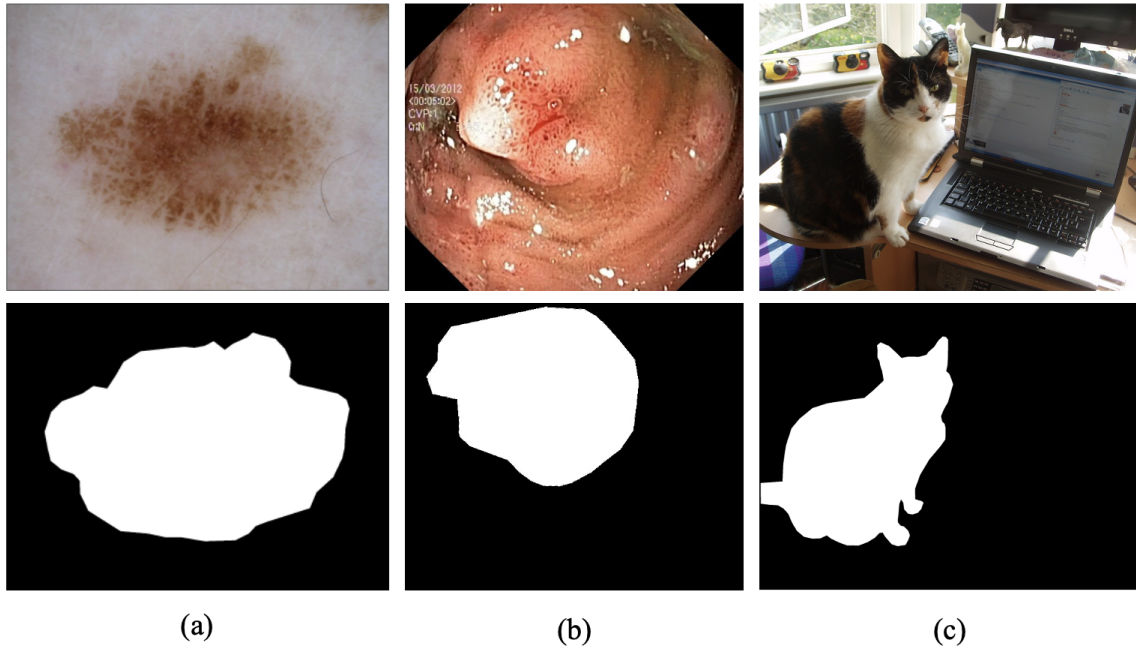


FIGURE 3.3: Sample images from images (top) and respective class-specific segmentation masks (bottom) from (a) ISIC-18, (b) KvasirSeg, and (c) COCO datasets. Note how the biomedical regions of interest are hardly described by their shape. In fact, looking at the masks in isolation, the only plausible inference one can make is that of the cat in (c).

and few-shot learning [108]. Bridging this gap for semantic segmentation is the main contribution of this paper, under the assumption that the inductive prior of a CNN naturally encodes information about the image’s scaling laws.

3.8.1 Multifractal analysis in Medical Imaging

Physiologic systems have been described as inherently (multi)fractal since they evolve in a self-seeding, cascading fashion [63]. In fact, the theory of fractal and self-organizing structures has been proposed as fundamental design principles of life itself [132]. *Fractalomics* [65] have been proposed to leverage statistical quantities based in fractal geometry due to its effectiveness at describing quantitatively morphological cellular structure complexity of cellular and biological tissues. Pathologies such as tumors have also described fractal properties in both space and time [123]. For example, in dermoscopy, the fractal dimension was used to associate the regularity of the boundary of dermatological lesions with their severity [126]. This was eventually extended to include lesion texture to enhance classification [127]. It has been used to segment calcification or regress bounding boxes in lesions captured in breast ultrasounds [133, 134]. It was also adopted for MRI

to segment or classify tumors [128–130]. Its power to derive texture features significantly contributed to the increase of classifier accuracy in several histopathology modalities [105, 124, 125] as well as polyp and mucosal characterization in endoscopy [70, 118].

In fact, fractal and multifractal intersect with virtually all medical modalities in such a profound way that a thorough review of the art would require its own dedicated chapter, so we recommend the interest reader to [131, 135]. In the case of medical images, although the number of use-cases is broad, the types of techniques is relatively narrow, and does not differ significantly from the rest of the computer vision community (see Section 3.6.1).

3.8.2 Implementation

Our implementation* is geared towards the U-Net (see Figure 3.4) but the principles set forth in this chapter maybe easily extended to other types of DNNs. Remember that the recalibration functions are always performed at the output of each encoder head, g_l , and a decoder head d_l , for $l = 1, 2, \dots, L$.

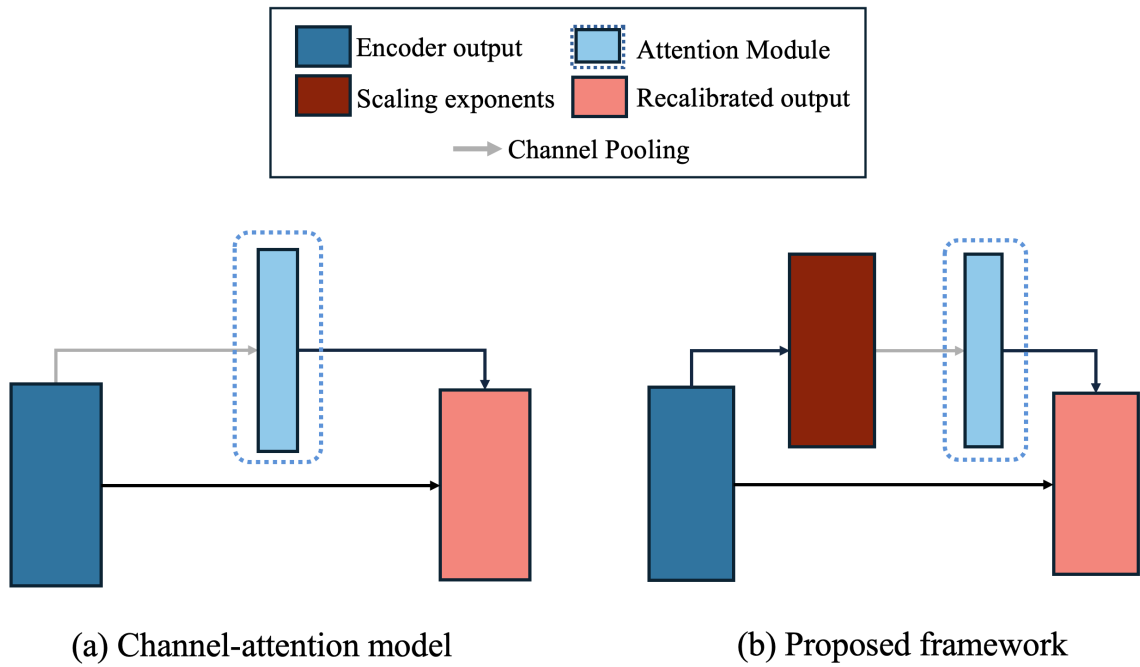


FIGURE 3.4: (a) The typical channel attention model where some statistics is pooled directly from the encoder output. (b) Our proposed approaches use statistics derived from the scaling exponents of the features maps.

The scaling exponents of each g_l are not normalized (see equation ??), which can compromise optimization during gradient descent[109, 136], so in this section H_l signifies the

*Implementation available: <https://github.com/miguelmartins/multifractal-recalibration>.

empirical singularity exponents after the action on a *distinct* batch-normalization layer for each l .

In order to compute the expectation in 3.11, and hence the attained Box-dimension, is used as the global channel descriptor, which is fed to a squeeze-and-excitation pathway [92]. The output of g_l will be recalibrated by a function $h : \mathbb{I}_l \rightarrow \mathbb{R}^{C_l}$ so that:

$$h^{\text{Mono}}(g_l(\mathbf{X})) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \text{GAP}(\mathbf{H}_l))), \quad (3.24)$$

where GAP stands for spatial global average pooling, $\mathbf{W}_1 \in \mathbb{R}^{\lfloor \frac{C_l}{s^*} \rfloor \times C_l}$, $\mathbf{W}_2 \in \mathbb{R}^{C_l \times \lfloor \frac{C_l}{s^*} \rfloor}$, for some $1 \leq s^* < C_l$, and σ, δ are the sigmoid and rectified linear unit (ReLU) activation functions, respectively. The output of the encoder at level l for $l < L - 1$ thus becomes:

$$g_l^{\text{Mono}}(\mathbf{X}) = g_l(\mathbf{X}) \odot h^{\text{Mono}}(g_l(\mathbf{X})). \quad (3.25)$$

This output is propagated to g_{l+1} , but also to the decoder in case of the U-Net by means of a skip connection.

Note that 3.14 allows for several characteristic scaling exponents that are calibrated as a function of the dependent variable (hence the term *Multifractal recalibration*). This allows for the characterization of the entirety of the scaling spectra (refer to Section 3.4.2, Figure 3.2). The morphology of the MFS is also relaxed due to the data-driven partitioning of $\mathbb{I}_l$, which is desirable when the MFS morphology differs from a perfect concave parabola [64, 88, 89].

We will characterize the scaling exponents for all channels of each g_l as Q jointly learned centers. Specifically, by direct application of 3.14:

$$p_l^{(q)}(\mathbf{H}_l) \propto \frac{\exp(-s_q^*(\mathbf{H}_l - \mathbf{H}_{lq}^*)^2)}{\sum_{q'} \exp(-s_{q'}^*(\mathbf{H}_l - \mathbf{H}_{lq'}^*)^2)}, \quad (3.26)$$

where $q \in \{1, \dots, Q\}$, and $s_l^{(q)} \in \mathbb{R}$, $\mathbf{H}_l^* \in \mathbb{R}^Q$ are learnable parameters. Each layer g_l is thus partitioned in Q stochastic level sets (see Figure 3.5). This definition implies parameter sharing across spatial and channel axes, so if one draws parallels with the *Histogram Method* [64, 90, 99], it means we learn the same quantization of the level sets for all measures in the same layer l . This pattern can be easily implemented using a dense layer with softmax activation function resemblant of [66, 137].

We surmised that each stochastic level set defined in equation 3.26 should be weighted differently depending on the downstream task. We opt for a weighted sum aggregation

and learn an orderless pooling function over q , a common approach in texture classification [66, 107, 137]. We propose to pool the response of H_l across the Q -axis, using a batch-normalization layer ϕ , so that:

$$\tilde{H}_l := h^{\text{Multi}}(g_l(\mathbf{X})) = \sigma \left(\sum_q \delta(\phi(p_l^{(q)}(\alpha_l^{(q)}))) \right), \quad (3.27)$$

where σ, δ are the sigmoid and ReLU activation functions, respectively. Without the activation functions, this aggregation is a mixture of Q Gaussians (as in Figure 3.5). A similar type of pooling is used in context encoding for semantic segmentation [137], and the permutation invariance it enforces is also a desired property when dealing with textural patterns [66].

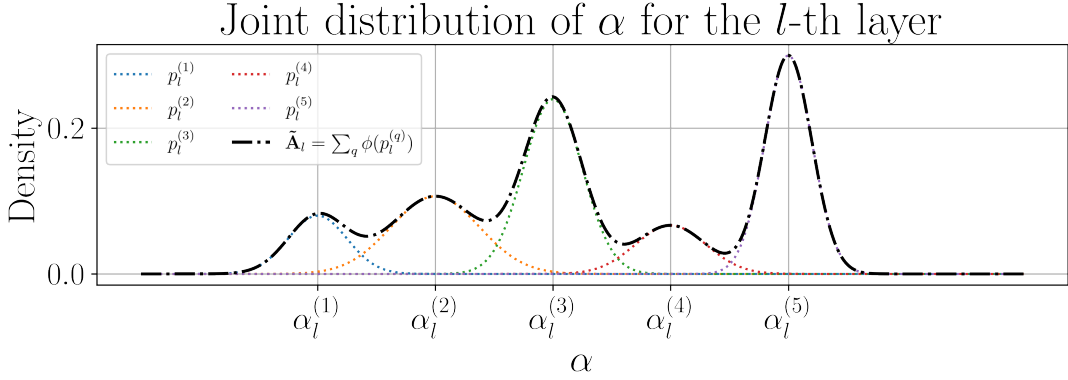


FIGURE 3.5: Visualization of the joint encoding of p_l at encoder g_l . Each level set is characterized by a Gaussian $p_l^{(q)}$ localized around $\alpha_l^{(q)}$.

The final step is to fuse this representation with the response of g_l . We opt for a similar strategy to [108]:

$$g_l^{\text{Multi}}(\mathbf{X}) = g_l(\mathbf{X}) + \tilde{H}_l, \quad (3.28)$$

where $+$ is the element-wise sum. Note that, by the linearity of convolution, in the l -th decoder head d_l , $g_l^{\text{Multi}} * d_{lc} = (g_{lc} * d_{lc}) + (\tilde{H}_{lc} * d_{lc})$, so in this way we ensure that singularity information may be easily disentangled by the decoder filter d_{lc} .

3.9 Experiments

We conducted experiments in three different public datasets from medical imaging modalities that vary significantly in terms of equipment and interventions used. These modalities also present challenging multi-scale characteristics, as well as variations in intensity, lighting, noise artifacts, and texture. Specifically, we selected the ISIC18 (dermoscopy [138]), KvasirSEG (endoscopy) [139], and BUSI (breast ultrasound) [140] datasets. As discussed in [141], using three datasets is the median number of recent contributions towards general models for medical image segmentation, provided that the datasets are varied and of sufficiently high quality (i.e., low inter-method and high intra-method variability).

3.9.1 Experimental setup

As prescribed by [141] we enforce a 5-fold cross-validation setup for all datasets. For each fold, we leave out 10% of each train partition to be used for early stopping. For the case of the BUSI dataset, each split is always stratified over the pathological classification variable since the ratio of benign to malign samples is roughly 2:1 [140]. For all datasets, all image/mask pairs were down-sampled to a 224×224 spatial resolution using bilinear interpolation and then normalized to have channel values in $[0, 1]$. Standard data augmentation was enforced using random horizontal and vertical flips.

We validate our approaches against the original squeeze-and-excitation (SE) module [92], the concurrent spatial and channel squeeze + channel excitation (scSE) proposed in [93], the style based recalibration (SRM) [94], and the frequency channel attention (FCA) [95] module. For FCA we used the best static configuration comprised of the 16 coefficients associated with the lowest frequencies of the DCT base [95]. Other than scSE and Multifractal recalibration, we call the other attention function *global* channel attention functions since they do not recalibrate filter responses at the “pixel level”.

We use the same baseline as [117], comprised of three encoder/decoder pairs with 32, 64, and 128 channels, and a bottleneck of 256 channels. In cSE, scSE, FCA, and Monofractal recalibration, the MLP architecture is the same, and we set $r^* = 2$ following [93]. For our proposed approaches, we set $\mathcal{R} = \{2, 3, 4\}$ to compute H_l (see Section 4.5.2). For Multifractal recalibration we learn $Q = 16$ stochastic level sets.

Our main research question is whether we can integrate the *multifractality* of $\mathbf{X}$ end-to-end to predict $\mathbf{Y}$. However, our target variables $\mathbf{Y}$ are very regular, even on the borders, likely due to limitations in the annotation procedures (see Fig. 3.6). We contend that,

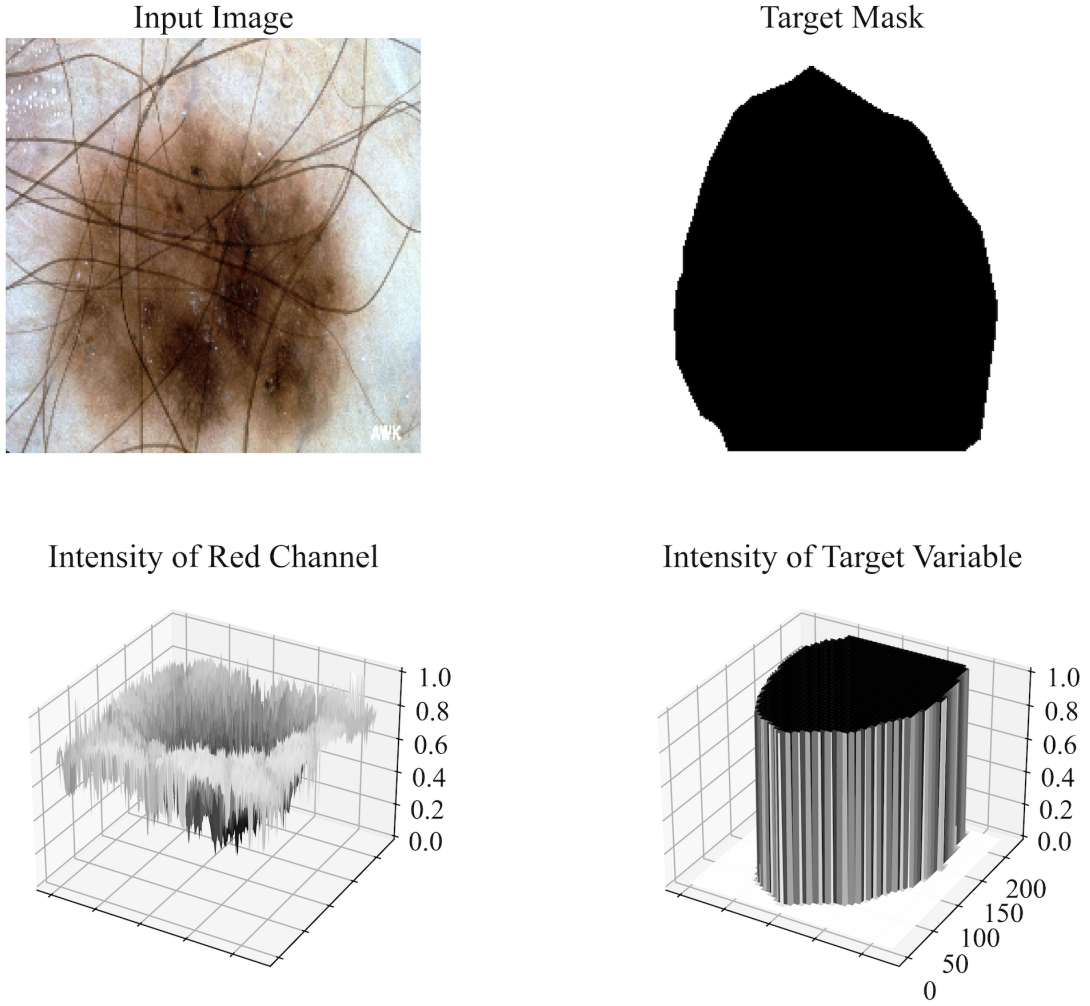


FIGURE 3.6: The intensity distribution of the input is highly irregular. Conversely, the target variable is very homogeneous and only relates information of the general region of interest. Sample taken from the ISIC18 dataset.

in isolation, the decoder is not explicitly encouraged to approximate an irregular function. With this in mind, and in order to streamline our experimental processes, we study the impact of the recalibration functions solely after each encoder output (before skip-connection and max-pooling as in [93, 113, 114]).

All models were trained using a mini-batch size of 16. Note that this value is empirically found to be the smallest mini-batch size that ensures that the modules that leverage batch normalization are not at a disadvantage [142]. The Adam gradient descent algorithm [143] was used with an initial learning rate of 1×10^{-4} over 400 epochs. A scheduler adjusted the learning rate by a factor of 0.5 should a plateau of the training procedure be detected with a patience of 10 epochs. We ensured that these fixed experimental parameters allowed all models to converge. All models were evaluated in terms of their

Dice-Sørensen score.

3.9.2 Results

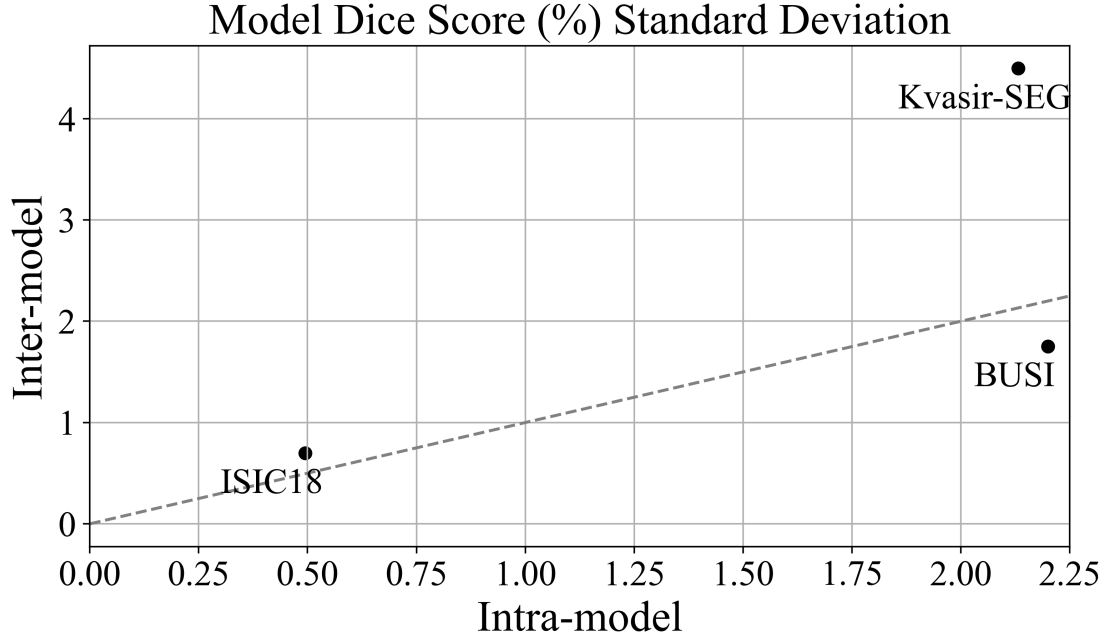


FIGURE 3.7: Dice score (%) inter and intra-model standard deviation per dataset. The dashed gray line marks a ratio of inter/intra model variability of 1.

The 5-fold experiments reveal that at least one of the proposed approaches always outperforms all others for all datasets (see Table 3.1). Both cSE and Multifractal recalibration are always beneficial on average when compared to the baseline. However, Multifractal recalibration is the only approach that improves the baseline in a statistically significant manner throughout all datasets. Both of the proposed approaches simultaneously show superior performance to all others in the ISIC18 and BUSI datasets.

It is important to analyze how one should weigh the evidence per dataset, so following [141] we analyze the ratios of intra and inter-model standard deviation for each dataset (see Figure 3.7). Both ISIC18 and Kvasir-SEG are above the threshold where the inter-model variance is of a higher average magnitude than intra-model. BUSI is below this threshold, but it is the only dataset where all of the approaches displayed a dramatic improvement over the baseline, with the highest differences being for SRM and Monofractal recalibration, +5.89% and +6.8% Dice, respectively. This dataset is thus less reliable for

TABLE 3.1: Mean $\pm$ standard deviation Dice score (%) of the cross-validation experiments. † and ‡ signify that the null-hypothesis of the pairwise t-test with regards to the U-Net baseline is rejected with $p \leq 0.05$ and $p \leq 0.01$, respectively. Best mean results are boldfaced.

Model	ISIC18	Kvasir-SEG	BUSI
U-Net [144]	85.40 $\pm$ 0.25	72.22 $\pm$ 1.82	62.20 $\pm$ 2.40
+cSE [92]	85.94 $\pm$ 0.36 †	72.72 $\pm$ 1.52	65.36 $\pm$ 1.36
+scSE [93]	85.92 $\pm$ 0.29 †	72.94 $\pm$ 1.07	64.82 $\pm$ 1.03
+SRM [94]	84.33 $\pm$ 1.27	61.13 $\pm$ 3.42	68.09 $\pm$ 3.14 †
+FCA [95]	86.19 $\pm$ 0.75	70.00 $\pm$ 2.51	66.27 $\pm$ 2.48
+Mono (ours)	86.24 $\pm$ 0.27 ‡	71.86 $\pm$ 2.37	69.00 $\pm$ 2.53‡
+Multi (ours)	86.26 $\pm$ 0.28‡	74.76 $\pm$ 2.20†	66.94 $\pm$ 2.45 †

assessing differences between attention functions, but it is still valid to show that attention functions can improve baseline performance drastically for this task.

It is noteworthy that our approaches are always superior to FCA, which is the most intricate statistical channel descriptor. SRM also captures second-order information, but it is only competitive in the BUSI dataset, and fails to even match baseline performance in the other cases.

Monofractal recalibration appears to have greater impact on performance when it works well, compared to cSE and even scSE. An interesting result is that cSE was superior on average to scSE in the BUSI dataset, and coincidentally, in the same dataset, Monofractal outperformed all other approaches, including Multifractal recalibration. We hypothesize that the speckle noise present in the ultrasound images [134] might lead to local singularities that are still propagated by g_l , which leads to diminishing returns faster for these spatial approaches. Keep in mind that in spite of this fact, the Multifractal approach still shows sizable statistical significant improvement over the baseline for this dataset, and that our analysis ranks BUSI lowest with regards to distinguishing differences between extensions of the baseline U-Net (see Figure 3.7). Despite this, Multifractal recalibration still increased baseline Dice score significantly.

Interestingly, Monofractal recalibration alongside the other sophisticated global statistical channel attention functions (SRM and FCA) under-performed when compared not only to cSE and scSE, but also to the baseline. On the other hand, the superior performance of Multifractal recalibration in the same dataset attests to the fact that higher order

TABLE 3.2: Mean $\pm$ standard deviation dice score (%) for values in $Q \in \{2, 4, 8, 16\}$ measured on the ISIC18 dataset. Best mean results are boldfaced.

$Q = 2$	$Q = 4$	$Q = 8$	$Q = 16$
86.21 ± 0.49	86.21 ± 0.32	86.15 ± 0.28	86.26 ± 0.28

statistical information is crucial for Kvasir-SEG.

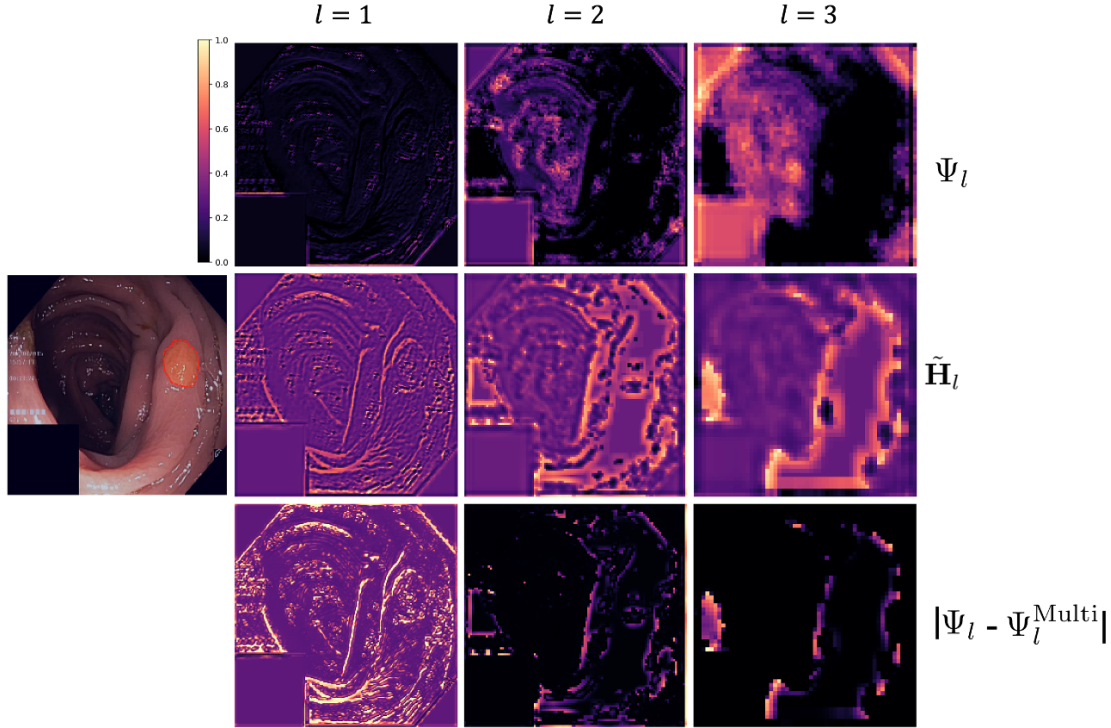


FIGURE 3.8: (Left-most) The input KvasirSEG image (from validation set) and target region of interest (highlighted in red). (Right) Layer depth $l \in \{1, 2, 3\}$ versus normalized response of g_l , H_l , and $|g_l - g_l^{\text{Multi}}|$. Note how H_l encodes complementary textural information of g_l for $l = 1$. For $l = 2, 3$ the preferred singularities relate more to luminance changes that highlight anatomical structures. This also illustrates the theory set forth by Vehel et al. [91], where each level set is associated with a distinct visual primitive.

We investigated what kind of information is captured in the singularity map H_l associated with g_l (see Figure 3.8). Qualitatively, the singularities appear to highlight general rugosity patterns latent in the feature maps. Multifractal recalibration effectively clusters this singularity information into macro-structures that become especially visible as $l \rightarrow L$.

3.9.3 Multifractal hyperparameter and aggregation analysis

In order to understand the impact of the number of soft bins in the histogram of Equation 3.26 we vary the value of the number Q mixtures so that $Q \in \{2, 4, 8, 16\}$. We measure the average and standard deviation of the Dice score in a 5-fold cross-validation

setup of the ISIC18 dataset. The results are collected in Table 3.2. Most of the performance is achieved immediately at $Q = 2$ histograms. This result becomes less surprising by realizing that this method shares some commonalities with the context encoding proposed in [137], where Q is always set to the number of classes. However, contrary to [137], we do not need any additional architectural changes nor loss terms during supervision. The increase in performance with bigger Q is not significant. However, at $Q = 16$ the average performance does increase slightly while simultaneously decreasing the amount of variance. We also found that this value of Q worked well in general for our experiments, as described in Section 3.9.2.

TABLE 3.3: Mean $\pm$ standard deviation of Multifractal of activation function pooling ablation in the Kvasir-SEG dataset (Dice (%)). The best mean results are boldfaced.

$\sum_q \phi(\cdot)$	$\sigma(\sum_q \phi(\cdot))$	$\delta(\sum_q \phi(\cdot))$	$\sigma(\sum_q \delta(\phi(\cdot)))$	$\delta(\sum_q \delta(\phi(\cdot)))$
73.91 ± 1.67	73.87 ± 1.38	73.66 ± 1.69	74.76 ± 2.20	69.82 ± 2.37

Next, we inquire about the design of the aggregation strategy of $\phi(p_l^{(q)}(\mathbf{H}_l))$ over axis Q . These results are collected in Table 3.3, which can be broken down into two parts: the choice of (non) linearity inside and outside the pooling aggregation. ReLU is the only candidate inner activation function, since it is already known to be empirically effective [137].

Two surprising facts about our results: the sigmoid in the outer activation function leads to the ultimate performance configuration, and the linear model is the second best in terms of average performance. Concerning the latter, as described in Section 3.7.3, this aggregation can be seen as a linear mixture of Gaussians, and since $Q = 16$ from the previous ablation experiment, this model should already be highly discriminative. However, the effectiveness of the former deserves more careful consideration. Remember that $g_l^{\text{Multi}} * \mathcal{F} = (g_l * \mathcal{F}) + (\tilde{\mathbf{H}}_l * \mathcal{F})$, where $\mathcal{F}$ can be either a decoder filter at depth l or an encoder filter at depth $(l + 1) \leq L$ (we omit MaxPooling2D here for simplicity). The output of this convolution depends linearly on g_l and $\tilde{\mathbf{H}}_l$, and since $g_l, \tilde{\mathbf{H}}_l \geq 0$, the sum aggregation bounds the response of this convolution. Choosing the sigmoid makes it so that $\tilde{\mathbf{H}}_l$ gates the point-wise magnitude of this convolution, similarly to an SE pathway, but in the spatial dimension. It also prevents the result of this operation from being dominated by $\tilde{\mathbf{H}}_l$.

Finally, we surmise that the effectiveness of the linear model reveals that the proposed aggregation is sensitive to the dying ReLU problem, which may negate the contribution

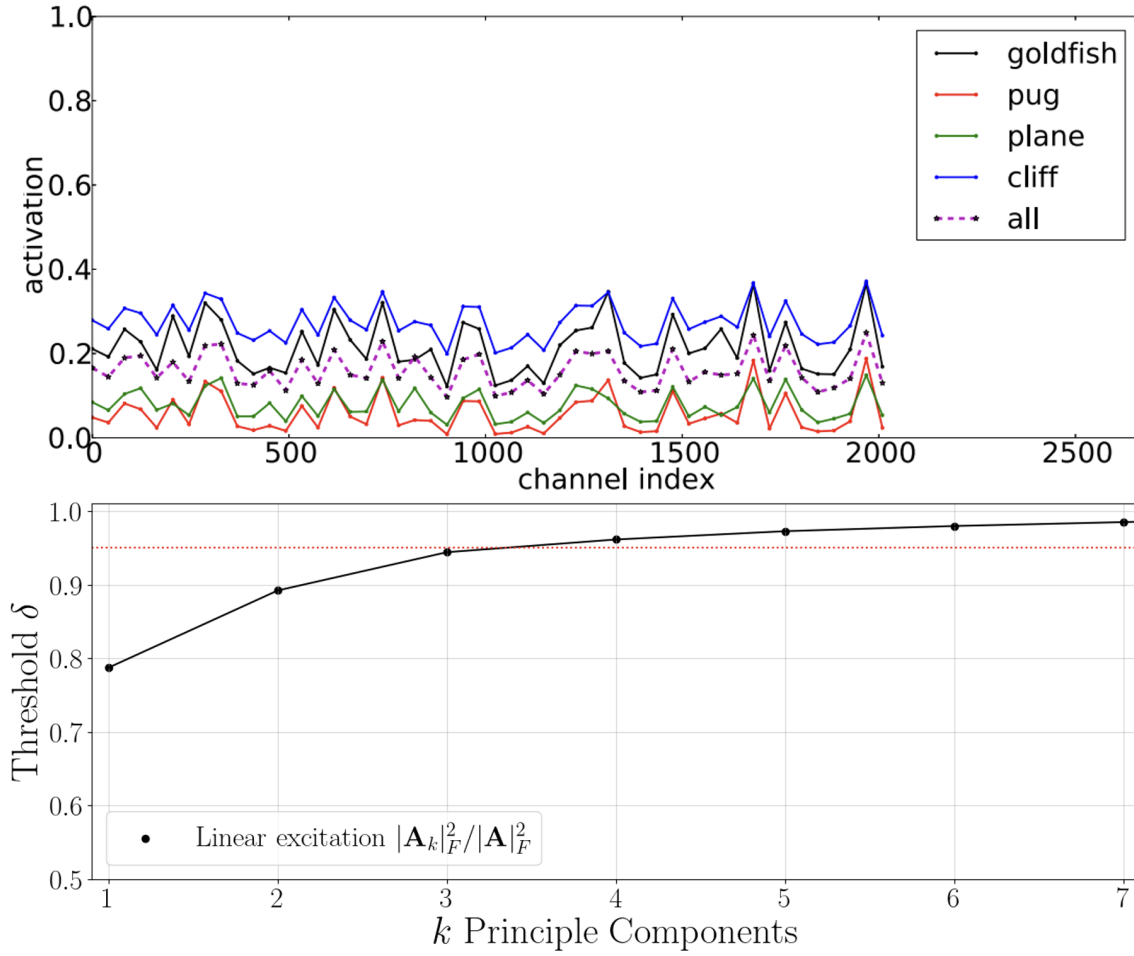


FIGURE 3.9: (Top) Excitation analysis of excitation magnitude on a ImageNet-1000 classification task, adapted from [92]. (Bottom) The number of principal components that capture $\delta = 0.95$ of the excitation variance on a binary segmentation task on the ISIC-18 dataset for the first encoder head of a U-Net deprived of its skip-connections.

or identification of informative level sets due to initialization of either equation 3.26 or ϕ [145]. Note that also that a general singularity spectrum is well defined for the entirety of the real numbers[119]. Architectural tweaks in this module should be further explored in future work.

3.9.4 Analysis of Global Channel Attention Functions

The authors of the original SE paper [92] argued that, for the task of image classification, the excitations tended to be more class-specific as $l \rightarrow L$, becoming (at least) visually separable before the output layer (see Figure 3.9). Since for segmentation each input location maps to each output, and this relationship becomes impossible to estimate given the lack of spatial resolution of the excitations g .

We use the linear estimation of the the covariance of the excitations $h \circ g_l$. We denote the covariance matrix as $A \in \mathbb{R}^{C_l \times C_l}$ such that

$$A := \frac{1}{n-1} (h \circ g_l)^\top (h \circ g_l), \quad (3.29)$$

where n is the size of a validation set.

Note that this aligns with recent trends in deep learning theory that state that, under mild assumptions, deep representations tend to display linear behavior after training [146, 147].

Since we deal only with binary tasks, we use how much information A_k , the singular value decomposition of A using only its top k singular values, captures about the variance of the neurons (see Fig. 3.9). We define the *linear excitation threshold*, $\delta \in (0, 1]$ as:

$$\operatorname{argmin}_k \frac{|A_k|_F^2}{|A|_F^2} \geq \delta. \quad (3.30)$$

In simple terms, for a fixed δ , if the excitations are linearly related to the target variable, then one expects the minimum $k \rightarrow 2$.

Analysis of this surrogate task with $\delta = 0.95$, for each model and dataset pair, yielded three main insights:

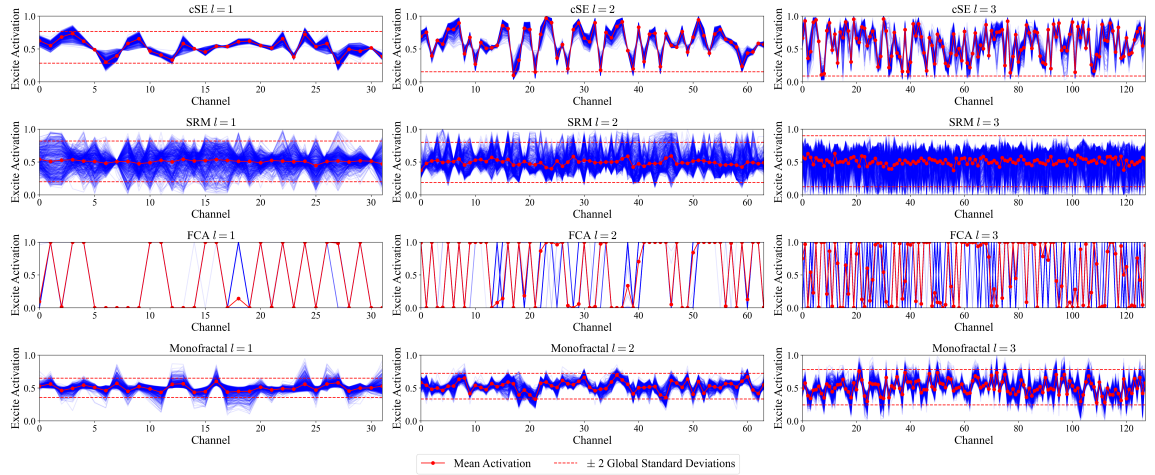


FIGURE 3.10: Test set excitation responses of cSE, SRM, FCA, and Monofractal recalibration for $l \in \{1, 2, 3\}$ in the first fold of the ISIC dataset. Each blue segment marks a response of the layer given an instance. SRM displays very heterogeneous behavior per instance. On the opposite end of the spectrum, FCA is almost instance-agnostic. cSE and Monofractal recalibration display similar instance variability, but markedly different excitation responses around the average.

(i) *There is no strong evidence of filtering of non-informative channels:* a common intuition behind channel attention functions is that they somehow learn to select only the most

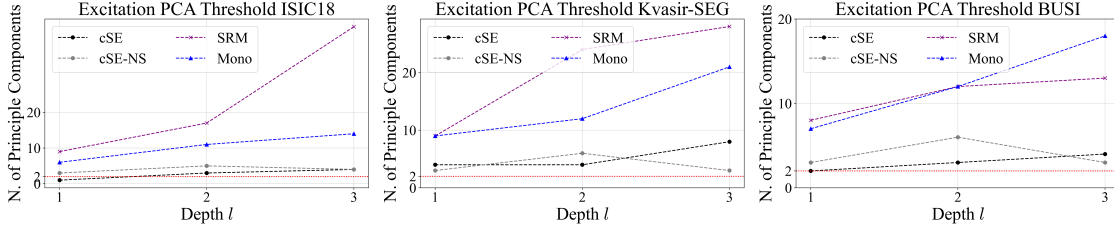


FIGURE 3.11: Number of principle components required to explain 95% of the variance for each model for every encoder depth l for each dataset. cSE-NS stands for cSE in a U-Net without skip-connections. FCA is not included due to its almost heterogeneous behavior. The U-Shape architecture makes it so that become less class-dependent as $l \rightarrow L$.

TABLE 3.4: Compute requirements. Inference and train times averaged over 1000 steps using a NVIDIA GeForce RTX 3090 GPU.

	U-Net[144]	+cSE [92]	+scSE [93]	+SRM [94]	+FCA [95]	+Mono	+Multi
Total parameters	1,946,881	1,968,721	1,968,948	1,947,783	3,022,417	1,989,105	1,967,553
Trainable parameters	1,946,881	1,968,721	1,968,948	1,947,335	1,968,721	1,969,169	1,947,521
Non-trainable parameters	0	0	0	448	1,053,696	19,936	20,032
Train time (ms/step)	60	73	104	77	76	165	450
Inference time (ms/step)	40	45	64	50	69	115	220

informative channels [93, 110, 117]. We were not able to measure such an effect in our experiments, at least for the cases when the attention modules actually increase baseline performance. For example, Monofractal recalibration displays the highest performance for the ISIC18 dataset (see Table 3.1), although most of the channels are not weighted aggressively by its excitations (see Fig. 3.10).

(ii) *The covariance of g depends monotonically on l due to its skip-connections:* we tested measuring the excitation threshold by removing the skip-connections of the U-Net (see Fig. 3.11, cSE-NS) and found that this would break the monotonic quasi-linear relationship between l and k . A similar behavior was identified by the original authors of the SE [92] for classification. Since $C_0 \ll C_l$, this also suggests that the linear manifold where the approximation of $\text{Cov}(g)$ resides increases with l , for a given δ . This is inline with current intuitions about how more ‘complicated features’ are learned in the later stages of a CNN [148, 149].

(iii) *Balanced instance variability appears to correlate with performance:* extreme instance variability consistently led to subpar performance, as it can be observed for SRM and FCA in 3.10. In contrast, both cSE and Monofractal recalibration appear to be predictable, although flexible instance variability.

3.9.5 Discussion and Conclusions

We introduce two empirical priors based on Fractal Geometry: Monofractal and Multifractal recalibration. We construct them in a principled way and integrate them successfully in practice. The proposed channel attention functions significantly enhanced a U-Net’s segmentation performance, while simultaneously outperforming other statistical informed channel attention functions [93–95]. Insofar as we are aware, this was the first time that Multifractal analysis was leveraged completely end-to-end for semantic segmentation.

The experiments were conducted on three public medical datasets suggest that although cSE, scSE [93] and Multifractal improve the baseline for all datasets, only Multifractal recalibration does so in a statistically significant manner for all cases.

Our analysis includes other statistical priors that extend traditional SE by including higher order information in some capacity, specifically the SRM [94] and FCA [95]. Surprisingly, they generally unperformed in our experiments. In terms of global channel statistical descriptors, with the exception of the Kvasir-SEG dataset, Monofractal recalibration offers superior performance.

Our empirical analysis also reveals that global channel attention function effectiveness is potential linked with instance variability, not to the degree to which these attention models “filter out” some channels [92, 110, 113, 150]. On the extremes, we observed instances where instance variability was excessive (SRM) or close to 0 (FCA), and in both cases, performance decreased on average. On the other hand, both cSE and Monofractal recalibration offer balanced instance variability but have markedly different average gating responses (refer to Figure 3.8). We also demonstrated through experiment that, for the U-Net architecture, the excitation responses seem to get more chaotic with the depth of the encoder, which can be likely attributed to an effect of its skip connections.

3.10 Chapter conclusion

In this chapter we set forth novel inductive priors aimed at describing highly irregular objects through their innate geometrical properties. Specifically, we geared our approach towards scale-free, self-seeding processes, such as those occurring in several medical imaging modalities. Two new attention functions describing cross-channel dynamics by deriving single or multiple statistics over the local scaling exponents of each filter are proposed

(see Figure 3.4). We show substantial improvements over several attention functions [93–95] in terms of segmentation performance in three public medical image segmentation datasets.

In comparison to previous end-to-end (multi)fractal approaches for the context of classification tasks [66, 107, 109], our methods are more theoretically grounded, but also more parameter efficient, and thus suitable for semantic segmentation. Notwithstanding, not all limitations were overcome. For instance, in order to keep computation time reasonable, we had to resort to a simple OLS estimate using a fixed small number of scales. Our recalibration modules require less compute than [66, 107, 109], yet they still take a considerable amount of extra computational time when compared to the other recalibration strategies studied in this chapter (see Table 3.4). On the other hand, our methods only require a diminutive set of parameters, barely increasing the baseline model size.

Due to the fact that we express these inductive priors as channel attention function, there is an implicit assumption that each filter will define a self-similar measure in order for the OLS step carried out to estimate a quantity that relates to its scaling exponents. We believe these issues may be solved in future work by further exploring the connections of CNNs with Wavelet Scattering [151] which potentially will allow to leverage more sophisticated multifractal formalisms [83–85].

3.10.1 Outcomes

- 4 publications in international peer-reviewed conferences [70, 71, 152, 153].
- 1 submission to a Q1 Journal [154].
- 1 unindexed national proceeding [155].

Chapter 4

Markov-based Neural Networks

Hitherto, we have focused on studying statistical regularities of random variables (images) whose joint distribution is permutation-invariant. However, it is possible that the random variables depend on some index t , for instance, time. In this Chapter we will precisely study regularities for this type of *sequential* data, with an emphasis on Markov models.

We will start by introducing the fundamental concepts of Markov and hidden Markov models, followed by our algorithmic contribution, Markov-based Neural Networks. It concludes with an extensive experimental section on heart sound segmentation, where we demonstrate the framework's potential when there is a strong prior match with the dataset.

We focus solely on the concepts of Markov models relevant to this chapter. We recommend [156] for an extensive treatise. Chapter 13 of [157] and even the seminal tutorial by Rabiner et al. [158] are also recommended if one is looking for material of a more accessible nature.

4.1 Markov and hidden Markov models

Markov models are a crucial statistical tool for modelling sequential data. Assume you have some set of data points $x \sim p$. The first assumption we make is that the joint distribution is indexed by some discrete t that induces some total or partial ordering on the data so that $x_{t-1} \preceq x_t$. Thus, for some $T \in \mathbb{N}^+$, the joint distribution may be written as

$$p(x) = \prod_{t=0}^{T-1} p(x_t | x_1, \dots, x_{t-1}). \quad (4.1)$$

If one imposes $x_t \perp x_{t-k}, \forall t, k \in \mathbb{N}^+$, then we have a k -th order Markov Chain:

$$p(x_t | x_1, \dots, x_{t-1}) = p(x_t | x_{t-k}, \dots, x_{t-1}) \quad (4.2)$$

Suppose that p is multinomial with L parameters and that the conditional distributions are stationary. The order of the Markov chain makes it so that each conditional needs to have $L - 1$ parameters for each of the categories or *states*, and this has to be unrolled as a function of the order so, for each conditional, we need $L(L - 1)^k$ [157].

The fact that x s depend on one another, and that imposing higher orders leads to an exponential growth of parameters, limits the practical expressability of distributions characterised by these Markovian assumptions. Fortunately, this can be mitigated by the more general family of *hidden Markov models*.

Definition 4.1 (Hidden Markov Model). Suppose that the observations x depend on latent factor z so that

1. The latent factor forms a Markov chain $z_{t+1} \perp z_{t-1} | z_t, \forall t \in \mathbb{N}^+$
2. The conditional is expressed as $p(x_t | x, z) = p(x_t | z_t)$ meaning that the observations at index t solely depend on z_t .

Then x, z are jointly distributed according to a (first-order) hidden Markov model.

Proposition 4.2. All of the Markov k -th order chains on some $\mathcal{M}$ can be expressed as a first order hidden Markov model (HMM) such that $\mathcal{Z} = \mathcal{M}^k$. The converse is not true.

Proof. Suppose that we have a k -th order Markov chain on $\mathcal{M}$. Define the set of states of the HMM as $\mathcal{Z} = \mathcal{M}^k$, and model all possible state transitions in Γ so that $p(z_t | x_1, \dots, x_t) = \delta(x_{t-k}, \dots, x_t)$, i.e., the transitions are chosen deterministically as a function of the last k observations, and so the emission distribution is exactly the same for both models.

The converse is not true. The Markov chain is a deterministic function of the last k symbols. Suppose $\mathcal{Z} = \{0, 1\}$, and you observe the following sequence of length $k + 1$: $00 \dots 01$. For a k -order Markov chain, observing 1 has to be a deterministic function of the first k zeroes. The HMM allows for two possible events at symbol $k + 1$ given the first k symbols of this sequence: either 0 with probability q or 1 with probability $1 - q$. $\square$

Remark 4.3. Note that the above proof serves the purpose of showing that we can model arbitrary long dependencies with a simple first-order HMM. That said, the implication step did not resolve the previously raised parameter issue. The main takeaway is that

observations are correlated on arbitrary length using a very diminutive set of parameters. Note that for the above example, the symbols are coupled as an exponential decay function q^t .

For the purposes of this chapter, the special case of first-order Markov chains is particularly important. Consider for instance, the *complete likelihood* of a first-order hidden Markov chain:

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) \xrightarrow{\text{Def. 4.1}} p(z_0) \prod_{t=1}^{T-1} p(\mathbf{x}_t|z_t) \prod_{t=1}^{T-1} p(z_t|z_{t-1}) \quad (4.3)$$

The likelihood of a given sequence given some underlying Markov chain thus follow by marginalizing over all possible sequences of length T for a given set of states $\mathcal{Z}$:

$$p(\mathbf{x}) = \sum_{\mathbf{z} \in \mathcal{Z}^T} \left(p(z_0) \prod_{t=1}^{T-1} p(\mathbf{x}_t|z_t) \prod_{t=1}^{T-1} p(z_t|z_{t-1}) \right) \quad (4.4)$$

4.2 Markov-Based Neural Networks

In the remainder of this chapter, it is important that we make explicit the set of parameters modeling each of the distributions.

Definition 4.4 (Parametrization of discrete Markov Chain). Suppose we have a countable state set $\mathcal{Z}$ and let $L = |\mathcal{Z}|$. To characterize $p(\mathbf{z})$ we need to define the initial state distribution $\boldsymbol{\pi} = (p(z) : z \in \mathcal{Z})$ such that $\boldsymbol{\pi} \in \mathbb{R}^L$, and a state transition matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{L \times L}$. Let $\lambda = \{\boldsymbol{\pi}, \boldsymbol{\Gamma}\}$ denote the parameters of this chain, then the likelihood of a given sequence $\mathbf{z}' = [z'_0, \dots, z'_{T-1}]$ is denoted by p_λ

$$p_\lambda(\mathbf{z}') = p_{\boldsymbol{\pi}}(z'_0) \prod_{t=1}^{T-1} p_{\boldsymbol{\Gamma}}(z'_t|z'_{t-1}).$$

So far, we mainly articulated how the distribution of $\mathbf{z}$ can be characterised using λ . However, we do not know *a priori* the class of distributions to which the so-called *emissions* $p(\mathbf{x}_t|z_t)$ * pertain. It is common to model this conditionally using generative models such as Gaussian Mixture Models [158].

Highly discriminant models such as neural networks only allow estimation of $p(z_t|\mathbf{x})$, but this limitation can be easily overcome by leveraging Bayes' theorem as such.

*Note that $\mathbf{x}_t$ is deliberately boldfaced, indicating that it can be a multi-dimensional vector.

Definition 4.5 (Markov-based neural network). Suppose you have a Markov chain $\lambda = \{\pi, \Gamma\}$ and a DNN parametrized by Θ modeling $p(z_t|\mathbf{x}_t)$. Let $\Psi = \{\lambda, \Theta\}$, then:

$$p_{\Psi}(\mathbf{x}_t|z_t) = \frac{p_{\Theta}(z_t|\mathbf{x}_t)p_{\Psi}(\mathbf{x}_t)}{p_{\lambda}(z_t)}.$$

The parameter set Ψ describes a Markov-based neural network.

When the DNN is discriminative one can only model $p_{\Theta}(z_t|\mathbf{x}_t)$, then one disregards $p_{\Psi}(\mathbf{x}_t)$ in the above formulation. This means one optimises a pseudo-likelihood of Ψ , a quantity that is proportional, but not equal to, the complete likelihood.

Definition 4.6 (Hybrid training). Let Ψ be the set of parameters of a Markov-based neural network (MNN), then suppose that the DNN model is discriminative. If one searches for Ψ that maximizes the following pseudo-likelihood:

$$\hat{p}_{\Psi}(\mathbf{x}, \mathbf{z}) = p_{\lambda}(\mathbf{z}) \prod_{t=0}^{T-1} \frac{p_{\Theta}(z_t|\mathbf{x}_t)}{p_{\pi}(z_t)},$$

then the model is said to be trained in a hybrid fashion.

Due to the tractability of our model, one can produce closed expressions for the observation likelihood $p_{\Psi}(\mathbf{x})$. However, if the DNN is not generative, then this quantity will also only be computed to a value that is proportional to the true likelihood, $\hat{p}_{\Psi}(\mathbf{x})$. In order to simplify notation, in the remainder of this chapter, we will use p and $\hat{p}$ interchangeably, since we will be focusing on hybrid MNN as postulated in 4.6.

4.2.1 Training

Suppose you have a dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{z}^{(i)})\}_{i=1}^N$, where each observation has length T^i . In this work, we focus solely on stochastic gradient descent (SGD) optimisation with the three distinct loss functions. The first is the direct optimization over the (pseudo) complete likelihood:

$$\mathcal{L}_{\text{CL}} = - \sum_{i=1}^N \log p_{\Psi}(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}). \quad (4.5)$$

A fine-tuning loss that leverages the tractability of the MNN:

$$\mathcal{L}_{\text{FT}} = - \sum_{i=1}^N \log p_{\Psi}(\mathbf{x}^{(i)}). \quad (4.6)$$

Finally, following Fritz et al. [159], the maximum mutual information (MMI) criterion loss:

$$\mathcal{L}_{\text{MMI}} = - \sum_{i=1}^N \log p_{\Psi}(\mathbf{x}^{(i)}, \mathbf{z}^{(i)}) - p_{\Psi}(\mathbf{x}^{(i)}). \quad (4.7)$$

One should make a distinction between equation 4.5 and equation 4.7. The latter is clearly equivalent to $\log p_{\psi}(\mathbf{z}|\mathbf{x})$, as intended to enhance the generative capabilities of the MNN. However, although it can be shown that this loss does not depend on the explicit specification of $p_{\Psi}(\mathbf{x}|\mathbf{z})$, this also means that the optimisation procedure is blind to the true emission model. When the model is hybrid, neither equation 4.5 nor equation 4.7 is equivalent to a fully specified model that allows for computation of the precise complete likelihood, and their performance depends more on the nature of the downstream task.

4.2.2 Parameter estimation

We will exploit the closed expression for maximum likelihood quantities of the parameters of the underlying chain λ . Specifically, we will focus on a family of HMMs that describe most stochastic processes that govern real-world phenomena, those that have a *steady state distribution*.

Definition 4.7 (Aperiodic irreducible Markov Chain). A Markov matrix Γ is a row-stochastic matrix, that is, $\sum_j \Gamma_{ij} = 1, \forall i$. An *irreducible* and *aperiodic* Markov chain is characterized by Γ so that:

1. Γ is irreducible if it is possible to visit all states, i.e., $\exists t \forall t' \geq t, \forall i, j : \Gamma_{ij}^{t'} > 0$.
2. Γ is aperiodic if one can not transition into the same state, i.e., $\Gamma_{ii} > 0, \forall i$.

Examples of different types of Markov chains are illustrated in Figure 4.1.

Proposition 4.8 (Stationary steady state distribution). *Any aperiodic and irreducible Markov chain converges to a state distribution π^* , regardless of the initial π .*

Proof. Since Γ is irreducible, then Γ has full rank so and a unique right eigenvector. Γ is also aperiodic, so Γ^t converges (see Theorem 4.9 of [160]).

Let there be a finite t , so that, for any $t' \geq t$, it holds that:

$$\pi^* = \pi \Gamma^{t'}$$

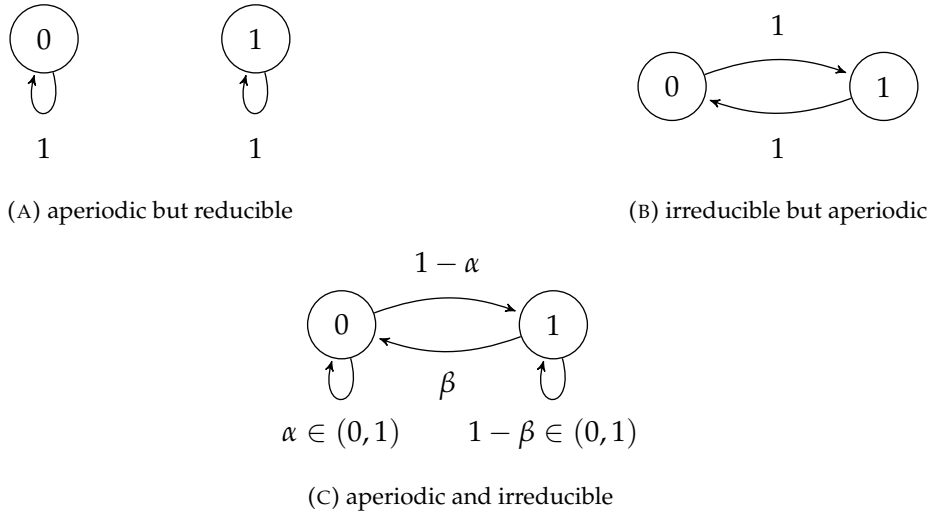


FIGURE 4.1: Examples of reducible/irreducible and periodic/apperiodic Markov chains.

for any π . Then, for $t + 1$:

$$\pi^* \Gamma = \pi \Gamma^t \Gamma \xrightarrow{\text{Def.}} \pi^* = \pi^* \Gamma.$$

Since Γ is aperiodic, its *steady state distribution* is stationary. $\square$

Remark 4.9. For aperiodic and irreducible Γ one can deduce the stationary distribution π^* by imposing the constraint $\pi \mathbf{1} = 1$, that is:

$$\pi = \pi \Gamma, \text{ s.t. } \pi \mathbf{1} = 1 \implies (\Gamma^\top - I) \pi^\top = \mathbf{0}, \text{ s.t. } \mathbf{1}^\top \pi^\top = 1$$

One can solve the following system in the familiar algebraic form, $Ax = b$ [161]:

$$\underbrace{\begin{bmatrix} \Gamma^\top - I \\ \mathbf{1}^\top \end{bmatrix}}_A \underbrace{\pi^\top}_x = \underbrace{\begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}}_b,$$

which has exactly one solution since $\Gamma^\top - I$ is invertible.

The parameters λ of the underlying Markov chain are initialised with their maximum likelihood estimates (MLE). Given access to a labeled train set, the MLE estimate of the transition matrix, Γ_{MLE} , can be attained directly from $\{\mathbf{z}^{(i)}\}_{i=1}^N$ by calculating the expected number of transitions between states [158].

Concerning the initial state distribution, we approximate it with the steady state probability vector $\pi^* \in \mathbb{R}^L$, by solving the linear system proposed in Remark 4.9 with $\Gamma =$

Γ_{MLE} . Henceforth, assume:

$$\lambda_{\text{MLE}} = \{\pi^*, \Gamma_{\text{MLE}}\} \quad (4.8)$$

The MNN's parametrisation Ψ is searched within a constrained optimisation scheme. Specifically, since λ has probabilistic parameters, the feasible set for π and the rows $\Gamma_i \in \mathbb{R}^L$ of Γ must be a subset of the canonical simplex $\mathbb{K}^L$ of $\mathbb{R}^L$:

$$\mathbb{K}^L = \left\{ x \in \mathbb{R}^L : x_i \geq 0, i = 0, \dots, L-1, \sum_{i=0}^{L-1} x_i = 1 \right\}. \quad (4.9)$$

Let $\|\cdot\|_2$ denote the ℓ_2 -norm. Following [162], we define the projective map $\Phi : \mathbb{R}^L \rightarrow \mathbb{K}^L$ so that:

$$\Phi(y) = \arg \min_{x \in \mathbb{K}^L} \left(\frac{1}{2} \|x - y\|_2^2 \right). \quad (4.10)$$

We are concerned with irreducible and aperiodic Markov chains, which means that we our feasible region is actually a subset of $\mathbb{K}$. More formally:

Definition 4.10. Let $\mathcal{U}_i^L$ be the point set defined as:

$$\mathcal{U}_i^L = \left\{ x \in \mathbb{R}^L \mid \sum_{i=1}^L x_i = 1, 0 < x_i < 1, i = 1, \dots, L \right\}, \quad (4.11)$$

for some $i \in \mathcal{Z}$. Then, let:

$$\mathbb{U}^L := \bigcup_{i \in \mathcal{Z}} \mathcal{U}_i^L.$$

Clearly, $\mathbb{U}^L \subseteq \mathbb{K}^L$. An irreducible and aperiodic Markov chain has transition matrix $\Gamma \in \mathbb{R}^{L \times L}$ so that $\Gamma = [\Gamma_0^T, \Gamma_1^T, \dots, \Gamma_{L-1}^T]^T$, where $\Gamma_i \in \mathbb{U}_i^L$, for all $i \in \mathcal{Z}$.

For each gradient descent update we set $\Gamma_i = \Phi(\Gamma_i + \eta \nabla_{\Gamma_i} \mathcal{L})$, given some learning rate η . In this way, we guarantee that the transition matrix stays in the feasible set $\mathbb{K}^L$ throughout training. Moreover, in each epoch, we derive π directly by solving the steady state linear system immediately after this projection step. The pseudo-code of the framework for gradient descent in an MNN is available in Algorithm 1, given dataset $\mathcal{D}$, some loss function $\mathcal{L}$, learning rate η , and maximum number of epochs m .

In practice, line 10 of Algorithm 1 might yield matrices that are reducible or periodic, which would violate the assumptions stated in Definition 4.8. By selecting a small ϵ , we guarantee that the maximum/minimum value of each row $\Phi(\Gamma_i)$ is $1 - \epsilon/\epsilon$.

The main advantage of an MNN is that it intrinsically models the (Markovian) character of the temporal regime through λ . However, the training procedure outputs a fixed parametrisation Ψ for all $x \in \mathcal{D}$. The fact that the underlying model is coupled to a strong

Algorithm 1: Supervised hybrid training of a generative MNN

Input: $\mathcal{D}, \mathcal{L}, \eta, m$
Output: Ψ

```

1  $\lambda \leftarrow \lambda_{\text{MLE}};$ 
2  $\Theta \leftarrow \text{GlorotUniform}();$  // As in [163]
3  $\Psi \leftarrow \{\lambda, \Theta\};$ 
4 for  $epoch = 1$  to  $m$  do
5    $D \leftarrow \text{shuffle}(\mathcal{D});$ 
6   for  $(x^{(i)}, z^{(i)}) \in D$  do
7      $\Theta \leftarrow \Theta + \eta \nabla_{\Theta} \mathcal{L}(x^{(i)}, z^{(i)}; \Psi);$ 
8      $\Gamma \leftarrow \Gamma + \eta \nabla_{\Gamma} \mathcal{L}(x^{(i)}, z^{(i)}; \Psi);$ 
9     for  $\Gamma_i \in \Gamma$  do
10       $\Gamma_i \leftarrow \Phi(\Gamma_i);$  // equation 4.10
11    end
12     $\pi$  s.t.  $\pi = \pi \Gamma, \pi \geq 0, \|\pi\|_1 = 1;$  // Remark 4.9
13     $\Psi \leftarrow \{\{\pi, \Gamma\}, \Theta\};$ 
14  end
15 end

```

statistical prior also makes it sensitive to observations sampled from a different temporal regime from that of its training set. Herein, we propose a routine that enables the model to adapt to unseen or unusual observations using an unsupervised loss function. Thus, we adapt Algorithm 1 so that it leverages equation 4.6 for the specific purpose of increasing the likelihood of some target observation x given baseline parametrization Ψ . Note that we fine-tune one model per observation and restart the procedure with baseline Ψ given a different datum. The unsupervised fine-tuning routine is described in Algorithm 2.

Algorithm 2: Fine-tuning pre-trained MNN Ψ on x

Input: Ψ, x, η, m
Output: Ψ_m

```

1  $\Psi_0 \leftarrow \Psi;$ 
2 for  $k = 1$  to  $m$  do
3    $\Theta \leftarrow \Theta + \eta \nabla_{\Theta} \mathcal{L}_{\text{FT}}(x; \Psi_{k-1});$ 
4    $\Gamma \leftarrow \Gamma + \eta \nabla_{\Gamma} \mathcal{L}_{\text{FT}}(x; \Psi_{k-1});$ 
5   Lines 9-12 of Algorithm 1;
6    $\Psi_k \leftarrow \{\{\pi, \Gamma\}, \Theta\};$ 
7 end

```

Hereafter, let Ψ_k be the parametrisation after k rounds of fine-tuning on the same observation x given baseline Ψ .

4.3 The Learning Triplet for Markov-Based Neural Networks

Recall the configuration framework $\mathfrak{C} = (\Phi, \mathcal{F}, \mathcal{L})$ defined in Chapter ?? . We can now precisely map the Markov-based Neural Network (MNN) approach to this tuple. This mapping highlights the specific hybrid nature of our inductive bias, which combines discriminative deep learning with generative statistical modelling.

- **Data Processing (Φ):** The input transformation Φ maps raw sequential data into a time-ordered sequence of feature vectors $\mathbf{x} = (x_1, \dots, x_T)$. Unlike standard i.i.d. processing, Φ here must preserve the temporal ordering t to satisfy the Markovian assumptions of the downstream model. The effectiveness of the MNN relies on Φ exposing features where the conditional independence assumption $p(x_t|z_t)$ holds approximately true.
- **Hypothesis Space ($\mathcal{F}$):** The hypothesis space is the set of all MNNs parametrised by $\Psi = \{\lambda, \Theta\}$. This space is characterised by:
 1. **Structural constraints:** We enforce a probabilistic graphical model structure where latent states evolve according to a first-order Markov process (the generative component λ), while observations are processed by a non-linear deep neural network (the discriminative component Θ).
 2. **Parametric Constraints:** The search space for the transition parameters $\lambda = \{\pi, \Gamma\}$ is constrained to the canonical simplex $\mathbb{K}^L$. This is rigorously enforced during optimisation via the projection operator (Eq. 4.10). The neural parameters Θ are unconstrained in $\mathbb{R}^d$ but regularised by the architecture of the DNN.
- **Loss Function ($\mathcal{L}$):** The choice of $\mathcal{L}$ defines the training regime and the specific properties of the posterior we wish to sharpen.
 - In the **supervised phase** (Algorithm 1), we employ the pseudo-complete likelihood $\mathcal{L}_{\text{CL}}$ or the Maximum Mutual Information loss $\mathcal{L}_{\text{MMI}}$. This aligns the model with the ground-truth state sequence provided by the training data.
 - In the **adaptation phase** (Algorithm 2), we switch to the fine-tuning loss $\mathcal{L}_{\text{FT}}$, which corresponds to the negative marginal likelihood. This allows the hypothesis Ψ to adapt to the specific statistics of a single test observation $\mathbf{x}$ without access to labels $\mathbf{z}$, effectively performing test-time unsupervised learning.

By explicitly defining this triplet, we see that the MNN is determined by the configurations where the loss function dictates the interaction between a Markovian generative hypothesis and a deep discriminative feature extractor.

4.4 Markov models in the deep learning era

Works that combine a Markovian prior with gradient descent date back to the 1990's [164, 165]. Noteworthy are the more contemporary Graph Markov Neural Networks [166] proposed. Of special interest to us is their success found in automatic speech recognition [167], especially the work of Fritz et al. [159].

4.5 Application: heart sound segmentation

Cardiovascular diseases (CVDs) are the number one cause of mortality worldwide, with the total number of deaths projected to be over 23 million by 2030, given the current prevalence of CVD risk factors, such as smoking, obesity, and physical inactivity, as reported by the World Heart Federation [168]. CVDs are most prevalent in low to middle-income countries, where limited expert human resources and operational conditions constrain the effectiveness of healthcare services. To mitigate the economic and humanitarian toll of CVDs (especially in underprivileged scenarios), early discovery and treatment are only actionable if the screening becomes more reliable, inexpensive, and fast. In this context, cardiac auscultation is particularly attractive since this examination has low cost, can detect several heart conditions, and is one of the simplest cardiac screening procedures [169]. Although accurate interpretation of the sounds collected during auscultation requires access to personnel with extensive qualifications and clinical experience, collecting the necessary data in the form of a phonocardiogram (PCG) for retrospective examination and diagnosis is considerably simpler, thus requiring less training [170–172]. Moreover, the advent of electronic stethoscopes, coupled with significant progress in machine learning, has brought about a revolution in cardiac auscultation, since computer-assisted decision systems are now capable of extracting meaningful information from PCG recordings with significant clinical relevance [173, 174].

In order to extract information from heart sounds one can first detect key events in the PCG recording, notably the two fundamental sounds present in each heart cycle: the first sound or S1, generated by the mitral and tricuspid valve vibrations from the systolic onset,

and the second sound or S2, which results from the aortic and pulmonary valve closure at the diastolic onset (an example of such fundamental heart sounds is reported in Figure 4.2). The task of delineating the boundaries of the S1 and S2 sounds (which consequently yields knowledge of the systolic and diastolic intervals) for every heart cycle in a PCG is called heart sound segmentation.

One should note that there are many recent solutions for CVD-related downstream tasks that do not require this explicit segmentation step (see, for example, [175–179]), and the community currently questions whether a dedicated segmentation component is required within CVD detection models [180]. Whether the implicit biases in modern neural networks are sufficient remains unclear. However, precise knowledge of the four fundamental segments in each heart cycle not only enables the detection and localisation of extra sound components (e.g., the third and fourth heart sounds, murmurs, ejection clicks, etc.), but it also allows analysis of the waveform morphology of the S1 and S2 sounds. Furthermore, appropriate reconstruction of the underlying heart sound sequence should constrain data-driven models’ responses to physiologically plausible results, enhancing explainability. Naturally, if algorithmic solutions integrating a strong prior should increase robustness and out-of-distribution generalisation.

This should extend to any downstream task on this modality. A good example may be found on the final standings on the George B. Moody PhysioNet challenge 2022 [181], in the task of heart murmur detection – one of the most top-ranked solutions used explicit segmentation of the fundamental heart sounds to enhance classification [182].

In the remainder of this section, we will show how MNNs provide all of these benefits, showing that they outperform other solutions that depend more on their implicit biases using a comparatively small set of parameters.

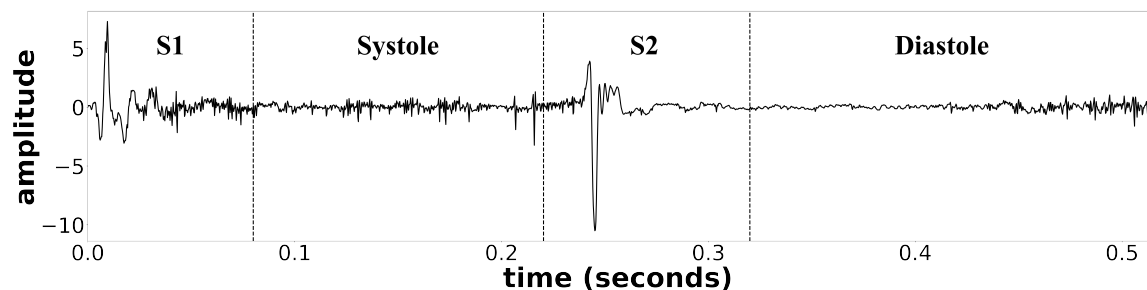


FIGURE 4.2: A segment of 0.51 seconds of a PCG sample from the CirCor DigiScope 2022 dataset displaying an entire heart cycle.

4.5.1 Literature review on heart sound segmentation

Existing literature for heart sound segmentation can be grouped into three main types of techniques: peak-picking, ad-hoc feature extraction coupled with statistical point-wise classifiers, and sequential models. Peak-picking strategies typically apply some transformation to the signal (either in the time or frequency domains) so that thresholding captures local maxima related to S1 and S2 sounds [183–188].

For the second type of strategies, the signal is typically preprocessed to facilitate the discrimination of S1 and S2 components, and hence features are extracted in the time [189], frequency [190], or Wavelet [191] domains. Then, different types of classifiers have been considered to detect the S1 and S2 components, ranging from k -means clustering [192, 193], decision trees [194], Support Vector Machines (SVMs) [191], to deep neural networks (DNNS) [195–197]. However, these methods require a candidate selection step to detect S1 and S2 sounds, which is typically provided by the above-mentioned peak-picking algorithms.

Finally, sequential models exploit a priori knowledge of the steady progression of the heart sounds during each heart cycle, i.e., the fact that the only permissible transitions are: S1 to systole, systole to S2, S2 to diastole, and diastole to S1. Most successful parametric sequential models for heart sound segmentation are grounded in either hidden Markov models (HMMs) or hidden semi-Markov models (HSMMs). The Markovian/semi-Markovian process governs the latent progression of the heart cycle and is typically coupled with a point-wise discriminator that models the signal’s emission distribution. Gamero *et al.* [198] and Gill *et al.* [199] combined HMMs with peak picking on an envelopogram of the PCG. Expanding this line of work, [200] modelled the emission distribution explicitly through Gaussian Mixture Models.

HSMMs enable parametrisation of the so-called sojourn time of each state, i.e., the distribution of the duration of each heart sound in the PCG sequence. Contrary to HMMs where the probability of repeating a state always decays exponentially, in HSMMs the state duration distribution is adaptive and may differ from state to state. HSMMs were first applied by Schmidt *et al.* for heart sound segmentation [201, 202]. Oliveira *et al.* [203, 204] developed methods to learn the sojourn time in the context of heart sound segmentation. In this line of work, there have also been proposals to improve the estimates for the emission distribution, such as the SVM and logistic regression proposed by Springer *et al.* [205, 206].

More contemporary literature points to the supremacy of data-driven methods exploiting DNNs. Renna et al. [207] introduced a sliding U-Net on envelopes extracted from heart sound recordings to estimate the emission distributions, which are then decoded by an HMM parameterised by the maximum likelihood estimates for the train set. In [208], Messner et al. suggested an end-to-end, fully data-driven method for simultaneous learning of the emission and latent state distributions, through a gated bidirectional Recurrent Neural Network (Bi-GRNN) over spectral and envelope features of the signals, which showed competitive results with the HSMM proposed by Springer et al. [206]. Fernando *et al.* [209] proposed a more capable bidirectional Long Short-Term Memory network coupled with an attention mechanism (Bi-LSTM+A) that acted on the *Mel Frequency Cepstral Components* (MFCCs) of the PCG. The attention mechanism can detect salient aspects in each sound, and its application is intended to enhance robustness against noisy or irregular recordings. Additionally, a temporal-framing adaptive (TFA) network was proposed in [210], which employs a specific transition loss function during training and can perform dynamic inference, enabling it to adapt to varying heart sound behaviors. The authors reported positive results when comparing to the HSMM method of Springer et al. [206] and the Bi-GRNN proposed by Messner *et al.* [208].

4.5.2 Implementation

At the beginning of the Chapter, we introduced formalism for a broader family of MNNs. Herein, we will instantiate an MNN explicitly for PCG fundamental heart sound segmentation. First, we establish an adequate feature extraction pipeline that serves the models throughout our experiments. Then we define the characteristics of the parameters in λ alongside their feasible set in the context of fundamental heart sound segmentation. We conclude this section with two distinct, relevant characterisations of MNNs for this task. The implementation may be found on: <https://github.com/miguelmartins/mnn>.

4.5.3 Encoding domain knowledge in the Markov model

Heart sound segments defined by fundamental heart sounds obey three important properties: (a) they occur in a cyclic fashion, (b) their occurrence order is unchanged throughout time, (i.e., S1, systole, S2, diastole), (c) no segment lasts forever. This domain information is expressed organically by a left-to-right HMM as illustrated in Figure 4.3. Hence-

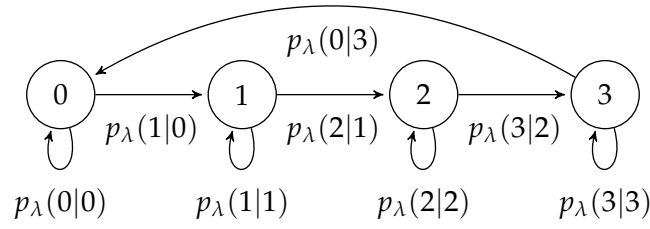


FIGURE 4.3: The left-to-right HMM parameterized by λ for PCG segmentation containing 4 states: S1 (0), systole (1), S2 (2) and diastole (3).

forth, assume that $L = 4$ and that $\mathcal{Z}$ maps the set of possible states: S1, systole, S2, and diastole (as in Figure 4.3).

The feasible set defined in Definition 4.10 can be further simplified to:

$$\begin{aligned} \mathbb{U}_i^L &= \left\{ \mathbf{x} \in \mathbb{R}^L \mid \forall j \in \mathcal{Z} : (j = i + 1 \mod L) \right. \\ &\quad \left. \implies (x_i, x_j > 0, x_i + x_j = 1) \text{ else } x_j = 0 \right\}, \end{aligned}$$

One can visualize this $\mathbb{U}_i^L$ as an edge of the L -dimensional canonical simplex $\mathbb{K}^L$ (see Figure 4.4 for an example in $\mathbb{R}^3$). In practice, computation of Φ in equation 4.10 can be greatly simplified. We are only concerned with solutions in $\mathbb{U}$ and not in $\mathbb{K}$, and Γ_i is always re-projected to the same edge $\mathbb{U}_i$.

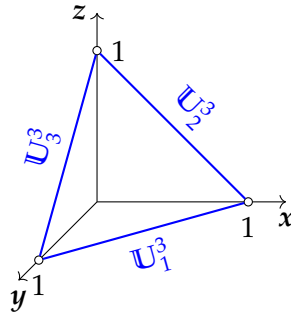


FIGURE 4.4: The feasible set for the columns of $\Gamma \in \mathbb{R}^{3 \times 3}$, where $\mathbb{U}^3 = \{\mathbb{U}_1^3, \mathbb{U}_2^3, \mathbb{U}_3^3\}$. Note how $\mathbb{U}^3 \subset \mathbb{K}^3$. More specifically, it is comprised of the edges of $\mathbb{K}^3$ while excluding its vertices.

We adapted Michelot's finite projection algorithm [162] so that $\Gamma_i \in \mathbb{R}^L \rightarrow \mathbb{U}_i$, by restricting it to the two non-zero components of each row of Γ .

4.5.4 Numerical computation of forward-backwards algorithm

In order to compute the pseudo-likelihood values, one needs to multiply $\mathcal{O}(T)$ valid probabilities, meaning that the value of $p(\mathbf{x})$ requires exponentially more precision.

As prescribed by Rabiner et al. [158], we suppose we compute the likelihood according to the following recursion.

$$p(\mathbf{x}_1, \dots, \mathbf{x}_t, z_t) = \sum_{z_{t-1} \in \mathcal{Z}} p(\mathbf{x}_1, \dots, \mathbf{x}_{t-1}, z_{t-1}) p(z_t | z_{t-1}) p(\mathbf{x}_t | z_t) \quad (4.12)$$

Note that we can compute equation 4.12 iteratively. This allows one to rescale it for each t :

$$\sum_{z_t \in \mathcal{Z}} \frac{p(\mathbf{x}_1, \dots, \mathbf{x}_t, z_t)}{\sum_{z_t \in \mathcal{Z}} p(\mathbf{x}_1, \dots, \mathbf{x}_t, z_t)} = 1, \forall t.$$

Then, let $Z_t = \frac{1}{\sum_{z_t \in \mathcal{Z}} p(\mathbf{x}_1, \dots, \mathbf{x}_t, z_t)}$. Then, by using the marginalisation property of probability mass functions $p(a_1, \dots, a_n) = \sum_b p(a_1, \dots, a_n, b)$, and the fact that Z_t depends solely of the quantity at time t :

$$\prod_{t=1}^T Z_t \sum_{z_T} p(\mathbf{x}_1, \dots, \mathbf{x}_T, z_T) = 1 \implies p(\mathbf{x}_1, \dots, \mathbf{x}_T) = \frac{1}{\prod_{t=1}^T Z_t},$$

and so, the log-likelihood is given by:

$$\log p(\mathbf{x}_1, \dots, \mathbf{x}_T) = - \sum_{t=1}^T \log Z_t. \quad (4.13)$$

4.5.5 Preprocessing

Replicating [207], the signals are first filtered using a Butterworth filter of order two with pass-band [25, 400] Hz and then down-sampled to 1000 Hz. After applying the spike removal algorithm proposed by [202] we extract two distinct feature maps: the envelopes, $\mathbf{X}_{\text{Env}}$, as in Renna et al. [207], and the static MFCCs, $\mathbf{X}_{\text{MFCC}}$, following Fernando et al. [209]. Our models use the same feature-extraction setup described in [206]. Specifically, $\mathbf{X}_{\text{Env}}$ is comprised of 4 channels (one for each feature): (i) the homomorphic envelope, (ii) the Hilbert envelope, (iii) the wavelet envelope, and (iv) the power spectral density envelope. A detailed explanation of this processing step can be found in Appendix C. These features are downsampled to 50 Hz and normalised so that each channel has zero mean and unit variance.

Concerning $\mathbf{X}_{\text{MFCC}}$, following Fernando et al. [209], we extract 6 static Mel-frequency Cepstral Coefficients (MFCC), alongside their first and second order frame differences, Δ and Δ^2 . Each model processes segments of 64 samples, which is equivalent to approximately 1.3 seconds. We assume that the label at the central position of each segment (i.e.,

$64/2 = 32$, or around 0.65 seconds) is the ground truth label for the entire frame. The value of this label is one-hot-encoded.

4.5.6 Inference and post-processing

We developed MNNs for PCG segmentation using a standard Viterbi decoder [158] to output the most likely sequence. Thus, the j -th sequence of the dataset is decoded as follows:

$$\hat{\mathbf{z}}^{(j)} = \max_{z_1^{(j)}, z_2^{(j)}, \dots, z_{T_j}^{(j)}} P_{\Psi}(\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, \dots, \mathbf{x}_{T_j}^{(j)}, z_1^{(j)}, z_2^{(j)}, \dots, z_{T_j}^{(j)}), \quad (4.14)$$

so that $z_1^{(j)}, z_2^{(j)}, \dots, z_{T_j}^{(j)} \in \mathcal{Z}$.

4.6 Experiments

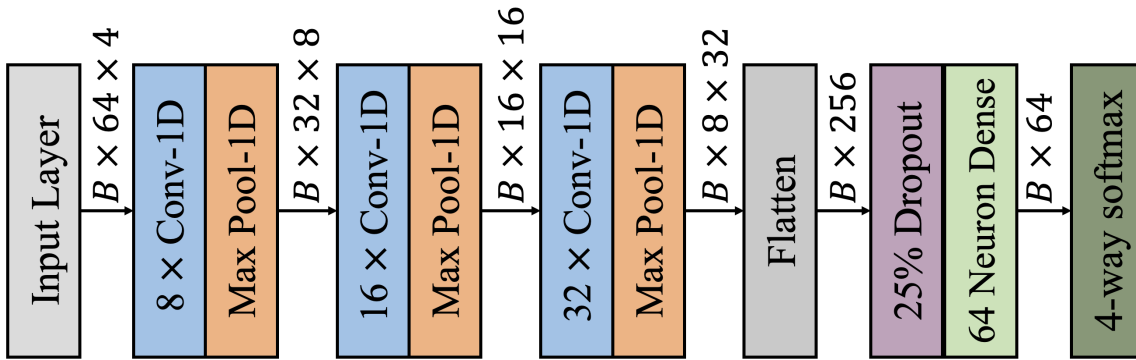


FIGURE 4.5: The template of the one-dimensional CNN backbone for the MNNs implemented in our experiments, assuming a mini-batch size B . The output layer is a 4-way softmax that yields posterior estimates for each PCG state.

We conducted our experiments on the 2016 PhysioNet and 2022 CirCor DigiScope datasets [211, 212]. Onwards, we will refer to each dataset as PhysioNet'16 and CirCor'22, respectively. Both databases contain noisy sounds collected in clinical and non-clinical environments (e.g., during screening campaigns). The raw datasets are fed independently into our preprocessing pipeline, where we extract $\mathbf{X}_{\text{Env}}$ and $\mathbf{X}_{\text{MFCC}}$. We adopt a 10-fold cross-validation setup throughout our experiments, where the folds have a fixed length of roughly 10% of the length of the dataset in its entirety. We set aside 10% of the out-of-fold dataset as a validation set so that we can employ early stopping within each fold. Thus, we will have a train, validation, and test split for each of the ten dataset partitions generated for cross-validation. We also ensure that all folds are mutually exclusive with regard to their patient identifiers. The sounds that have matching patient identifiers to those of

each fold are removed from the respective train and validation sets. Our experimental methodology can be separated into 3 main components: (Section 4.6.3) establishment of an MNN baseline in PhysioNet'16, (Section 4.6.4) performance assessment of the baseline MNN versus a statistical sequential model and two existing data-driven solutions in PhysioNet'16 and CirCor'22, and (Section 4.6.5) cross dataset assessment of the performance of the unsupervised fine-tuning given an MNN pre-trained on either PhysioNet'16 or CirCor'22.

We selected a simple one-dimensional Convolutional Neural Network (CNN) architecture (see Figure 4.5) as the DNN component of our MNN. It is built from a convolutional stem comprising three blocks of one-dimensional convolutions with kernel size 3 and rectified linear unit (ReLU) activation. These are interlaced with 2×1 max-pooling layers. Each block stacks 8, 16, and 32 convolutional filters, respectively. Finally, we employ a 25% dropout bottleneck layer, followed by a fully connected layer with 64 neurons and a ReLU activation. The output layer implements a 4-way softmax. An illustration of our experimental pipeline can be found in Appendix C.1. We compare our results with the well established HSMM approach by Springer et al. [206] and two recent data-driven PCG segmentation solutions: the sliding U-Net followed by Viterbi algorithm proposed by Renna et al. [207], and the Bi-LSTM+A proposed by Fernando et al. [209]. We use the preprocessing configurations associated with the best results reported by each author; hence, we use $\mathbf{X}_{\text{Env}}$ for the HSMM and U-Net, and $\mathbf{X}_{\text{MFCC}}$ for the Bi-LSTM+A.

The MNNs and U-Net were trained with a mini-batch size of 1, and Bi-LSTM+A with a mini-batch size of 32. We used the Adam gradient descent algorithm [143]. The MNNs and U-Net were trained using a learning rate of 0.001. The Bi-LSTM+A was trained with an initial learning rate of 0.002, and 0.0002 from epoch 10 onwards, as proposed by the authors [209]. Both the U-Net and the Bi-LSTM+A were trained to minimize the cross-entropy. Training was performed during 50 epochs throughout each fold and the value of the loss function in the validation set was used as the early-stopping criterion in all experiments.

4.6.1 Materials

The PhysioNet 2016 Dataset (PhysioNet'16) The public dataset used for the 2016 Computing in Cardiology (CinC)/PhysioNet Challenge [211] served as the initial baseline for

our experiments. It is a repository of 9 different heart sound databases collected by different international research groups. It amasses a sizeable number of PCG recordings from aortic, pulmonary, tricuspid, and mitral auscultation sites. The dataset spans 2435 recordings from 1297 healthy or pathological patients, the latter spanning a variety of conditions, including heart valve and coronary artery diseases, aortic stenosis, and mitral regurgitation. It includes a fairly large amount of noisy recordings with real-world acquisition conditions. These recordings were originally re-sampled at 2000 Hz with anti-aliasing [211]. Of the original recordings, we use only the 792 heart sounds (181 healthy, 611 pathological) from 135 patients that have an associated ECG recording*.

The fundamental heart sound sequence was estimated through analysis of the synchronous ECG recordings following [206]. The duration of each recording in this set ranges from 1 to 35.5 seconds. We discarded 39 samples, accounting for the cases where the signal lasted less than 1 second or had noisy labels (i.e., illegal state sequences, such as those that allow transitions from state S1 directly to state S2).

The CirCor DigiScope Dataset (CirCor’22) The CirCor DigiScope Dataset [212] (CirCor’22) is currently the largest publicly available pediatric heart sound dataset, and it was featured in the 2022 George B. Moody PhysioNet Challenge [181]. The focus of the tasks enabled by this dataset was that of cardiac murmur detection and classification. The CirCor’22 dataset was assembled using data collected from two different screening campaigns in Northeast Brazil with a specific focus on the pediatric population [212]. The study included all subjects who volunteered for screening within the study period. Patients younger than 21 years of age with a consent form signed by their parents (when appropriate) were included, with no further exclusion criteria. Two cardiac physiologists independently analysed the collected sounds and discarded PCG recordings based on a signal quality standard. Thus, the database spans 5282 PCG recordings collected from the aortic, pulmonary, tricuspid, and mitral locations from 1568 healthy or pathological subjects (1144 healthy, 305 pathological, 119 indiscernible).

The signals were sampled at 4000 Hz with 16-bit resolution, with durations ranging from 5.3 to 80.4 seconds. The ground truth labels were provided by the same two blinded cardiac physiologists, also tasked with reviewing and correcting the automatic annotation recommendations proposed by the automatic segmentation algorithms of Springer et al.

*The dataset is available at <https://PhysioNet.org/physiotools/hss/>

[206], Oliveira et al. [204], and Renna et al. [207]. Of the 5282 recordings, we used the 3279 samples publicly available in the training set for the 2022 PhysioNet Challenge*.

4.6.2 Metrics and evaluation

We follow the instructions of [202] to compute the confusion matrix in our experiments. Specifically, a prediction is considered a true positive if the centre of an S1 (S2) prediction is closer than 60 ms to the next S1 (S2) sound in the true sequence. All other predicted S1 or S2 segments are considered false positives. After computing the confusion matrix, the sensitivity is computed as:

$$S = \frac{\text{number of TP sounds}}{\text{total number of S1 and S2 sounds}}$$

and positive predictive value (PPV) as:

$$\text{PPV} = \frac{\text{number of TP sounds}}{\text{number of TP sounds} + \text{number of FP sounds}}.$$

These metrics are only tractable if no illegal transitions are present in the model prediction sequence (e.g., skipping systole by allowing the transition $S1 \rightarrow S2$). Thus, our implementation of the Bi-LSTM+A enforces temporal framing of the prediction sequences via Viterbi decoding using the maximum-likelihood estimates from the training set.

4.6.3 MNN baseline benchmark

We start by ablating the joint search of $\Psi = \{\lambda, \Theta\}$ described in Algorithm 1 in the PhysioNet'16 dataset. We define the static formulation of an MNN where the parameters of the underlying HMM stay fixed to $\lambda = \lambda_{\text{MLE}}$ throughout training, and a hybrid MNN where λ is initialized with λ_{MLE} but is jointly-learned with Θ throughout training.

We denote the static model as $\text{MNN}_{\mathcal{L}}^{\Theta}$ and its hybrid counterpart as $\text{MNN}_{\mathcal{L}}^{\Psi}$. At the same time, we evaluate the difference between setting $\mathcal{L} = \mathcal{L}_{\text{MMI}}$ (equation 4.7) and $\mathcal{L} = \mathcal{L}_{\text{CL}}$ (equation 4.5). Inspecting Figure 5.1, concerning $\mathcal{L} = \mathcal{L}_{\text{MMI}}$ in particular, the hybrid model $\text{MNN}_{\mathcal{L}_{\text{MMI}}}^{\Psi}$ displays less variability between folds. Moreover, this variant appears more robust in terms of sensitivity, while its static counterpart measures a higher median PPV. Notwithstanding, our pair-wise t -test ($\alpha = 0.5$) showed no statistical difference in their inter-fold performance metrics (see Figure 4.7).

*The original dataset is available at <https://physionet.org/content/circor-heart-sound/1.0.3/>.

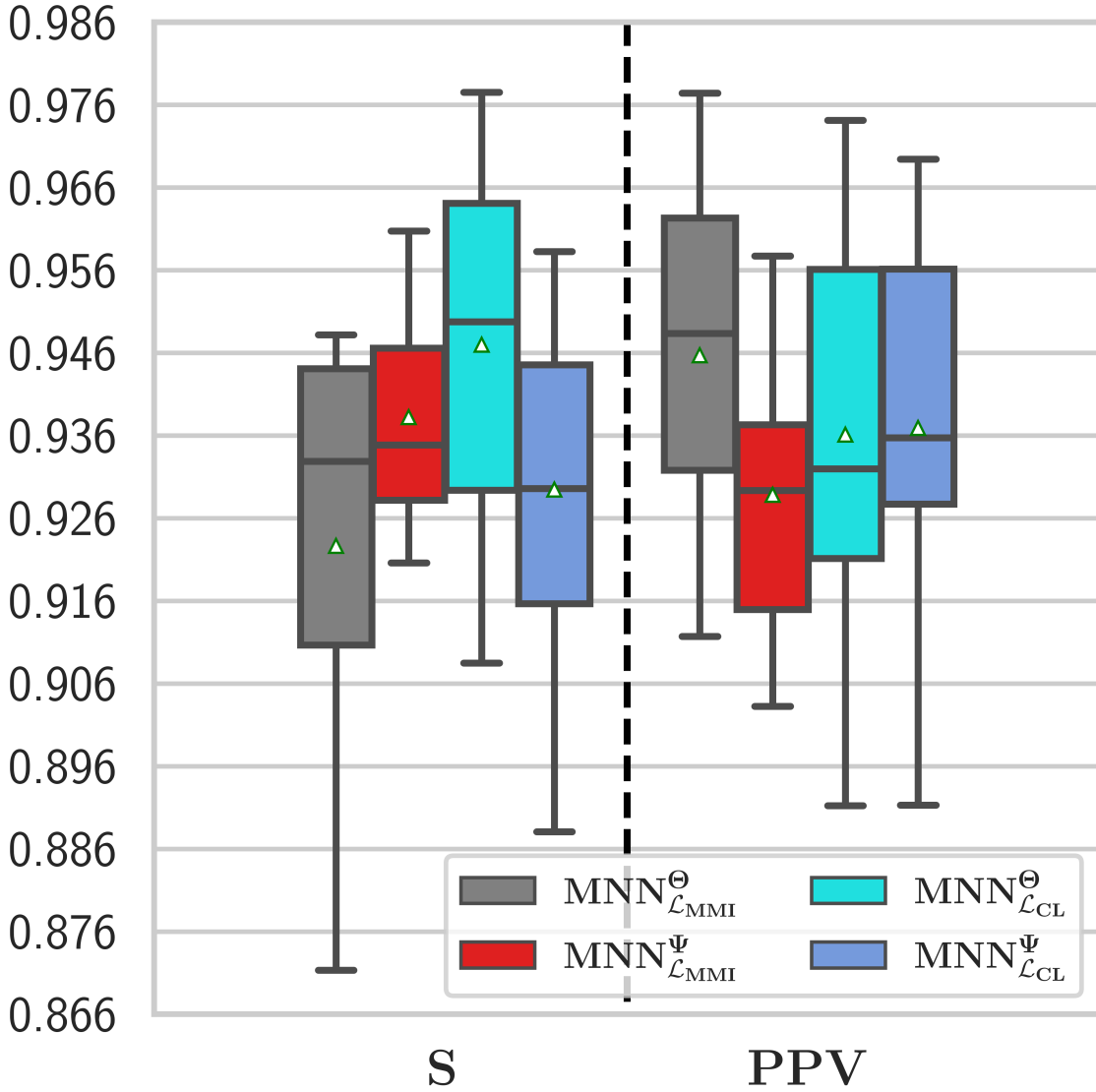


FIGURE 4.6: Measured sensitivity (S) and positive predictive value (PPV) in the 10-fold cross-validation experiment of the PhysioNet'16 dataset for different static/hybrid MNNs, under different loss functions $\mathcal{L} = \mathcal{L}_{MMI}/\mathcal{L} = \mathcal{L}_{CL}$.

When one looks at the results for $\mathcal{L} = \mathcal{L}_{CL}$, the static MNN is considerably more sensitive than its hybrid counterpart while being just marginally inferior in terms of PPV. There was a significant difference between $MNN^{\Theta}_{\mathcal{L}_{CL}}$ and $MNN^{\Psi}_{\mathcal{L}_{CL}}$ in terms of sensitivity. However, the null hypothesis could not be rejected in terms of PPV.

Finally, a general view of the pair-wise tests in Figure 4.7 (in conjunction with the box-plots of Figure 5.1) also reveals a statistical difference between $MNN^{\Theta}_{\mathcal{L}_{CL}}$ and $MNN^{\Theta}_{\mathcal{L}_{MMI}}$ in terms of sensitivity. Concerning PPV, both models optimizing $\mathcal{L}_{CL}$ were statistically different to $MNN^{\Theta}_{\mathcal{L}_{MMI}}$. Thus, $MNN^{\Theta}_{\mathcal{L}_{CL}}$ appears to be the superior configuration regarding sensitivity, while $MNN^{\Theta}_{\mathcal{L}_{MMI}}$ is better suited for optimal PPV. Henceforth, we choose

$\text{MNN}_{\mathcal{CL}}^{\Theta}$ as the baseline for the remaining experiments, since it offers the most efficient paradigm, as $\mathcal{L}_{\mathcal{CL}}$ can be calculated in $\mathcal{O}(T)$, where T is the length of the sound, without the forward-backward algorithm. Moreover, a static MNN does not need the costly gradient re-projection step.

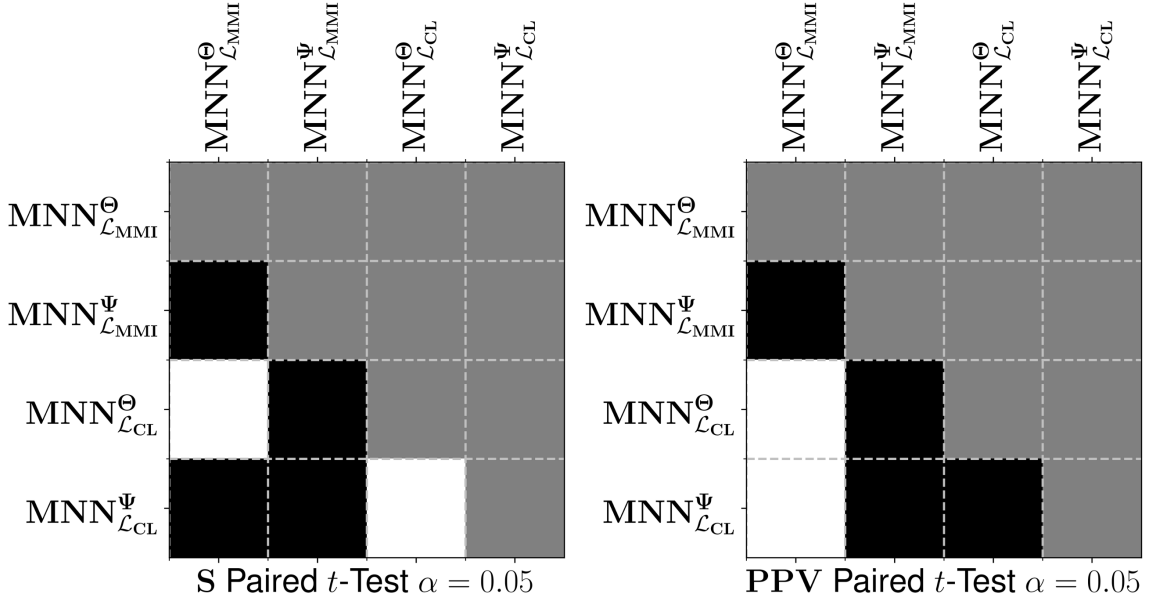


FIGURE 4.7: Pairwise t -test ($\alpha = 0.05$) between S and PPV of $\text{MNN}_{\mathcal{L}_{\text{MMI}}}^{\Theta}$, $\text{MNN}_{\mathcal{L}_{\text{MMI}}}^{\Psi}$, $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Theta}$ and $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Psi}$ in the PhysioNet'16. White cells signify that we can reject the null hypothesis for a pair of models, black cells mark the converse. The grey colour denotes uninformative cells.

4.6.4 Model performance comparison

We now compare $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Theta}$ to the HSMM with logistic regression by Springer et al. [206], the U-Net proposed by Renna et al. [207], and the Bi-LSTM+A introduced by Fernando et al. [209]. We will be performing measurements on 10-fold cross-validations of the PhysioNet'16 and CirCor'22 datasets. We recorded the average performance on each fold in the box-plots of the 10-fold cross-validated experiments in Figures 4.8 and 4.10, alongside pairwise t -tests with a significance level $\alpha = 0.05$ (Figures 4.9 and 4.11).

Concerning PhysioNet'16, an inspection of the box-plots (Figure 4.8) reveals that the models that leverage DNNs are substantially more performant than the HSMM. Our proposed MNN displays superior robustness with regards to both S and PPV. In concrete terms, the proposed $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Theta}$ scored an average S of 0.950 ± 0.022 and PPV of 0.937 ± 0.025 . It scored higher on average than all other models. However, in terms of S , we could not find a significant difference between the Bi-LSTM+A, which measured

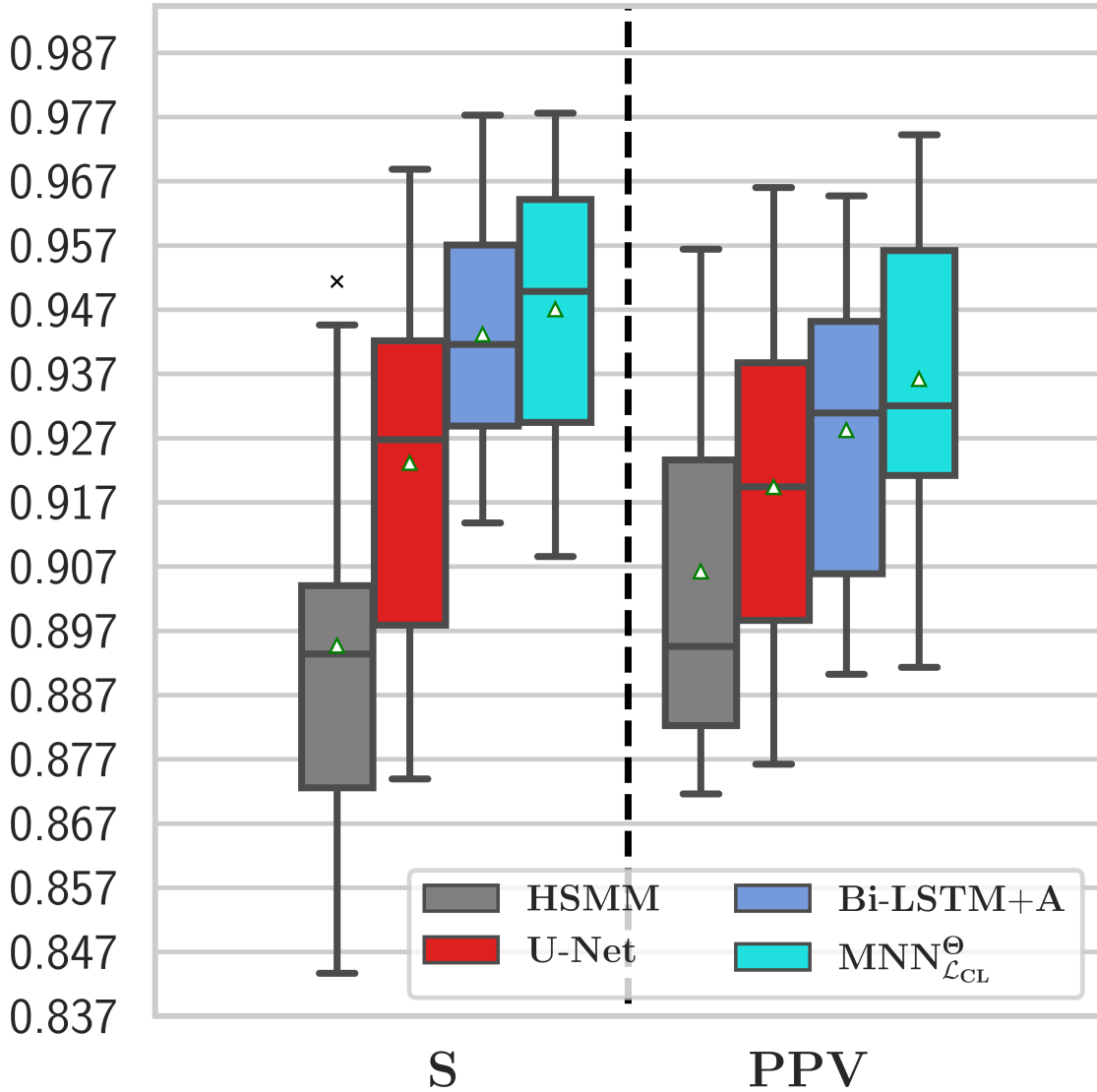


FIGURE 4.8: Measured sensitivity (S) and positive predictive value (PPV) statistics in 10-fold cross-validation on the CirCor'22 dataset between the HSMM by Springer et al. [206], the U-Net by Renna et al. [207], the Bi-LSTM+A by Fernando et al. [209], and $MNN^{\Theta}_{\mathcal{L}_{CL}}$. The green edged triangles mark the mean of the distribution, while the black crosses signify outliers. Note how the MNN with fixed λ during gradient descent using the complete likelihood loss ($MNN^{\Theta}_{\mathcal{L}_{CL}}$) yields the best trade-off between the two metrics.

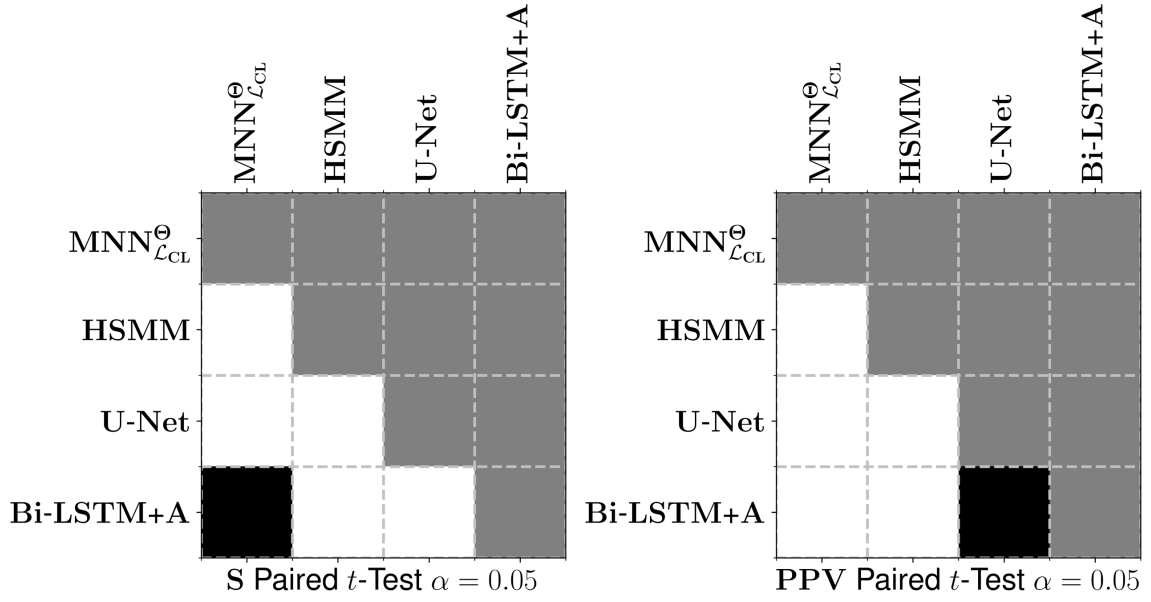


FIGURE 4.9: Pairwise t -test between $MNN^{\Theta}_{\mathcal{L}_{CL}}$, the HSMM by Springer et al. [206], the U-Net by Renna et al. [207], and the Bi-LSTM+A by Fernando et al. [209] with significance $\alpha = 0.05$ in the PhysioNet'16 dataset. White cells signify that we can reject the null hypothesis for a pair of models, black cells mark the converse. The gray color denotes uninformative cells.

an average of 0.944 ± 0.020 S . On the other hand, with respect to PPV, it was significantly superior to all other approaches by a considerable margin, with an average 1.5% improvement over the 0.929 ± 0.026 PPV observed in the Bi-LSTM+A, the second best scoring model.

With respect to the CirCor'22 dataset, the gap between the HSMM and the data-driven solutions becomes even wider. The measurements in Figure 4.10 reveal that the superiority of our model is even more pronounced in this (arguably) more challenging dataset (see Figure 4.10). Our model scored 0.95 ± 0.01 S and 0.94 ± 0.01 PPV. Concerning S , we could not prove statistical difference with the U-Net (see pairwise t -tests in Figure 4.11), which scored 0.94 ± 0.02 . However, one should note that the mean and variance of $MNN^{\Theta}_{\mathcal{L}_{CL}}$ are still higher. Moreover, the U-Net has a sizeable outlier in one of the folders, where it scored just 0.90 S . Concerning PPV, the $MNN^{\Theta}_{\mathcal{L}_{CL}}$ is significantly superior to all other models, scoring an additional 1.1% PPV on average over the U-Net, its runner-up, which registered an average $0.93 \pm 0.0w$ PPV.

4.6.5 Fine-tuning MNN on unseen data

In this section we measure the effects of a pre-trained static MNN in the PhysioNet'16/CirCor'22 as the *source* dataset when fine-tuned to CirCor'22/PhysioNet'16 as the *target* dataset. For

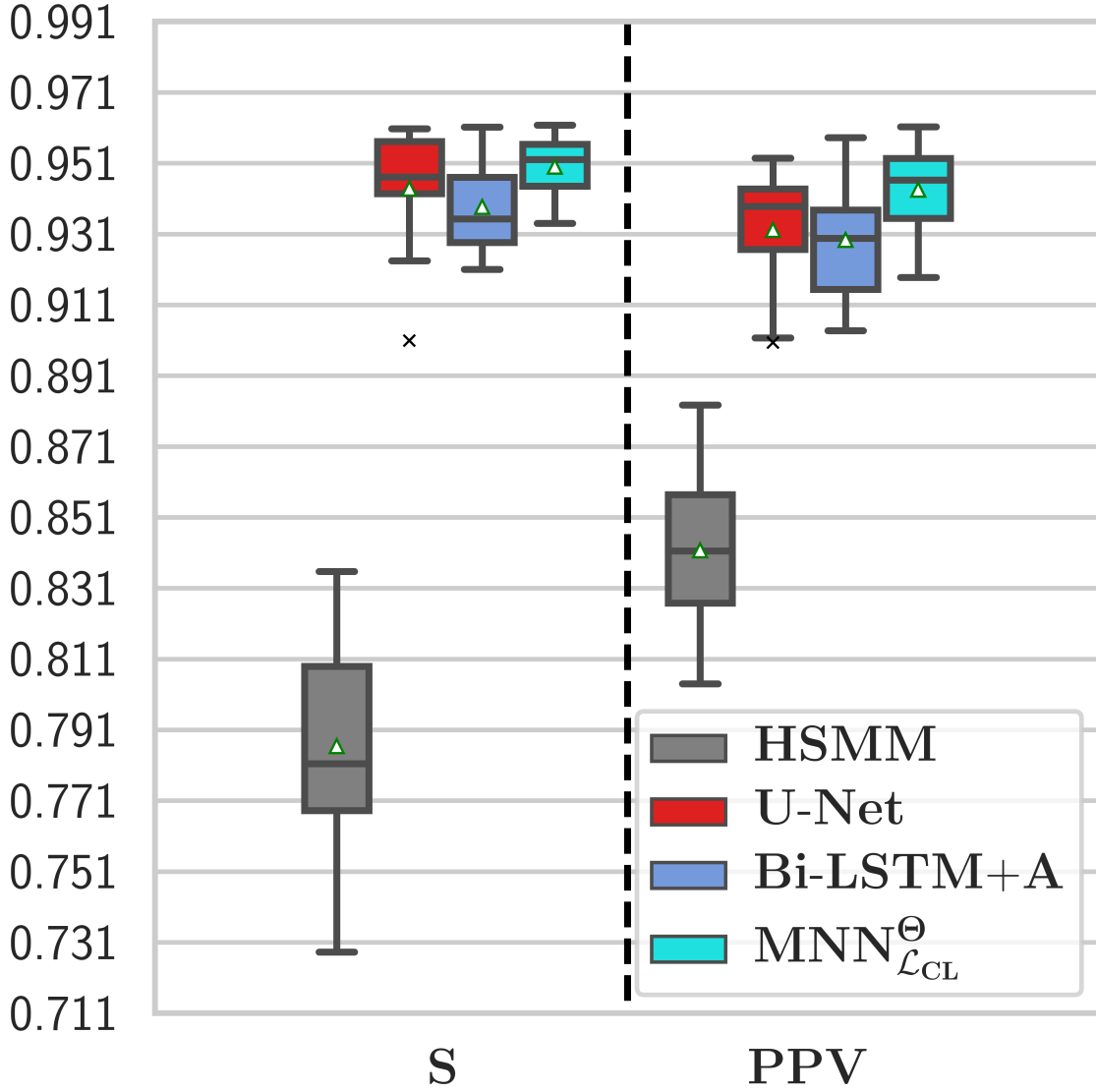


FIGURE 4.10: Measured sensitivity (S) and positive predictive value (PPV) statistics in 10-fold cross-validation on the CirCor'22 dataset between the HSMM by Springer et al. [206], the U-Net by Renna et al. [207], the Bi-LSTM+A by Fernando et al. [209], and $MNN^{\ominus}_{\mathcal{L}_{CL}}$. The green edged triangles mark the mean of the distribution, while the black crosses signify outliers. Note how the MNN with fixed λ during gradient descent using the complete likelihood loss ($MNN^{\ominus}_{\mathcal{L}_{CL}}$) yields the best trade-off between the two metrics.

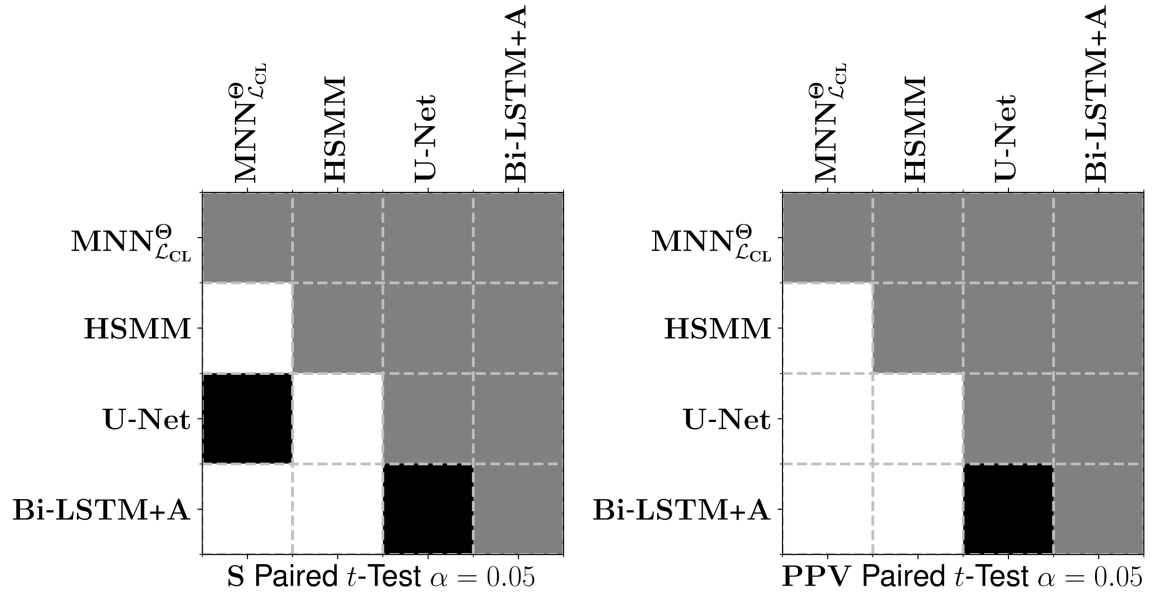


FIGURE 4.11: Pairwise t -Test between $MNN^{\Theta}_{\mathcal{L}_{CL}}$, the HSMM by Springer et al. [206], the U-Net by Renna et al. [207], and the Bi-LSTM+A by Fernando et al. [209] with significance $\alpha = 0.05$ in the CirCor'22 dataset. White cells signify that we can reject the null hypothesis for a pair of models, black cells mark the converse. The gray color denotes uninformative cells.

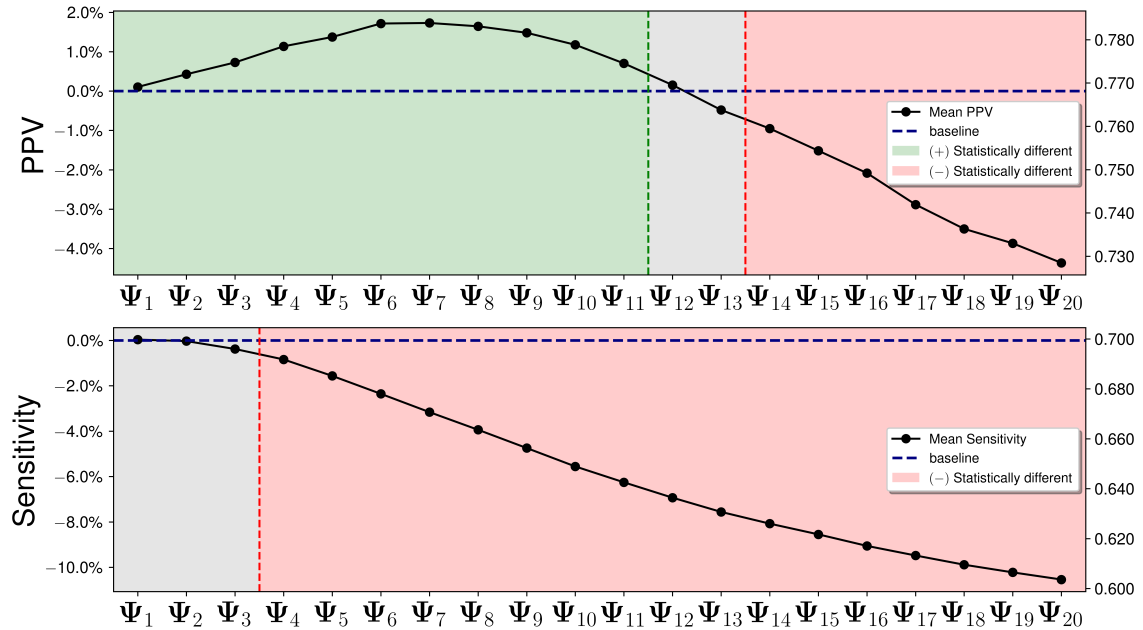


FIGURE 4.12: Fine-tuning average performance of $MNN^{\Psi_0}_{\mathcal{L}_{CL}}$ compared with $MNN^{\Psi_i}_{\mathcal{L}_{FT}}$, for $1 \leq i \leq 20$. The baseline is pre-trained on PhysioNet'16 and fine-tuned throughout different numbers of epochs on the CirCor'22 dataset. (Top) PPV. (Bottom) Sensitivity. A pairwise t -test between the baseline metric distribution and the outcome after i epochs was made to assess statistically significant differences. The left vertical axis displays average relative changes in percentage, and the right horizontal axis the average measured values.

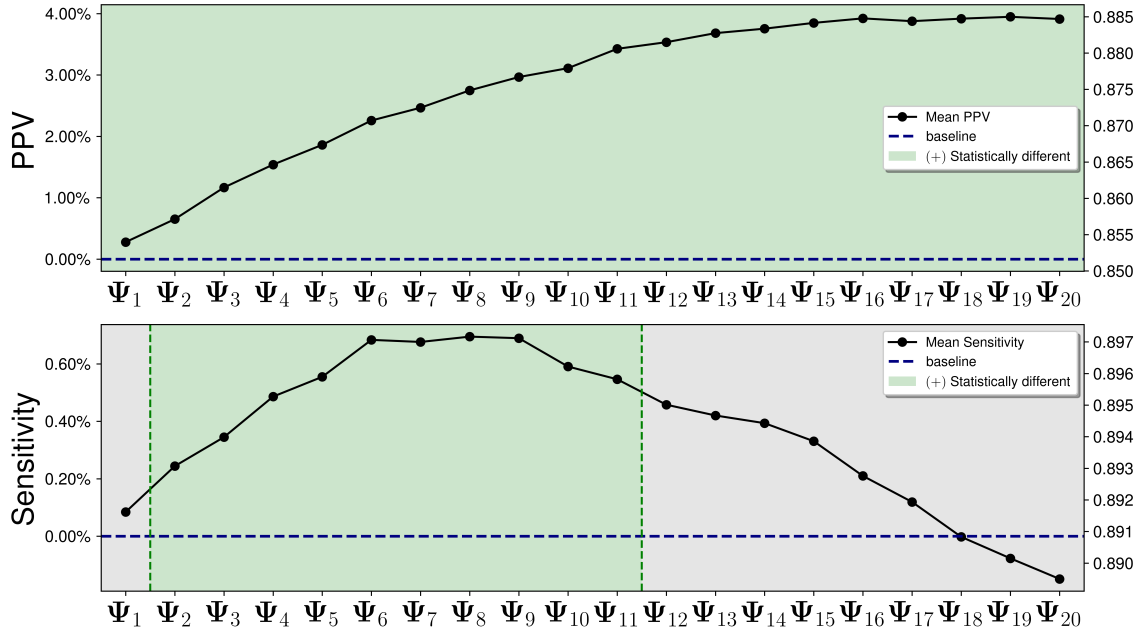


FIGURE 4.13: Fine-tuning average performance of $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Psi_0}$ compared with $\text{MNN}_{\mathcal{L}_{\text{FT}}}^{\Psi_i}$, for $1 \leq i \leq 20$. The baseline is pre-trained on CirCor'22 and fine-tuned throughout a different number of epochs on the PhysioNet'16 dataset. (Top) PPV. (Bottom) Sensitivity. A pairwise t -test between the baseline metric distribution and the outcome after i epochs was made to assess statistically significant differences. The left vertical axis displays average relative changes in percentage, and the right horizontal axis the average measured values.

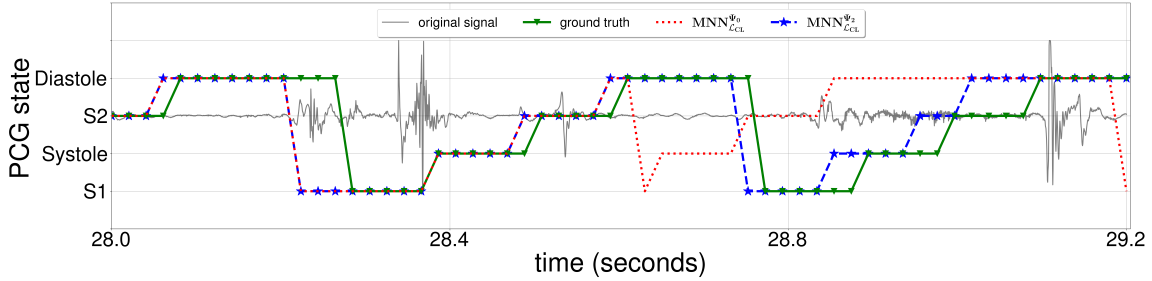


FIGURE 4.14: Predicted state sequences of an unseen CirCor'22 sound using a pre-trained $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Psi_0}$ on PhysioNet'16 as baseline (dotted red line) and an MNN fine-tuned before the sensitivity significance threshold $\text{MNN}_{\mathcal{L}_{\text{FT}}}^{\Psi_2}$ (dashed blue line with star markers). Note how $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Psi_0}$ deviates from the underlying regime, specifically from the diastole in the interval $[28.4, 28.8]$ seconds onward. This behavior is mitigated by $\text{MNN}_{\mathcal{L}_{\text{FT}}}^{\Psi_2}$, which effectively reduces the differences in rhythmic structure between the real and predicted state sequences.

each experiment, we set $\mathcal{L} = \mathcal{L}_{\text{CL}}$ and use random holdout with 80/10/10 split with early-stopping at the best loss value. The parameters of the model attained at this pre-training stage are denoted as Ψ_0 , so that $\Psi_0 = \{\lambda_{\text{MLE}}, \Theta\}$ depends on the train set split of the source dataset. Then, using Algorithm 2, we fine-tuned the model in a hybrid fashion on each observation and recorded the impact of each additional round of fine-tuning

until $k = 20$ epochs (Figures 4.12 and 4.13). We measure the mean performance of each metric in the same folders of the previous experiments for each $\text{MNN}_{\mathcal{L}_{\text{FT}}}^{\Psi_i}$. Pairwise t -tests between $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Psi_0}$ and $\text{MNN}_{\mathcal{L}_{\text{FT}}}^{\Psi_i}$, $1 \leq i \leq k$ with significance $\alpha = 0.05$ (Figure 4.12) were performed to grasp the statistical significance in PPV and S .

When PhysioNet'16 is used as the source and CirCor'22 as the target datasets, the fine-tuning procedure tends to result in an increase in PPV at the cost of S . The results in Figure 4.12 show that PPV increases or stays statistically the same for $i \leq 13$. Afterwards, each subsequent round of fine-tuning is detrimental to the performance of the model. More specifically, the baseline $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Psi_0}$ mean scores are 0.77 PPV and 0.70 S . At iteration $i = 2$ it is possible to have a small increase in average PPV, 0.78 (0.80% statistically significant increase), compromising S to 0.70 (0.01% statistically non-significant decrease). For larger numbers of i , the trade-off in S is always significant.

The results are more promising when CirCor'22 is the source and PhysioNet'16 the target dataset (see Figure 4.13). A steady increase in PPV is observed throughout all i with a positive or negligible impact in S . In fact, baseline $\text{MNN}_{\mathcal{L}_{\text{CL}}}^{\Psi_0}$ scored a mean PPV of 0.85 and mean S of 0.89, while the model $\text{MNN}_{\mathcal{L}_{\text{FT}}}^{\Psi_{20}}$ scores the best average PPV at 0.89 (3.90% statistically significant increase) with mean sensitivity of 0.89 (0.20% statistically non-significant decrease), which is a substantial improvement in performance.

4.6.6 Discussion and conclusion

We instantiated left-to-right MNNs for the downstream task of phonocardiogram fundamental heart sound segmentation, demonstrating the superiority of the novel framework over two recent fully data-driven architectures. The experiments were based on results from two publicly available datasets: PhysioNet'16 and CirCor'22. In the supervised case, we showed that the simplest MNN using one-dimensional feature maps is superior to both U-Net and Bi-LSTM+A. It was also shown that an MNN is adaptive to new data sampled from a dissimilar distribution via the proposed unsupervised fine-tuning procedure.

Below we summarise our main findings within the scope of this downstream application.

(a) The Markov model is a strong inductive prior for heart sound segmentation: The fact that the proposed MNN algorithm supported by a very simple CNN (18785 parameters) consistently outperforms models with 9.52 and 3.40 times the number of parameters,

such as the U-Net (178828 parameters) and Bi-LSTM+A (63908 parameters), highlights the impact the Markovian inductive bias has for the task of PCG heart sound segmentation. Note that this is true for both PhysioNet'16 and CirCor'22, regardless of the sizeable covariate shift in age groups, the heart sound duration statistics (see Figure 4.15), and the overall average length of the recordings between the two datasets. Furthermore, even for CirCor'22, where one would expect the larger dataset (roughly 4.12 times the number of samples in PhysioNet'16) to benefit more complex networks, our experiments show that MNNs remain superior.

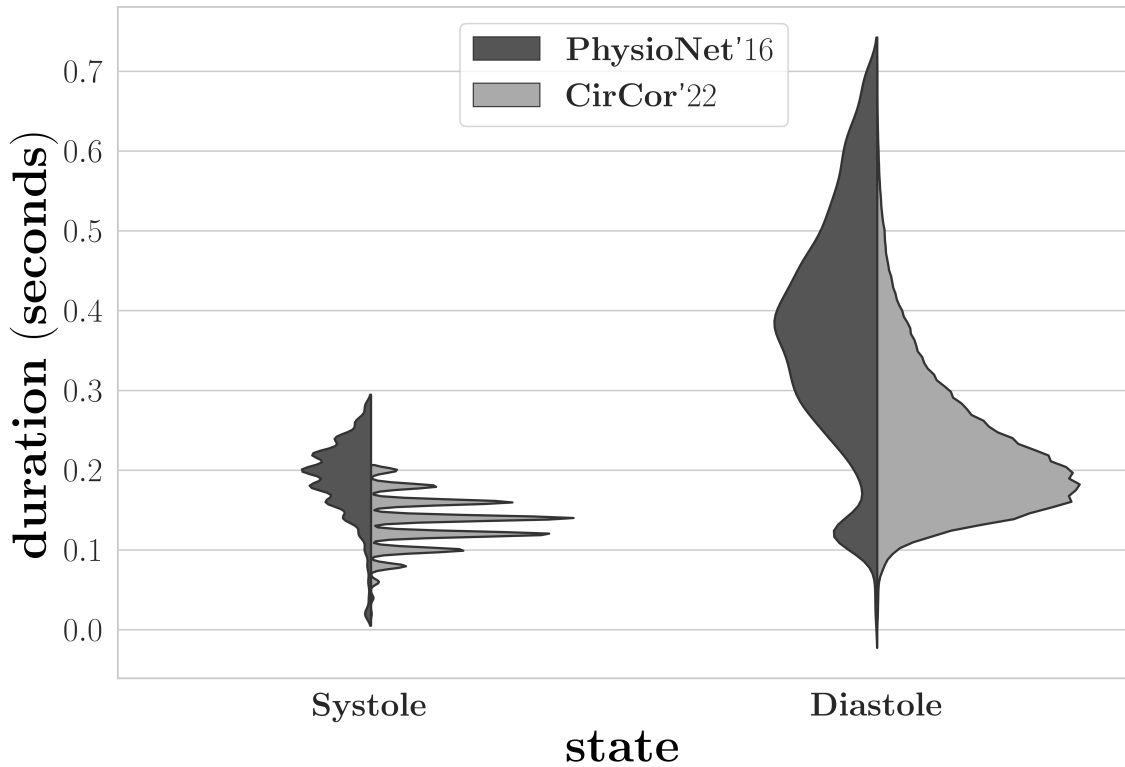


FIGURE 4.15: Distributions of the systole and diastole heart duration in the PhysioNet'16 (dark gray) and CirCor'22 (light gray). To improve readability, we plotted only values within two standard deviations of the mean.

We surmise that MNN performance could be further improved by using increasingly complex discriminators or feature-extraction components. We believe this is especially likely in the unsupervised fine-tuning case, since we found that the gradient of $\mathcal{L}_{CL}$ depended mostly on the CNN, with the underlying HMM parameters undergoing only very subtle changes during this type of training.

(b) MNNs allow for subject-driven fine-tuning, but the stop criterion needs to be defined: Algorithm 2 effectively allows MNNs to adapt to the rhythmic structure of unseen sounds sampled from different latent distributions, as can be observed in Figure

4.14. On the other hand, our experiments reveal that the success of this procedure heavily depends on the source and target datasets.

When CirCor'22 is used as the source domain, we did not observe a significant decrease in performance as the number of epochs increased. We hypothesize that pre-training on a larger, more varied source dataset may yield a more robust discriminator, which in turn leads to a more stable fine-tuning procedure. When PhysioNet'16 was used as the source domain, the model quickly became over-tuned as the number of epochs increased. It is still unclear how this over-tuning phenomenon could be avoided in real-world scenarios where access to labeled data might be impossible.

(c) The Markovian prior is strong, but it is not perfect: We note that the parameter set λ can be seen as a global descriptor of the statistics of the sequences produced by the MNN, but lacks local information. We envision using statistics on local signal invariances as additional prior information for the MNN to mitigate over-tuning. Other limitations of this study are: the fact that we assume that fundamental heart sounds occur in an unchanged cyclic fashion, which is true from a strictly physiological perspective. However, these assumptions are too strong once we appreciate that mistakes may occur during sound collection (e.g., accidentally lifting the stethoscope). Note that the proposed MNN framework is already sufficiently generic to account for this phenomenon. One needs only to include an extra state for such an occasion, and change the underlying automaton from left-to-right to a more densely connected first-order HMM. Furthermore, we only studied HMMs as the inductive bias for the time modulation of PCG heart signals. Notwithstanding, our framework is easily extended to support more sophisticated priors, such as an adaptation of the HSMM model proposed by Springer et al. [206].

(d) The sequential nature of the likelihood functions can be a computational limiting factor: Finally, we note that our framework has some computational performance barriers, especially during training. The proposed loss functions (Equations 4.5, 4.4 and 4.7) require sequential computation of the exact pseudo-likelihood estimates, which are still not effectively parallelizable using tensor algebra (more details in Appendix C.4). Consequently, although the model is less complex and its training is theoretically more time efficient, it is nonetheless slower to train in practice compared to the more complex U-Net. On the other hand, our MNN instantiation is faster during inference than the U-Net and the Bi-LSTM+A given its very simple discriminator and feature extraction pipeline. We note that in order to guarantee physiologically valid predictions, all models share the

same limiting factor: the decoder. In fact, we are contemplating non-sequential extensions of our framework for real-time screening, since the Viterbi algorithm requires complete observation of a sound in order to decode its output sequence. A possible next step in this direction is to couple our MNN with an online short-time Viterbi decoder [213], thereby enabling the MNN to emit valid sequences in real time.

4.7 Chapter conclusion

In this chapter, we introduced MNNs by formalising a set of principles that enable the embedding of an underlying HMM into a highly discriminative, data-driven neural network through a unifying optimisation framework. We proposed algorithms that encapsulate gradient-descent strategies for supervised and unsupervised learning and guarantee that the Markovian assumption holds via a projective strategy for the gradient updates.

For future work, end-to-end estimation of the model’s exact likelihood is of particular interest. This allows for a full specification of the model. When the Markovian prior is not mismatched with the distribution, the fine-tuning stage can have an actual stopping criterion based on the loss function itself, a limitation highlighted in point (b) of Section 4.6.6.

Note that, as stated in Definition 4.5, we do not compute $p(x)$, resulting in an optimisation of a *pseudo*-likelihood. Of special interest for these ends is leveraging recent advances in nonlinear identifiability theory [214–216], where the likelihood values of the observed measurements factorise over a latent factor $\mathbf{z}$. Another interesting direction is to combine this task with existing labels (such as the heart state labels for the task in Section 4.5), since adequate labeling of the dataset can help better identify the generating model [217].

4.7.1 Outcomes

- 1 publication in an international peer-reviewed conference [218].
- 1 publication in a Q1 Journal [219].
- 1 international [220] and 1 national [221] best presentation awards.
- 5 international and national tutorials on heart sound segmentation [222–225] (check if ENEEB 5/09/25 is going forward).

- A HuggingFace dataset [\[226\]](#) integration of CirCor'22 dataset with the raw data, and feature extraction [\[227\]](#).

Chapter 5

No free lunches for self-supervised learning

When we started this thesis it was not clear that just continuing scaling the models would be an effective direction, and the community raised some concerns about the possibility of a second AI winter [228–231].

With recent advancements in self-supervised learning (SSL), not only are able to leverage unlabeled large-scale datasets at an unprecedented ammount [232, 233], but SSL methods appear to learn regularities that can be expressed in their representation spaces without any constraints in the training trask [234–237]. As of writing this thesis, the days of fearing another AI winter are far gone.

Yet, despite the frontier-level performance achieved by models trained with SSL, and in particular contrastive-based schemes, their inner mechanisms remain poorly understood. Prior work [238, 239] has studied the regularities that different methods induce in representation space, with some studies suggesting a convergence towards a set of Platonic Representations [240].

We lack an understanding of the types of possible manifolds that emerge in nature [241, 242], or even whether current machine learning methods could invert or even learn patterns in such manifolds [243, 244]. Notwithstanding, for most practical applications, it is fair to assume that, at the very least, the way the architectures and optimisation frameworks were designed does indeed capture inherent properties of the data that seemingly match Human intuition [245].

Given the model-based *logos* of this Thesis, we decided to focus on the implications of inverting, or identifying patterns on some hypothetical data generating process (DGP) of the world, and most importantly, how this relates to downstream performance.

As a first step toward outlining the limitations of these models, we study contrastive learning methods on an arbitrary downstream classification task. We set forth to demonstrate that these models need to obey the “no free lunch” theorems alluded to in Chapter 2.

Overall, our contributions are summarized as follows:

1. We reframe the contrastive learning paradigm as **solving an inverse problem with very limited prior knowledge**, where data augmentation acts as a regularizer.
2. We introduce a DGP that accounts for **how augmentation affects the distance between instances in latent and observable spaces**. Within this DGP, we articulate no-free-lunch theorems for contrastive learning, relating the DGP’s invertibility to its downstream performance on classification tasks.
3. We observe a **trade-off between classification performance and identifiability**, both in controlled numerical simulations, but also in disentanglement benchmarks of images where the complete effect of the augmentations cannot be known a priori.

An overview of our framework’s intuition is presented in Figure ??.

5.1 Generalised contrastive learning problems

Key to the notion of self-supervised learning are the following conditions:

1. The observations $\mathbf{x} \in \mathcal{X}$ depend on some factorizable latent distribution $\mathbf{z} \in \mathcal{Z}$
2. The latents are $\log p(\mathbf{z}|\mathbf{u}) = \sum_{i=1}^n q_i(z_i, \mathbf{u})$, where $\mathbf{u} \in \mathcal{U}$
3. Learning occurs by contrasting $(\mathbf{x}, \mathbf{u})$ vs. $(\mathbf{x}, \mathbf{u}^*)$

Crucial to this chapter is the duality of contrastive learning with instance discrimination tasks. We closely follow the treatise established in [246] to demonstrate their equivalence in concrete terms.

Definition 5.1 (General Contrastive Learning Problem). Let $\mathbf{X} \in \mathcal{X}$, where $\mathbf{X} = (\mathbf{x}, \mathbf{x}^+, \mathbf{x}_1^-, \dots, \mathbf{x}_M^-) \sim p_{\mathbf{X}}$. By construction, $\mathbf{x}$ is the *anchor point* with *positive view* $\mathbf{x}^+$ and *negative views* $\mathbf{x}_i^-$. We assume:

i) **Mutual Independence:** The views are mutually independent given the anchor:

$$\mathbf{x}^+ \perp\!\!\!\perp \mathbf{x}_1^- \perp\!\!\!\perp \cdots \perp\!\!\!\perp \mathbf{x}_M^- \mid \mathbf{x}$$

ii) **Negative views are identically distributed w.r.t. anchor:** given anchor $\mathbf{x}$:

$$p(\mathbf{x}_i^- \mid \mathbf{x}) = p(\mathbf{x}_j^- \mid \mathbf{x}) \quad \forall i, j$$

Let $\mathcal{U}$ be a class of bi-variate functions $u : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. Consider the following objective:

$$\min_{u \in \mathcal{U}} \mathcal{L}_g(u) := \mathbb{E} \left[-\ln \frac{e^{u(\mathbf{x}, \mathbf{x}^+)}}{e^{u(\mathbf{x}, \mathbf{x}^+)} + \sum_{j=1}^M e^{u(\mathbf{x}, \mathbf{x}_j^-)}} \right]. \quad (5)$$

Due to the presence of the positive view in the numerator, u is usually called a *similarity* function. The tuple $(\mathcal{X}, p_X, \mathcal{U}, \mathcal{L}_g)$ is called a *general contrastive learning (CL) problem*.

TODO: instantiate this using our notation $\mathfrak{C}$.

Note that by Definition 5.1:

$$p(\mathbf{x}, \mathbf{x}^+, \mathbf{x}_1^-, \dots, \mathbf{x}_M^-) = p(\mathbf{x}) p(\mathbf{x}^+ \mid \mathbf{x}) \prod_{j=1}^M p(\mathbf{x}_j^- \mid \mathbf{x}). \quad (5.1)$$

Definition 5.2 (Instance Discrimination Problem). The *instance discrimination data* $(\mathbf{x}, \tilde{\mathbf{x}}_0, \dots, \tilde{\mathbf{x}}_M; K)$ corresponding to a general CL problem $(\mathcal{X}, p_X, \mathcal{U}, \mathcal{L}_g)$ is defined by the following process:

- $\mathbf{x} \sim p_X$
- $K \sim \text{Uni}(\{0, 1, \dots, M\})$
- $\tilde{\mathbf{x}}_K \sim p(\mathbf{x}^+ \mid \mathbf{x})$
- $\forall j \in \{0, \dots, M\} \setminus \{K\} : \tilde{\mathbf{x}}_j \sim p(\mathbf{x}^- \mid \mathbf{x})$.

where Equation 5.1 can be written as:

$$p(\mathbf{x}, \tilde{\mathbf{x}}_0, \dots, \tilde{\mathbf{x}}_M, K = i) = p(\mathbf{x}) p(K = i) p(\tilde{\mathbf{x}}_i \mid \mathbf{x}) \prod_{0 \leq j \leq M, j \neq i} p(\tilde{\mathbf{x}}_j \mid \mathbf{x}). \quad (5.2)$$

An *instance discrimination problem* comes from optimizing for the following objective.

$$\min_{u \in \mathcal{U}} \mathcal{L}^{id}(u) := \mathbb{E} \left[-\sum_{i=0}^M \delta_{K=i} \ln \frac{e^{u(\mathbf{x}, \tilde{\mathbf{x}}_i)}}{\sum_{j=0}^M e^{u(\mathbf{x}, \tilde{\mathbf{x}}_j)}} \right], \quad (5.3)$$

where δ_E is the indicator variable of event E . Solving this task is equivalent to solving a generalised contrastive learning problem (Definition 5.1), with positive views $\tilde{\mathbf{x}}_k$ and negative views $\tilde{\mathbf{x}}_j, j \neq k$.

It might not be immediately apparent that solving an instance discrimination task is equivalent to a contrastive learning problem. Intuitively speaking, encouraging a higher response of the similarity function u at $(\mathbf{x}, \mathbf{x}^+)$ does indeed amount to at least a smooth instance discrimination of the positive pair. The formal statement is described below.

Theorem 5.3. $\mathcal{L}_g(u) = \mathcal{L}^{id}(u)$, for any $u \in \mathcal{U}$.

Proof. First, using the law of total expectation $\mathbb{E}[A] = \mathbb{E}[\mathbb{E}[A|B]]$, we expand $\mathcal{L}^{id}$ w.r.t. K :

$$\mathcal{L}^{id}(u) = \mathbb{E}_K \left[\mathbb{E} \left[- \sum_{i=0}^M \delta_{K=i} \ln \frac{e^{u(\mathbf{x}, \tilde{\mathbf{x}}_i)}}{\sum_{j=0}^M e^{u(\mathbf{x}, \tilde{\mathbf{x}}_j)}} \middle| K \right] \right] \quad (12)$$

$$= \frac{1}{M+1} \sum_{k=0}^M \mathbb{E} \left[\sum_{i=0}^M \delta_{K=i} \ln \frac{e^{u(\mathbf{x}, \tilde{\mathbf{x}}_i)}}{\sum_{j=0}^M e^{u(\mathbf{x}, \tilde{\mathbf{x}}_j)}} \middle| K = k \right]. \quad (13)$$

For each term, the δ function will set to 0 all indices $i \neq k$, so one can state:

$$\mathbb{E} \left[- \sum_{i=0}^M \delta_{K=i} \ln \frac{e^{u(\mathbf{x}, \tilde{\mathbf{x}}_i)}}{\sum_{j=0}^M e^{u(\mathbf{x}, \tilde{\mathbf{x}}_j)}} \middle| K = k \right] = \mathbb{E} \left[- \ln \frac{e^{u(\mathbf{x}, \tilde{\mathbf{x}}_k)}}{\sum_{j=0}^M e^{u(\mathbf{x}, \tilde{\mathbf{x}}_j)}} \middle| K = k \right] \quad (14)$$

Note also that

$$p(\mathbf{x}, \tilde{\mathbf{x}}_0, \dots, \tilde{\mathbf{x}}_M | K = k) = p(\mathbf{x}) p(\tilde{\mathbf{x}}_k | \mathbf{x}) \prod_{0 \leq j \leq M, j \neq k} p(\tilde{\mathbf{x}}_j | \mathbf{x}).$$

Therefore, the expectation given that we set index k for the instance discrimination task is exactly the same as selecting some $\mathbf{x}^+ \in \mathcal{X}$ for $\mathcal{L}_g(u)$ in Equation 5, so:

$$\mathcal{L}^{id}(u) = \frac{1}{M+1} \sum_{k=0}^M \mathcal{L}_g(u) = \mathcal{L}_g(u). \quad (15)$$

□

Note that this particular articulation is used in most recent identifiably proofs in CL [246–248]. In fact, this precise articulation relates to the framing of CL as an instance discrimination task and has led to crucial theoretical and practical advancements, for instance, by enabling CL to be conducted with linear projectors and small batch sizes [249–252].

5.2 No free lunches for Self-supervised Learning

We focus on the inversion of the data generation process (DGP) and frame it within the broader context of the inverse problem literature. Of special interest to us are recent results by [253–255] that relate to Smale’s 18th problem [256]: even if a neural network exists that coincides with the underlying true map between the latents in the DGP and the observables, can our methods actually learn that map?

Unlike prior work that focuses on the performance of SSL in downstream tasks [257–259], we draw from recent identifiability theory of CL and model the latents $\mathbf{z}$ of the observation $\mathbf{x}$ explicitly [246, 248, 260, 261]. Our contrastive learning task includes the space of latents and observables:

Assumption 5.4. *The DGPs we study obey the following conditions:*

1. *There is $\mathcal{I}$, a set of indices, and $\mathcal{C}$, a set of downstream classes. There exists a surjective labeling function $c : \mathcal{I} \rightarrow \mathcal{C}$. $c(i) \equiv c_i$ denotes the class of instance i .*
2. *The indices i are sampled uniformly $p_{\mathcal{I}} \sim \text{Uniform}(\mathcal{I})$.*
3. *Let $(\mathcal{Z}, \|\cdot\|_{\mathcal{Z}})$ and $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be normed vector spaces.*
4. *There is a continuous, injective (and invertible on its image), possibly non-linear, map $g : \mathcal{Z} \rightarrow \mathcal{X}$.*
5. *Let $\mathcal{T}$ be a set of transformations $t : \text{Im}(g) \rightarrow \text{Im}(g)$. An augmentation τ is an element of $\mathcal{T}$, the monoid with respect to function composition^{*}, $\langle \mathcal{T} \rangle$.*
6. *Each instance $i \in \mathcal{I}$ has an instance vector $\mathbf{z}_i \in \mathcal{Z}$.*
7. *Each class $c \in \mathcal{C}$ has a unique class vector $\mathbf{z}_c \in \mathcal{Z}$.*

For the purposes of this chapter, it is instructive to consider the specific case when $\mathcal{I}, \mathcal{C}$ are finite.

Assumption 5.5. *The data generation process $G : \mathcal{Z} \rightarrow \mathcal{X}$ unfolds as follows: when instantiated, sample index i from $p_{\mathcal{I}}$ and τ from $p_{\langle \mathcal{T} \rangle}$. Each i is assigned exactly to one latent vector $\mathbf{z}_i \in \mathcal{Z}$, with corresponding observation $\mathbf{x}_i = g(\mathbf{z}_i) \in \mathcal{X}$. Finally, output $\tau(\mathbf{x}_i)$.*

^{*}The set of all possible function compositions of $t \in \mathcal{T}$ including $\text{id}_{\mathcal{X}}$.

We re-instantiate the instance discrimination task defined in 5.1 within the constraints of our DGP:

$$\mathcal{L}^{id} := \mathbb{E}_{\substack{\mathbf{x} \sim p_{\mathcal{X}}, i \sim p_{\mathcal{I}} \\ \tau \sim p_{\langle \mathcal{T} \rangle}}} \left[-\ln \frac{\exp(u(\tau(\mathbf{x}_i), i))}{\sum_{j \in \mathcal{I}} \exp(u(\tau(\mathbf{x}_j), j))} \right]. \quad (5.4)$$

This instantiation of a contrastive learning loss better fits our notation and aligns with the methods we selected for our empirical experiments [250].

We rely on the notion of $\mathcal{T}$ -connectedness introduced in [259], but instantiate it in the context of normed vector spaces:

Definition 5.6 ($\mathcal{T}$ -connectedness). Consider a set of instances $\mathcal{I}' \subseteq \mathcal{I}$, if for $\delta > 0$:

$$\begin{aligned} \mathcal{I}' = \{i \in \mathcal{I} : & \|\tau_1(\mathbf{x}_i) - \tau_2(\mathbf{x}_j)\|_{\mathcal{X}} \leq \delta, \\ & j \in \mathcal{I}, \tau_1, \tau_2 \in \langle \mathcal{T} \rangle\}, \end{aligned}$$

then $\mathcal{I}'$ is $\mathcal{T}$ -connected with diameter at most δ .

Now that we have specified the tools of analysis for the augmentations in our DGP, we are ready to extend the no-free-lunch results from inverse imaging, specifically Theorem 4.5 of [255].

Theorem 5.7 (Identifiably-stability tradeoff). Consider the case where we get $f : \mathcal{X} \rightarrow \text{Im}(g)$ by minimizing $\mathcal{L}_{\text{GCL}}$. Suppose that $\Psi = \Phi \circ f$, where $\Phi : \text{Im}(g) \rightarrow \mathcal{Z}$ is some continuous, affine (possibly non) linear map. If for $i, j, k \in \mathcal{I}$:

$$\|\Psi(G(\mathbf{z}_i)) - \mathbf{z}_i\|_{\mathcal{Z}} < \delta, \quad \|\Psi(G(\mathbf{z}_j)) - \mathbf{z}_k\|_{\mathcal{Z}} < \delta \quad (5.5)$$

and

$$\|G(\mathbf{z}_i) - G(\mathbf{z}_j)\|_{\mathcal{X}} \leq \delta, \quad (5.6)$$

then the inversion of the generation process G by Ψ is unstable, i.e., its Lipschitz constant is bounded below by $\|\mathbf{z}_i - \mathbf{z}_k\|_{\mathcal{Z}} - 2\delta$.

Proof. The proof follows closely the original arguments of [255]. Define the ϵ -Lipschitz constant of Ψ at $\mathbf{z}$ as

$$L^\epsilon(\Psi, \mathbf{z}) = \sup_{\|\epsilon\|_{\mathcal{X}} \leq \epsilon} \frac{\|\Psi(G(\mathbf{z})) - \Psi(G(\mathbf{z}) + \epsilon)\|_{\mathcal{Z}}}{\|\epsilon\|_{\mathcal{X}}} \quad (5.7)$$

Suppose that $\epsilon \geq \delta$. We know that Ψ reconstructs $\mathbf{z}_i$ within up to δ . We also know that, G maps $\mathbf{z}_i$ and $\mathbf{z}_j$ to be close to one another on $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$. The inversion of $\mathbf{z}_i$ will lie close

to that of $\mathbf{z}_k$, since $\|G(\mathbf{z}_i) - G(\mathbf{z}_j)\|_{\mathcal{X}} \leq \delta \leq \epsilon$. Remember that $\epsilon \geq \delta$, so

$$\begin{aligned} L^\epsilon(\Psi, \mathbf{z}_i) &\geq \frac{1}{\delta} \|\Psi(G(\mathbf{z}_i)) - \Psi(G(\mathbf{z}_j))\|_{\mathcal{Z}} \\ &\geq \frac{1}{\delta} \|\Psi(G(\mathbf{z}_i)) - \mathbf{z}_i - (\Psi(G(\mathbf{z}_j)) - \mathbf{z}_k) - (\mathbf{z}_k - \mathbf{z}_i)\|_{\mathcal{Z}} \end{aligned}$$

using the reverse triangle inequality, we have:

$$\begin{aligned} &\geq \frac{1}{\delta} \left(\|\Psi(G(\mathbf{z}_i)) - \mathbf{z}_i\|_{\mathcal{Z}} \right. \\ &\quad \left. - \|\Psi(G(\mathbf{z}_j)) - \mathbf{z}_k\|_{\mathcal{Z}} - \|\mathbf{z}_k - \mathbf{z}_i\|_{\mathcal{Z}} \right) \\ &> \frac{1}{\delta} (-2\delta + \|\mathbf{z}_k - \mathbf{z}_i\|_{\mathcal{Z}}) \end{aligned}$$

Taking into account the strict inequality and continuity of Ψ , the global Lipschitz constant is bounded below as a function of $\|\mathbf{z}_i - \mathbf{z}_k\|_{\mathcal{Z}}$, which can be arbitrarily larger than δ . $\square$

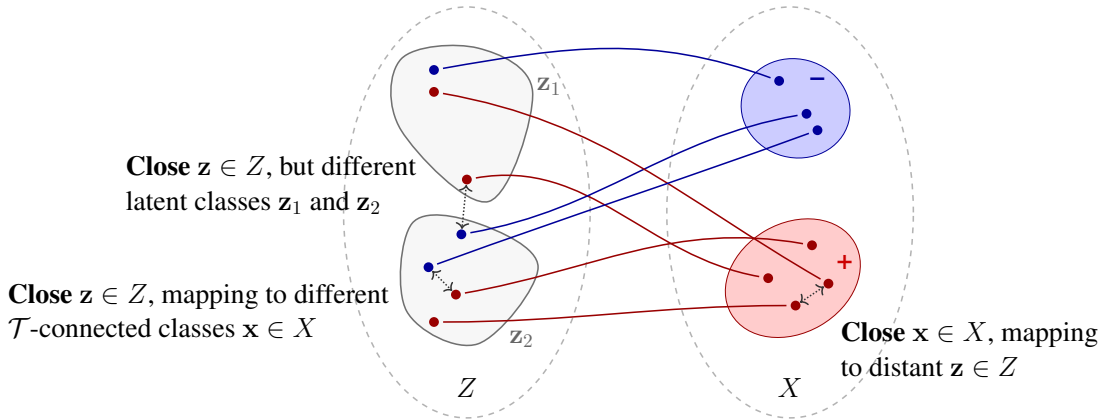


FIGURE 5.1: We study the relationship between identifiability and downstream performance of a model trained using contrastive learning. Illustrated is a particularly problematic mapping for the contrastive learning task: if the model achieves high accuracy on the classes in $\mathcal{X}$, it likely comes at the expense of inverting certain latent factors. On the other hand, when the model is very capable of inversion, then downstream performance is likely compromised. If the model is simultaneously capable of achieving high accuracy and identifying its latent factors, the only remaining outcome is instability, i.e., sensitivity to small perturbations.

Remark 5.8 (Reconstructing latents collapsed by G increases instability). The main driving factor for instability is that G can *collapse* latents that are far apart on $\mathcal{Z}$ to lie very close together on $\mathcal{X}$. Two cases require closer attention as given in the statement of Theorem 5.7:

- For $\mathbf{z}_j = \mathbf{z}_k$: consider $i, j \in \mathcal{I}$ so that $\|\mathbf{z}_i - \mathbf{z}_j\|_{\mathcal{Z}} > 2\delta$, then observing a *collapsed* version from observation space $\mathcal{X}$ makes the inversion of G gets more and more unstable. In fact, it gets worse as Ψ becomes better at the simultaneous inversion of $\mathbf{x}_i$ and $\mathbf{x}_j$.
- For $\mathbf{z}_j \neq \mathbf{z}_k$: then if $\|\mathbf{z}_i - \mathbf{z}_j\|_{\mathcal{Z}} \leq \delta$, inversion will be inconsistent with respect to $(\mathcal{Z}, \|\cdot\|_{\mathcal{Z}})$; Ψ will invert $\mathbf{x}_i$ correctly but fail to invert $\mathbf{x}_j$. This happens in spite of the fact that they lie close to one another in the latent space $\mathcal{Z}$. This problem gets worst as $\|\mathbf{z}_i - \mathbf{z}_k\|_{\mathcal{Z}}$ increases.

As training unfolds, one simply has access to the observations $\mathbf{x}$. Note that G depends on $p_{\langle \mathcal{T} \rangle}$, so this means that the stability depends on one's guess of an appropriate $\mathcal{T}$. This uncertainty can only increase when considering the interplay of $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ with $(\mathcal{Z}, \|\cdot\|_{\mathcal{Z}})$. It thus stands to reason that predicting $\mathbf{z}$ by f optimised in these settings should be very unexpected. Indeed, observing high downstream performance while simultaneously inverting the DGP is likely a symptom of instability of the learned representation, as we will now show.

Theorem 5.9 (Accuracy-identifiability trade-off). *Assume that we learn a continuous f after performing some GCL task. Consider Φ^{id}, Φ^{cls} as continuous, affine (possibly non) linear maps of the form $\text{Im}(g) \rightarrow \mathcal{Z}$, and their composition with f denoted as Ψ^{id}, Ψ^{cls} , respectively. Suppose that the following $\mathcal{T}$ -connected class $c \in \mathcal{C}$.*

$$\sup_{\mathbf{z}_k \in \mathcal{I}'} \|G(\mathbf{z}_k) - G(\mathbf{z}_c)\|_{\mathcal{X}} < \delta.$$

with

$$\|\mathbf{z}_i - \mathbf{z}_c\|_{\mathcal{Z}} > 2\delta$$

Then, predicting c with high accuracy, i.e.

$$\|\Psi^{cls}(\mathbf{x}_i) - \mathbf{z}_c\|_{\mathcal{Z}} < \delta \tag{5.8}$$

implies instability in the identification task.

Proof. If Ψ^{cls} is accurate then it means that for $k \in \mathcal{I}' \subseteq \mathcal{I}$, with $\mathcal{I}'$ being $\mathcal{T}$ -connected, then:

$$\sup_{\mathbf{z}_k \in \mathcal{I}'} \|f(G(\mathbf{z}_k)) - G(\mathbf{z}_c)\|_{\mathcal{X}} < \delta.$$

On the other hand, if $j \in \mathcal{I}'$ with $\|\mathbf{z}_j - \mathbf{z}_c\|_{\mathcal{Z}} > 2\delta$ then we know from Theorem 5.8 that Ψ^{id} is not a stable inversion of the DGP around $G(\mathbf{z}_c)$. $\square$

Remark 5.10 (Simultaneous identifiability and downstream accuracy should arise suspicion). A slightly more nuanced consequence of Theorems 5.7 and 5.9 is the case when Ψ^{id} recovers well instances **belonging to separate classes** that were collapsed by G , while Ψ^{cls} is simultaneously accurate on both of them (see Figure ??). This means that there exist instances i, j that are $\mathcal{T}$ -connected to a class that do not align with $\mathbf{z}_c$. Such samples may lie on an ϵ -ball around $G(\mathbf{x}_i)$ and $G(\mathbf{x}_j)$.

Remark 5.11 (Entangled classifiers). Consider the conditions in Theorem 5.9. Suppose that, for each class c_1, c_2 , there are $\mathcal{T}$ -connected sets of instances $\mathcal{I}_{c_1}$ and $\mathcal{I}_{c_2}$. If these classes are not well separated enough in $\mathcal{X}$, their error may depend on $\|\mathbf{z}_i - \mathbf{z}_k\|_{\mathcal{Z}}$. If $\mathbf{z}_i$ and $\mathbf{z}_k$ map to distinct classes in $\mathcal{X}$, then f necessarily entangles them.

5.3 Related work

Despite the plurality of SSL models, the most popular methods essentially solve the same underlying task [262, 263]. A striking connection is that the InfoNCE loss [264] can be reduced to an instance discrimination task, as recently shown by [248]. The same instant discrimination objective has been used to strip away most of the methodological complexity of SSL, while maintaining performance comparable to more “over-engineered” SSL methods [250, 252]. In fact, [233] recently disclosed a position paper arguing that researchers should focus on the commonalities and not the dissimilarities of the models in order to focus research efforts in the field more effectively.

While all of these results are promising, Reizinger et al. [233] also outlined that most theoretical results [248, 260, 261] rely on very strict assumptions over the nature of the DGP to invert. Validating this, Rusak et al. [246] showcased contradictory empirical results with regard to theoretical predictions and numerical simulations derived from such strong assumptions. In practice, models with increased identifiability appear to exhibit lower downstream performance. This is reminiscent of the issues faced by previous works attempting to explain the success of contrastive learning (CL) performance [258, 259], i.e., that the underlying assumptions do not reflect the effectiveness of methods used in practice.

5.3.1 Identifiably of self-supervised learning

Recent advancements bridging non-linear independent component analysis (ICA) [265] to contrastive learning are promising [233, 245, 246, 248, 260]. The idea is that, under reasonable assumptions about a data-generation process (DGP), which is comprised of an invertible non-linear map that emits observations as a function of well-defined latent factors, the minimiser of a contrastive loss coincides with recovering the original latent sources of the distribution up to admissible indeterminacies (e.g., affine transformations). However, there is a lack of consensus on what the precise articulation of the DGP should be or whether these identifiability results are useful in practice [233, 245]. Notably, Rusak et al. [246] did observe an identifiability-accuracy trade-off in practice, precisely because there was a mismatch of the expectations of the theory with the actual practical experimental setup.

Even from a classical nonlinear ICA perspective, these methods do not require any prior knowledge of the joint distribution of the observations and their latent variables. It is well known that in this case, contrastive models tend to favour the recovery of the sources that dominate the first moments of the data, to the detriment of the inversion of those that dictate higher-order variations [266, 267].

5.3.2 Identifiability theory contradicts well-established contrastive learning design principles

The idea that identifiability could be a design principle in CL implies a degree of *equivariance* with respect to transformations between representations and their latents, which is counterintuitive to principles drawn from information theory. A recurring principle is to design the pre-text task so that it imposes *invariance* over a set of possibly non-invertible transforms, so as any variation along the axes that are not desired for the downstream task is minimized [257, 258, 261, 268–270].

5.3.3 Oversimplified effect of data augmentations in identifiability theory

As outlined by Reizinger et al. [233], current DGPs used to study the identifiability of contrastive learning either do not account for, or are too simplistic with respect to, the effects of data augmentation. For instance, most DGP models assume that the instances of the observables $\mathbf{x}_i$ are conditionally independent of the labels [233, 246, 260, 261]. However, as argued by Wang et al. [259], this claim is easily refuted in practice: views of the

same instance should be more similar to one another than to views generated from other instances in the dataset. Moreover, if this conditional independence holds, there is a possibility that the learned representation easily degenerates to a trivial solution, i.e., fares no better than random chance for any downstream task [259, 271].

Empirical evidence supports this: Morningstar et al. [272] observed that the design of the data augmentations is more relevant for downstream performance than any other algorithmic choices. This sensitivity to the set of augmentations has been extensively verified in empirical studies [273–275].

Furthermore, unlike supervised learning, CL lacks a supervisory signal from ground-truth labels to correct mis-specification in its pre-text task [276]. Thus, selecting the wrong set of augmentation can be catastrophic downstream, especially if these transformations are biased towards certain classes or vary as a function of spurious or adversarial regularities in the data [275, 277].

In general, when considering the plethora of augmentations and inductive biases, general downstream performance guarantees of self-supervised methods remain a challenging open problem [278], in spite of promising developments that strive to ameliorate some of these issues [250, 252].

Instantiation of the Learning Triplet $\mathfrak{C}$. We can now map this definition to the unified learning triplet $\mathfrak{C} = (\Phi, \mathcal{F}, \mathcal{L})$ introduced in the previous chapter:

- **Data Processing (Φ):** Unlike deterministic preprocessing, here Φ accounts the stochastic transformation $\tau \sim p_{\langle \mathcal{T} \rangle}$.
- **Hypothesis Space ($\mathcal{F}$):** This corresponds to the class of similarity functions $\mathcal{U}$ (after discarding the projector head).
- **Loss Function ($\mathcal{L}$):** The loss is instantiated as the functional $\mathcal{L}_g$ (Equation 5), which enforces a metric depending on the “strength” of $\langle \mathcal{T} \rangle$.

5.4 Experiments

5.4.1 Numerical simulations

The outcomes of these experiments are presented in Table 5.1, and we immediately found a striking pattern: *identifiability and best possible linear probe accuracy are negatively correlated.*

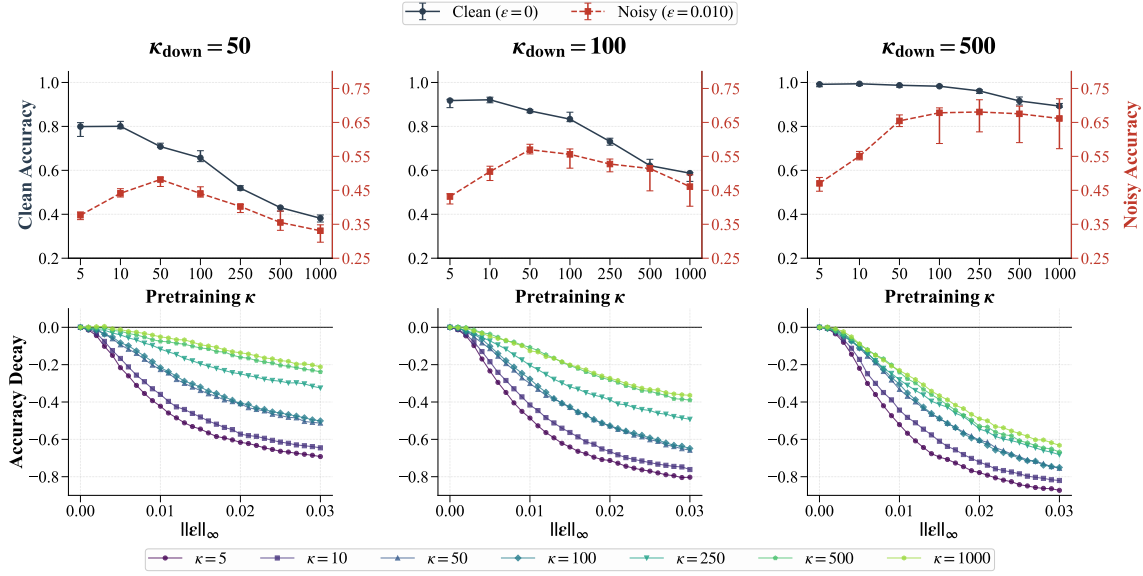


FIGURE 5.2: Downstream accuracy pairs of $(\kappa_{\text{pre}}, \kappa_{\text{down}})$. Models with higher identifiability scores (Table 5.1) tend to be very accurate on the clean validation set. However, there are no free lunches, so they are sensitive to small perturbations.

5.4.1.1 Inverting the wrong DGP leads to instability

TABLE 5.1: Supervised linear classification performance on $\mathbf{z}_i \sim p(\mathbf{z}|\mathbf{z}_c, \kappa)$ and identifiability score of $f \circ G_\kappa$. $p(\mathbf{z}|\mathbf{z}_c, \kappa)$ is a von Mises-Fisher distribution around $\mathbf{z}_c$ with concentration parameter κ . **Green** indicates the best performance; **red** indicates the worst. Mean and standard deviation computed over 10 random seeds on validation sets unseen during training.

κ	Acc. $\phi(\mathbf{z}_i, c_i)$	$R^2(f(G_\kappa(\mathbf{z}_i)), \mathbf{z}_i)$
5	0.10 $\pm$ 0.01	0.99 $\pm$ 0.00
10	0.25 $\pm$ 0.01	0.99 $\pm$ 0.00
50	0.83 $\pm$ 0.02	0.91 $\pm$ 0.00
100	0.95 $\pm$ 0.01	0.89 $\pm$ 0.00
250	0.99 $\pm$ 0.00	0.89 $\pm$ 0.00
500	1.00 $\pm$ 0.00	0.88 $\pm$ 0.00
1000	1.00 $\pm$ 0.00	0.85 $\pm$ 0.00

In light of the previous results, our theory suggests that inverting G in the high-concentration κ regime may be problematic. Note that κ is a surrogate for $\mathcal{T}$ -connectedness and, in order to be linearly separable, the classes need to be nicely separated in the sphere (see Figure 5.3).

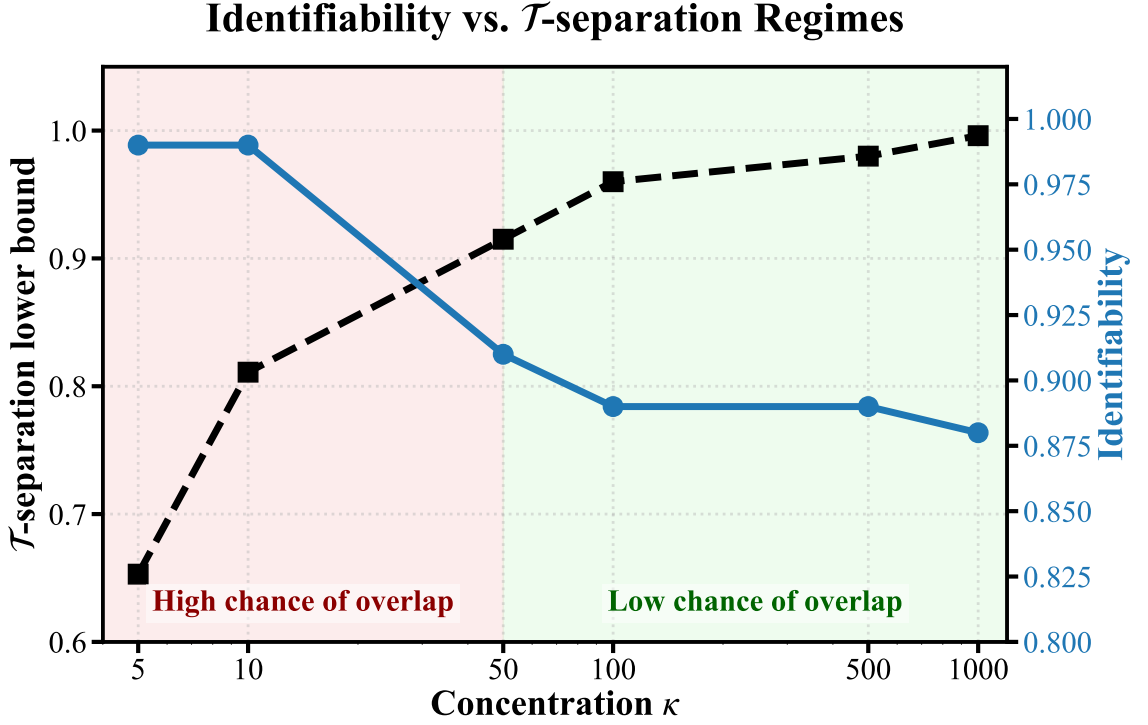


FIGURE 5.3: Relationship between Identifiability with an empirical lower bound for class separation. Since samples are drawn by first selecting $\mathbf{z}_c$, we compute the average cosine similarity between $\mathbf{z}_i$ and $\mathbf{z}_{c_i}$.

We slightly adapt the isotropic DGP proposed by [248] to show that, even when the augmentations induce $\mathcal{T}$ -connected classes that align with the ground truth of interest to us, it will necessarily come at the cost of instability (i.e., either very low accuracy or very low robustness). It can be expressed as a special case of our DGP (Definition 5.5), with the following specific conditions:

1. $Z = \mathbb{S}^{d-1}$ and $X = \mathbb{R}^d$.
2. $|I| = 1000$, $|C| = 100$, $d = 5$.
3. $\mathbf{z}_c$ are uniformly distributed in $\mathbb{S}^{d-1}$ for all $c \in C$.
4. The latents are generated by $p(\mathbf{z}) = \frac{1}{|C|} \sum_{c \in C} p(\mathbf{z}|\mathbf{z}_c, \kappa)$, where p is a von Mises-Fisher distribution [?] with concentration κ .
5. G is a non-linear, invertible, well-conditioned linear map so that $\mathbf{x}_i = G_\kappa(\mathbf{z}_i)$, with $\mathbf{z}_i \in \mathbb{S}^{d-1} \sim p(\mathbf{z}|\mathbf{z}_c, \kappa)$.
6. There are training and downstream generators, $G_{\kappa_{\text{pre}}}, G_{\kappa_{\text{down}}}$, and they differ solely in the concentration parameter controlling the distribution of $\mathbf{z}$

Note that κ shares a strong connection with the value δ used to denote the diameter of the $\mathcal{T}$ -connected classes.

By the data processing inequality [279], given that the generator is a deterministic function, we can control the uncertainty of the position of each instance around its semantical direction. The concentration parameter thus acts as a surrogate for how aggressive the augmentations imposed during training are. It also dictates how concentrated, i.e., separated, the downstream observations generated by $G_{\kappa_{\text{down}}}$ can be. The divergence between the values of κ_{pre} and κ_{down} allows one to mimic how good our guess with respect to the adequate data-augmentations was.

The generator is a fixed invertible multilayer perceptron (MLP) [248], mapping input $\mathbf{z} \in \mathbb{S}^{d-1}$ to $\mathbb{R}^d$. We train $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$, another MLP with rectified linear unit activations and linear, ℓ_2 normalized projector, $W \in \mathbb{R}^{|I| \times d}$, by minimization of $\mathcal{L}_{\text{PID}}$ (see Equation 5.4) via gradient-descent.

5.4.1.2 On the relation of identifiability and best possible downstream performance

We first see how these functions invert $G_{\kappa_{\text{pre}}}$, by measuring the R^2 score [280] of $f(G_{\kappa_{\text{pre}}}(\mathbf{z}))$ with respect to $\mathbf{z} \sim p(\mathbf{z}|\kappa_{\text{pre}})$.

To further approximate these experiments to a realistic scenario, we evaluate downstream performance on a validation set generated with $G_{\kappa_{\text{down}}}$. We selected regimes where moderate to very high classification accuracy was possible: $\kappa_{\text{down}} \in \{50, 100, 500\}$. We also perturb observations with ℓ_∞ noise to assess the models' stability. One of these two possible outcomes is expected, according to our theory:

- Highly identifiable models will not be accurate in the downstream task with high accuracy
- Highly identifiable models can be accurate, but this will imply that the classifier is unstable, i.e., sensitive to small perturbations.

The results in Figure 5.2 are surprising in some sense: not only are the models trained on pretext tasks that stray away from the true value of κ_{pre} accurate on $G_{\kappa_{\text{down}}}$, but without perturbations, their performance exceeds where $\kappa_{\text{pre}} = \kappa_{\text{down}}$. Our theory predicts that, in this case, the classifier becomes unstable, and this is exactly what we observed.

We also verified that the relative drop in downstream performance reveals the expected ordering of these models Table 5.1), i.e., to be ordered by their value κ_{pre} .

TABLE 5.2: Effect of label-flip probability on regression performance and identifiability. A precision of 4 decimal places was used extraordinarily to make the trade-off between robustness and identifiability clearer. Note that κ is the same downstream.

p_{flip}	Downstream R^2				Identifiability R^2	
	$\epsilon = 0$		$\epsilon = 0.03$		SL	SSL
	SL	SSL	SL	SSL		
0.0	0.9256	0.9244	0.2289	0.2605	0.9931	0.9824
0.1	0.9285	0.9152	0.2572	0.7325	0.9920	0.9602
0.2	0.9308	0.9013	0.3873	0.8076	0.9910	0.9455
0.3	0.9291	0.8894	0.5094	0.8024	0.9869	0.9248
0.4	0.9234	0.8885	0.7188	0.8267	0.9783	0.9252
0.5	0.9179	0.8628	0.8008	0.8575	0.9642	0.8819

5.4.2 Label flipping augmentations have a regularising effect

Thus far, we have not considered the case of label flipping data augmentations $\tau \in \langle \mathcal{T} \rangle$, since they are non-invertible by their very own nature. However, they are unquestionably effective in practice: random crops and cutouts are crucial in most self-supervised learning methods, and some authors claim they eliminate the need to craft domain-specific augmentations [281]. Our claim is not as extreme; we posit that such augmentations can have a regularising effect similar to that of SGD noise [282, 283].

To test this connection with SGD-noise, we evaluate performance on a regression task $(f(\mathbf{x}_i), \mathbf{z}_{c_i})$, while simulating their effect numerically by flipping the index label randomly using i.i.d. Bernoulli trials for each instance. We also train a supervised model in which we flip $\mathbf{z}_{c_i}$ to another $\mathbf{z}_{c_j}$ using the same criterion. The results are displayed in Table 5.2.

Not only is an increase in robustness observed in both cases, but a trade-off with instability is also observed. The trade-off varies in proportion to what we expect from the free lunch theorems we set forth.

Possible theoretical argument. Write the CL problem in its ID form and assume that $\langle \mathcal{T} \rangle$ is perfectly aligned with the downstream class $\mathbf{z}_c \in \mathcal{C}$:

$$\min_u \mathcal{L}(u) = \mathbb{E}_{i,\tau} \left[- \sum_{i \in I} \log \frac{\exp(u(\tau(\mathbf{x}_i), i))}{\sum_{j \in I} \exp(u(\tau(\mathbf{x}_i), j))} \right]. \quad (5.9)$$

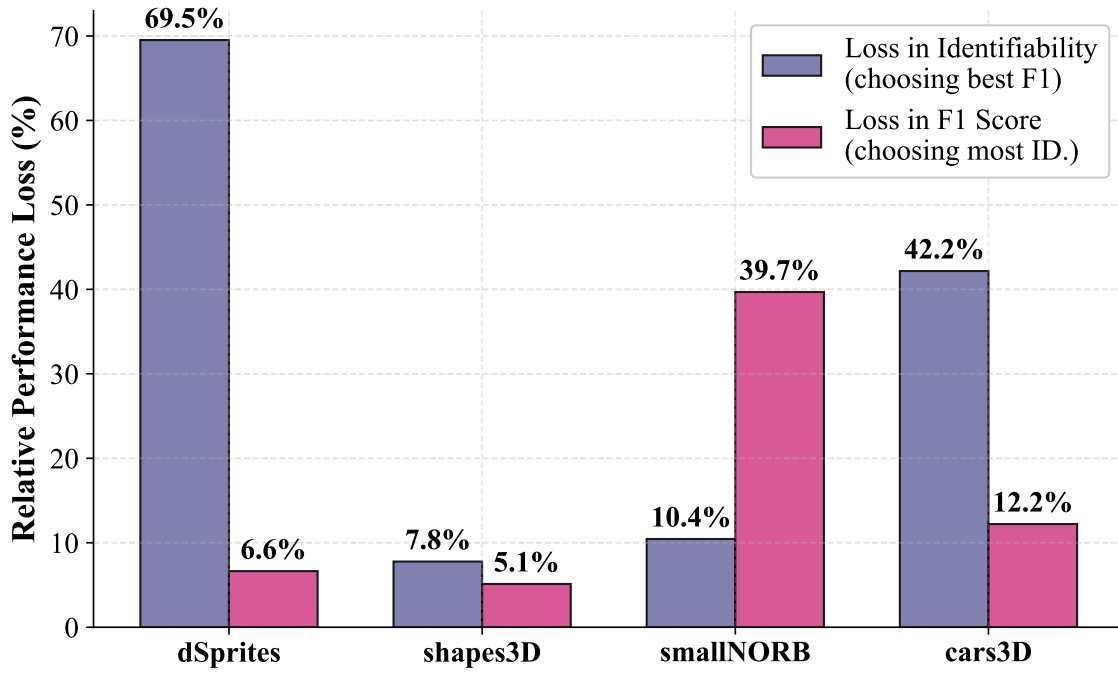


FIGURE 5.4: identifiably-stability (downstream) trade-off. As predicted by Theorem 5.9, one at the extremes, unconstrained CL learns solutions that manifest the trade-off, as a consequence of the interplay of $\mathcal{T}$ on Z and X .

Then adding τ to $\mathcal{T}$ so that $i \mapsto j$ with $i \neq j$ and $j \in I$ is equivalent to inducing stronger SGD-noise regularisation during gradient descent [282]. Assume τ is such that it flips each index label according to a Bernoulli trial t with probability $p \leq \frac{1}{2}$ of assigning to another index. Then by supposing that we want to solve $\mathcal{L}(\theta) = \frac{1}{|I|} \sum_{i \in I} \frac{1}{2} (f_i(\theta) - c_i)^2$. Suppose one is observing the k -th mini-batch during training so that $B_k \subset I$, then:

$$\mathcal{L}(\theta) = \frac{1}{|B_k|} \sum_{j \in B_k} \frac{1}{2} \left(u_j(\theta) - j - t(j) \right)^2$$

so that $\theta_{k+1} = \theta_k - \eta \nabla \mathcal{L}(\theta_k)$. Since the classes are $\mathcal{T}$ -connected, by the law of large numbers, flipping the instances corresponds to flipping the classes with probability $\sigma = \frac{p|C|}{|I|}$ of instances swapped to the wrong class. In the numerical experiments, the distribution is also isotropic [248]. The resulting long-running effect will be similar to the negative force theorised by Jha et al. [263], and this sets the necessary conditions to state the problem in a way that all the SGD noise results follow from [282].

5.4.3 DisLib benchmark

Evaluating the interplay between identifiably and downstream performance is particularly challenging, since one scarcely has access to the true generating factors of nature

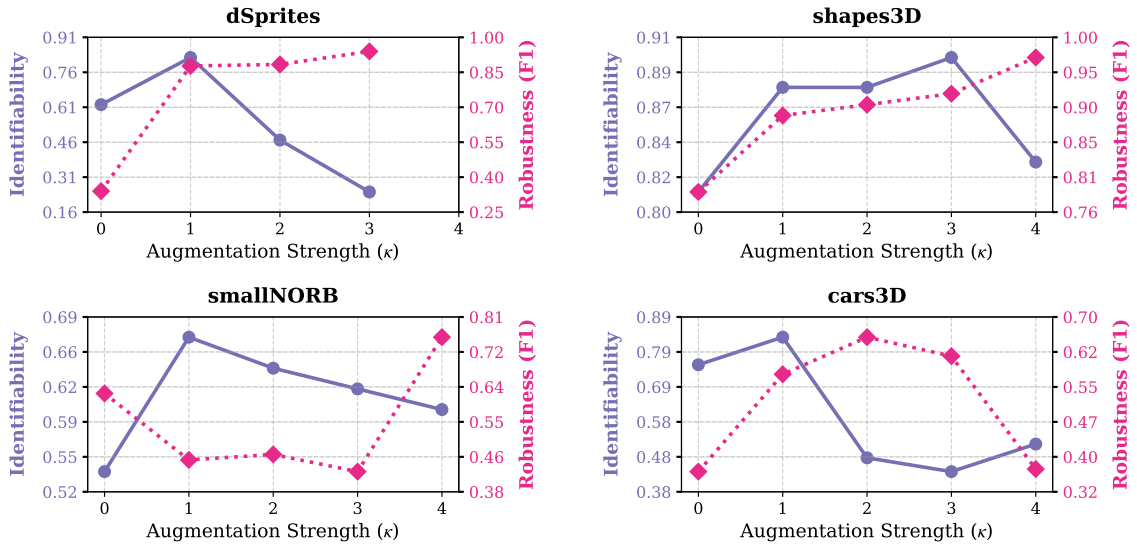


FIGURE 5.5: F1 measure between the MCC of the latents in the clean and noisy datasets versus the MCC of the remainder of the latent factors.

[246].

We follow [248] and use the DisLib benchmark [284] because it provides a set of orthogonal latent factors that completely determine the rendered image. The benchmark comprises four tasks, each with its own set of unique latents and images. The main regularity across the benchmarks is that a single latent variable always dictates the observed object’s shape, akin to $\mathbf{z}_c$ in our notation. The relationship between this factor’s prediction and the complete recovery of the factors is the main focus of this section.

Our f will be a ResNet18 optimized under the DIET framework using the original suggested hyperparameters for training [250] for 100 epochs.

We abuse previous notation and denote κ as the “strength” of the augmentation applied to each image. Since the category is defined as the shape of highly symmetrical objects across datasets, we designed our experiments to leverage a tuple of augmentations with an increasingly expansive set of compositions of transformations that impose geometrical invariances as well as color-jittering inspired by SimCLR style augmentations [268] (see description of Table 5.3).

We then predict the mean Pearson correlation coefficient between each latent. For each dataset, we use the mean Pearson correlation (MCC) of the latent variables that do not control the category as a surrogate for an identifiable score. For each run, the validation is conducted on both a clean and a corrupted version of the same holdout set. We chose ℓ_∞ noise for this purpose. The results are collected in Table 5.3 (the complete version

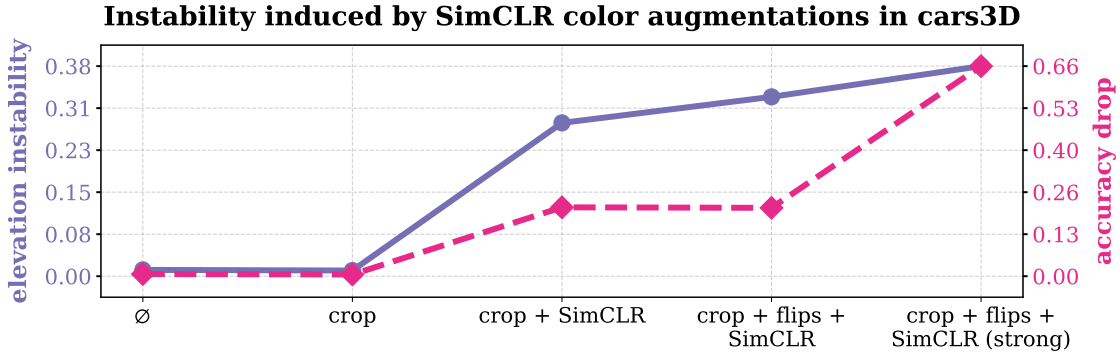


FIGURE 5.6: SimCLR color augmentations induce very high Pearson correlation coefficient scores on the validation set with respect to the latent **elevation**. However, it varies drastically in the noisy dataset, correlating with the drop in the coefficient controlling the shape.

of the Pearson correlation score of each latent can be found on 5.3 in the supplementary material).

We remark that [248] studied the identifiability of the factors of this dataset with supervision of $\mathbf{z}_c$, without any augmentations or noise. Our experiments are unsupervised.

5.4.3.1 There is a trade-off between identifiability and downstream performance

The overall trend was that the downstream task’s performance increased proportionally to the “strength” of the augmentations used during training. Note that this stands in contrast with the experimental results using supervision with the latent factor of shape in [248]. As recently predicted by [276], unsupervised methods are highly dependent on augmentation to achieve competent performance on a plausible downstream task.

In Figure 5.4, we can observe the extent of the trade-off between identifiability and downstream performance. We computed the F1 measure of the correlation between the clean and noisy datasets to gauge robustness (see also Figure 5.5). With the exception of shapes3D, a very strong and asymmetrical concession is required. In fact, even with shapes3D, one still has to sacrifice 7.5% of the average identifiable score to achieve the best downstream performance.

5.4.3.2 Miss-specified augmentations lead to instability

Finally, we identified a particular case in the results for the cars3D dataset (see Table 5.3), where noisy accuracy plummeted when color jittering augmentation was added. We decided to inspect the impact of this augmentation specifically for this dataset (see

Figure 5.6). We observed a strong association between downstream performance and the variability of the elevation latent variable in the noisy dataset, suggesting potential entanglement. Note that this is entirely possible, since Theorem 5.7 bounds the Lipschitz constant of the inversion below proportional to $\|\mathbf{z}_i - \mathbf{z}_j\|_Z$. From Remark 5.10, if two disjoint classes are $\mathcal{T}$ -connected, then not only is the map unstable, but the representation also entangles the latent factors.

5.4.4 Relationship with Equivariant-SSL

We simulated the experiments of [285] and observed that our predictions hold for their experiments in the CIFAR-10 dataset (see Table 5.4). Indeed, Wang et al. [285] shows a connection between equivariant SSL and instance discrimination, although it is grounded solely in an informational-theoretic perspective and lacks identifiable guarantees to date.

The objective is relatively simple, instead of predicting the correct instance index for each sample of the dataset, one has to predict the parameters of $\tau \sim p_{\langle \mathcal{T} \rangle}$ that generated the positive views. Note that, as claimed by [285], instance discrimination is an extreme case of equivariant SSL, since the action space of instance discrimination (which is in practice at least in the order of thousands) under τ yields a much larger action space than predicting the parameters of τ (which is in the order of dozens of parameters). Adequate formalisation of this statement is delegated to future work.

TABLE 5.3: Identifiability and category mean correlation coefficient (MCC) on clean and noisy holdout validation sets for disLib [284]. Results are reported as mean $\pm$ standard deviation by selecting distinct train/validation splits using three random seeds. The augmentation strength $\kappa = i$ denotes the first i transformations from the ordered set ($\{\text{id}_X\}, \{\text{crop}, \text{x-flip}\}, \{\text{crop}\}, \{\text{crop}, \text{x-y-flip}\}, \{\text{crop}, \text{x-y-flip}, \{\text{color-jitter}\}\}$). Identifiability is the mean Pearson correlation coefficient over all latent factors excluding the category. Noise was selected to be 8/55 for RGB images, and 4/255, otherwise. **Green** indicates the best performance; **red** indicates the worst.

Dataset	Latent MCC	$\kappa = 0$	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$
dSprites	scale	0.70 ± 0.03	0.94 ± 0.00	0.94 ± 0.01	0.95 ± 0.00	-
	orientation	0.10 ± 0.02	0.46 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	-
	x coordinate	0.85 ± 0.00	0.94 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	-
	y coordinate	0.84 ± 0.01	0.94 ± 0.01	0.94 ± 0.00	0.05 ± 0.00	-
	identifiability	0.62 ± 0.02	0.82 ± 0.01	0.47 ± 0.01	0.25 ± 0.00	-
	category (clean)	0.34 ± 0.00	0.93 ± 0.00	0.92 ± 0.01	0.96 ± 0.00	-
	category (noisy)	0.34 ± 0.00	0.83 ± 0.04	0.85 ± 0.05	0.92 ± 0.01	-
shapes3D	floor color	0.84 ± 0.05	0.86 ± 0.02	0.88 ± 0.03	0.91 ± 0.01	0.79 ± 0.02
	wall color	0.81 ± 0.00	0.87 ± 0.02	0.89 ± 0.00	0.91 ± 0.01	0.92 ± 0.01
	object color	0.75 ± 0.03	0.83 ± 0.02	0.83 ± 0.03	0.83 ± 0.07	0.84 ± 0.00
	size	0.81 ± 0.02	0.93 ± 0.00	0.92 ± 0.00	0.94 ± 0.01	0.98 ± 0.00
	azimuth	0.82 ± 0.03	0.91 ± 0.02	0.88 ± 0.02	0.91 ± 0.01	0.62 ± 0.06
	identifiability	0.81 ± 0.01	0.88 ± 0.01	0.88 ± 0.01	0.90 ± 0.01	0.83 ± 0.01
	category (clean)	0.79 ± 0.01	0.90 ± 0.02	0.92 ± 0.01	0.93 ± 0.00	0.99 ± 0.00
	category (noisy)	0.78 ± 0.01	0.88 ± 0.01	0.89 ± 0.04	0.91 ± 0.01	0.95 ± 0.01
smallNORB	instance	0.33 ± 0.01	0.50 ± 0.01	0.52 ± 0.01	0.49 ± 0.01	0.58 ± 0.01
	elevation	0.68 ± 0.02	0.85 ± 0.01	0.85 ± 0.01	0.83 ± 0.01	0.76 ± 0.01
	rotation	0.31 ± 0.01	0.42 ± 0.02	0.28 ± 0.01	0.25 ± 0.01	0.35 ± 0.01
	lighting	0.83 ± 0.02	0.92 ± 0.00	0.92 ± 0.00	0.92 ± 0.00	0.72 ± 0.01
	identifiability	0.54 ± 0.02	0.67 ± 0.01	0.64 ± 0.01	0.62 ± 0.01	0.60 ± 0.01
	category (clean)	0.62 ± 0.01	0.80 ± 0.01	0.82 ± 0.01	0.82 ± 0.01	0.89 ± 0.00
	category (noisy)	0.62 ± 0.01	0.32 ± 0.10	0.33 ± 0.09	0.29 ± 0.13	0.66 ± 0.03
cars3D*	elevation	0.74 ± 0.01	0.87 ± 0.01	0.84 ± 0.02	0.77 ± 0.02	0.89 ± 0.01
	azimuth	0.76 ± 0.00	0.79 ± 0.01	0.11 ± 0.02	0.11 ± 0.03	0.14 ± 0.02
	identifiability	0.75 ± 0.01	0.83 ± 0.01	0.48 ± 0.02	0.44 ± 0.03	0.52 ± 0.02
	category (clean)	0.37 ± 0.01	0.58 ± 0.02	0.66 ± 0.00	0.64 ± 0.03	0.59 ± 0.02
	category (noisy)	0.36 ± 0.01	0.57 ± 0.02	0.65 ± 0.00	0.59 ± 0.04	0.27 ± 0.08

TABLE 5.4: CIFAR-10 replication of Equivariant SSL using equivariant ResNet-18 backbone [285]. Results are reported for baseline ($\kappa = 0$), random + crop ($\kappa = 1$), and SimCLR ($\kappa = 2$). The % improvement represents the relative gain in F1 score over the $\kappa = 0$ baseline.

Metrics	$\kappa = 0$	$\kappa = 1$	$\kappa = 2$
identifiability	1.00	1.00	0.93
category (clean)	0.72	0.79	0.83
category (noisy)	0.65	0.73	0.79
robustness (F1)	0.68	0.76	0.81
% improve in F1	-	10.85%	18.10%

5.5 Chapter conclusion

In this chapter, we have effectively extended the “No Free Lunch” theorems to the domain of contrastive learning. We showed that identifiability—the recovery of the true data-generating process (DGP)—does not automatically translate into superior downstream performance.

Our theoretical analysis also considers the impact of data augmentations in a more realistic way. While strong augmentations are required to concentrate observations around meaningful semantic directions for classification, they predictably compromise the extent to which the DGP can be inverted. A more subtle finding concerns cases where models achieve the highest accuracy on a validation set while simultaneously being highly competent at inverting the DGP. Interestingly, this precise condition dictates their vulnerability to perturbations, leading to an *identifiably-stability trade-off*.

These findings suggest that the community should exercise caution when using identifiability to gauge representation quality. Future work should focus on formally characterising the “sweet spot” between these seemingly opposing forces – identifiability and downstream performance. The way this trade-off unfolds in the intermediary layers of the neural network is also still unanswered. The extent to which different inductive biases and downstream tasks (e.g., dense prediction) may stir this trade-off is also an interesting open question.

5.5.1 Outcomes

1. 1 submission to a top-tier AI conference, ICLR.
2. Several invited talks in seminars: CentraleSupélec, CEA, ELLIS Max Planck Institute, Telecom Paris.

Chapter 6

Conclusion

The goal of this Thesis was to investigate the extent to which deep learning models can benefit from additional prior knowledge in downstream tasks where domain knowledge is available and to test this hypothesis in a set of specific applications in biomedical signal and image processing. The central argument was formulated through the *learning triplet*,

$$\mathfrak{C} = (\Phi, \mathcal{F}, \mathcal{L}),$$

where Φ denotes the data-processing map, $\mathcal{F}$ the hypothesis space induced by the model architecture and parametrisation, and $\mathcal{L}$ the objective function that determines which properties of the data-to-target relation are optimised. Under this framework, prior knowledge is not understood as an external addition to learning, but as an intervention on one or more components of $\mathfrak{C}$. The overarching conclusion of this Thesis is that downstream gains obtained from domain knowledge are never free: they are achieved by imposing additional assumptions. When those assumptions are well matched to the data-generating process and the downstream task, they can improve feasibility, performance, stability, or interpretability. When they are mismatched, they can introduce brittleness, instability, or loss of generality.

In Chapter 2, we asked:

RQ1: To what extent can domain knowledge, expressed through the processing of the input signal, simplify the learning problem and make a clinically relevant downstream task feasible?

This question was answered affirmatively, but conditionally. We showed that when one has strong prior knowledge about where the minimal clinically relevant region of

interest lies, Φ can be designed to discard large portions of the input signal. In the application studied, upper gastrointestinal endoscopy, this reduced the model’s exposure to nuisance variables such as instrumentation, bodily fluids, illumination changes, irrelevant anatomical structures, and artifacts of the heads-up display. For the specific task of gastric intestinal metaplasia, this intervention was motivated by the self-similarity hypothesis: if the pathological pattern is sufficiently self-similar, then local patches may preserve the discriminative information required for classification. In this setting, domain knowledge expressed through Φ simplified the learning problem and made the downstream task feasible with relatively standard deep learning machinery.

However, the answer to RQ1 is not that aggressive preprocessing is generally desirable. Rather, the chapter shows that manipulating Φ transfers complexity away from the model and into the signal-processing stage. This makes the learning problem easier only if the discarded information is genuinely irrelevant to the task. In practice, such an assumption depends on acquisition standardization, on the reliability of the region-of-interest selection process, and on the validity of the self-similarity hypothesis for the pathology under study. Thus, Chapter 2 answered RQ1 by showing that domain knowledge can simplify learning through input processing, but also that this simplification is paid for by a stronger dependence on acquisition conditions and task-specific assumptions.

Chapter 3 moved from preprocessing to structural inductive bias and asked:

RQ2: Can geometrical priors derived from fractal and multifractal analysis improve the segmentation of pathological patterns whose visual manifestation is self-seeding, cascading, and scale-dependent?

The answer was also affirmative, although more nuanced. We argued that irregular pathological phenomena with self-seeding, cascading, and scale-dependent structure are common across medical imaging modalities and are poorly described by ordinary Euclidean intuitions about shape. For such phenomena, the relevant discriminative information may not be captured by local appearance alone, but by how image structure behaves across scales. Chapter 3 therefore articulated fractal and multifractal geometry as structural priors on $\mathcal{F}$, rather than merely as handcrafted descriptors.

This led to the proposal of monofractal and multifractal recalibration mechanisms, which use attention to reweight convolutional responses according to scaling-law information. The experiments showed that these recalibration strategies can improve segmentation performance over the baseline and can outperform other higher-order channel

attention functions. The findings also suggested that different imaging modalities and datasets benefit from different levels of geometrical modelling: in some cases, a global fractal descriptor was sufficient; in others, the multifractal formulation was necessary to capture richer textural heterogeneity. Therefore, RQ2 was answered by showing that geometrical priors can improve the modelling of irregular pathological objects, provided that the prior is aligned with the type of scale-dependent structure present in the images.

At the same time, the results caution against an overly simplistic interpretation of attention as merely selecting informative channels. The evidence did not support the view that successful attention modules simply suppress non-informative channels. Instead, performance appeared to depend on a more delicate recalibration of feature responses, with balanced instance variability being preferable to either rigid or excessively unstable excitation behaviour. Thus, the contribution of Chapter 3 is not only that fractal priors can improve performance, but that the way such priors interact with $\mathcal{F}$ is itself a non-trivial modelling choice.

In Chapter 4, we turned to temporal structure and asked:

RQ3: What is the impact of encoding domain knowledge as a statistical sequential prior, expressed simultaneously in the model structure and in the loss function, for learning to segment rhythmic, quasi-periodic biomedical signals?

This chapter showed that, for sequential biomedical data with known temporal regularities, prior knowledge may need to intervene in the entirety of $\mathcal{C}$. In the heart-cycle application, Φ was required to partition the raw signal into finite events with adequate temporal resolution, preserving the sequential structure needed for the Markov assumption to be meaningful. The hypothesis space $\mathcal{F}$ was then defined as a hybrid model combining a discriminative neural network with a generative Markovian component. Finally, $\mathcal{L}$ determined how the discriminative and generative components interacted, through pseudo-complete likelihood, maximum mutual information, and marginal-likelihood-based fine-tuning objectives.

The answer to RQ3 is therefore that statistical sequential priors can make learning more coherent with the underlying physiology, especially when the observed process is rhythmic, quasi-periodic, and state-based. By explicitly modeling transition structure, the Markov-based Neural Network does not treat the heart sound as an arbitrary sequence of independent observations. Instead, it constrains predictions according to the expected progression of physiological states. This can improve temporal consistency and enables a

principled form of unsupervised test-time adaptation through the marginal likelihood of a target observation.

However, Chapter 4 also showed that stronger hybridization is not automatically superior. The interaction between $\mathcal{F}$ and $\mathcal{L}$ determines the practical behavior of the model. Some configurations improved sensitivity, others favoured positive predictive value, and the fully hybrid variants introduced additional computational and optimisation costs. In particular, the results suggested that the most theoretically complete model is not necessarily the most efficient or robust configuration for the downstream task. Thus, RQ3 was answered by showing that statistical sequential priors are useful when they match the temporal regime of the data, and our analysis suggests that this benefit depends on the chosen point in the trade-off between discriminative performance, generalization, and computational cost.

Finally, Chapter 5 studied the case in which models are often assumed to require less explicit domain knowledge and asked:

RQ4: Are self-supervised representations truly task-agnostic, or does their apparent generality conceal instability when transferred to unseen in-domain datasets?

The answer was negative: self-supervised representations are not truly task-agnostic in the unconstrained sense often implied by their use as general-purpose feature extractors. Chapter 5 re-framed contrastive learning as an inverse problem with limited prior knowledge, where data augmentation acts as a regulariser. Under this view, the pretext task is not neutral. Positive and negative pair generation implicitly defines Φ , and therefore imposes assumptions about which variations should be preserved and which should be discarded.

The theoretical and empirical results showed that contrastive learning obeys a no-free-lunch principle. Increasing the degree to which a representation identifies the latent factors of the data-generating process does not necessarily improve downstream classification performance. Conversely, augmentations that improve downstream discrimination may do so precisely by destroying or suppressing information that would be needed for identifiability. This creates a trade-off between explaining the data-generating process and optimising performance for a particular downstream task. In this sense, self-supervised models that appear to simultaneously explain and discriminate the data are not exempt

from the constraints described throughout this Thesis. Their constraints are simply less explicit.

This result closes the argument initiated in Chapter 2. Even when domain knowledge is not deliberately encoded through a biomedical model, geometrical prior, or statistical sequential assumption, the learning configuration still induces a bias. In contrastive learning, this bias is largely inherited from both the training objective, and the implicit invariances of the pretext task. Therefore, Chapter 5 answered RQ4 by showing that apparent generality can conceal instability: the more variation a representation attempts to capture, the more fragile it may become under perturbations or transfer to unseen in-domain data; conversely, the more invariant it becomes, the more information it necessarily discards.

Taken together, the four research questions may be jointly framed in the context of a certain (deep learning) configuration $\mathcal{C}$ that determines how the data is processed, what functions are admissible, and what objective is optimised. Chapter 2 showed that intervening on Φ can make a difficult clinical task feasible by suppressing nuisance regions of the input data. Chapter 3 showed that intervening on $\mathcal{F}$ through geometrical inductive bias can improve the modeling of irregular, scale-dependent pathological structures. Chapter 4 showed that intervening jointly on $\mathcal{F}$ and $\mathcal{L}$ is possible for encoding temporal domain knowledge in a hybrid statistical-deep learning model. Chapter 5 showed that even methods presented as broadly task-agnostic still rely on implicit assumptions, and that these assumptions induce measurable trade-offs between identifiability, stability, and downstream performance.

The central contribution of this Thesis is therefore not a claim that domain knowledge always improves deep learning. Rather, it is a framework for understanding how, where, and at what cost domain knowledge constrains learning. The most successful models in this Thesis were not those with the fewest assumptions, but those whose assumptions were best aligned with the structure of the downstream task. This also explains the limitations observed across chapters: aggressive preprocessing can become brittle under acquisition shift; geometrical priors may help only they can express the nature of the relevant signal ; Sequential priors are valuable only when the temporal regime is well matched; and self-supervised objectives can fail when the pretext task conflicts with the downstream task.

This stance aligns with recent critiques of representation learning and identifiability theory [233, 245]. As deep learning increasingly shifts toward large-scale self-supervised

and foundation-model paradigms, future work should adopt theoretical models that more closely reflect how these systems are trained and used in practice. Benchmark performance remains useful, but it is weaker than principled guarantees that explain when a model should work, when it should fail, and which assumptions are responsible for that behavior.

A promising future direction is to connect domain-specific model-based methods with data-driven frontier models, such as Large Language Models and other foundation models. This may be especially important in critical domains that operate in large data regimes but still require interpretability, reliability, and certification, such as intensive care, astrophysics, and diplomatic or military decision-making. In such settings, the lesson of this Thesis is particularly relevant: scale may expand what models can represent, but it does not remove the need to understand the assumptions under which learning takes place.

Appendix A

Deep Learning with prior knowledge

A.1 Extended literature review

A.1.1 Classical computer vision

In the early 2010s, the work of Riaz and Coimbra pioneered research in the CH and NBI modalities by introducing classical computer vision techniques to identify and characterise lesions in UGI tissue [286–289]. Riaz *et al* [287] developed a Support Vector Machine (SVM) classifier tailored towards classifying chromoendoscopy and NBI images from Gabor Texton [290] feature vectors. They collected 176 chromoendoscopy and 221 NBI images in the Portuguese Institute of Oncology, Porto (IPO). The annotations were defined by the same clinician who performed the endoscopy. The choice of Gabor filters was motivated by their ability to minimise uncertainty in the spatial and frequency domains and their analogy to mammalian visual cortex cells. A filter bank was designed using self-similar Gabor filters using [?] to minimise the number of redundant filters in the dictionary. The multi-resolution coverage of the filters within the dictionary is exploited, since the tissues in the human body have important multi-scale characteristics [291]. The filter responses are then clustered using *K*-means due to the redundancy still present in the filter bank responses. A SVM is used to classify the frames in 3 (CH) or 4 (NBI) classes. The authors report an accuracy of 82.3% accuracy and a True Positive Rate (TPR) of 0.90 for CH images. However, this algorithm outputs a high number of false negatives, which is especially worrisome when diseased tissue is misclassified as healthy. Concerning NBI, an accuracy of 83.9% and TPR of 0.88 was achieved, with the same difficulties of assigning false negatives to non-healthy observations.

Coimbra *et al.* [286], ablated normalised cuts [292], mean-shift [293] and variation level sets [294] as the unsupervised segmentation step with the end-goal of improving a binary static lesion detection classifier. They found that mean shift gave the best overall results for both NBI and CH modalities in the task of tissue characterisation. However, artificial improvements were still subpar compared with the SVM under *perfect segmentation* (using its own ground-truth masks as a preprocessing step). However, in the pre-deep learning era of computer vision, segmentation was crucial to improving downstream task performance, as evidenced by the SVM classifier achieving the worst results when this step was skipped in the experiments. In the same spirit, but now tailored towards achieving the best segmentation results, Riaz *et al.* [289] expanded the segmentation step work in 2014, by introducing a set of highly discriminant features based on texture, creaseness and colour on a set of 142 CH and 224 NBI images. They used a divisive clustering approach based on the Graph Cut methodology, *NCUT*, to mitigate oversegmentation and noise sensitivity [292, 295]. The image was represented as a weighted adjacency matrix of an undirected graph. The weight or affinity was calculated for each vertex using the Hadamard product of the edge, creaseness, and colour responses. Edge information was extracted using multi-scale cues [296] and adapted to a single spatial response following [297]. To address creaseness, multi-local level set extrinsic curvature with enhancement by structure tensor (MLSEC-ST) [298], a partial derivative tensor using all directional combinations of the smoothed image gradient. These features prioritise the global direction of the image gradient instead of local textures. Should the eigenvalues of these measures of directional variation be similar, one could conclude that the smooth neighbourhood of a pixel is isotropic [289]. Finally, the colour features were extracted from the $\mathbf{l}^* \mathbf{u}^* \mathbf{v}$ colour spaces in order to capture human perceptual changes to colour in the image using algebraic distances [299, 300]. The chrominance components ($\mathbf{u}$ and $\mathbf{v}$) were discarded since colour variation is not a major property in GI data, and thus colour information was extracted solely from the luminance channel ($\mathbf{l}$). The output segmentations were compared to 28 images annotated by 3 experts, showing an average *DICE* coefficient (DSC) of 0.73 for CH and 0.8 for NBI frames. Comparing the agreement with the *NCUT* algorithm, the average DSC agreement 0.69 for CH and 0.77 for NBI modality. It should be noted that these were not external test images, and they do not indicate the robustness of this parametrisation to unseen data.

The work of Coimbra and Riaz suffers from a lack of modern error analysis and cross-validation schemes. Moreover, it is crucial to have a histological golden standard, especially when collecting data from pathologies with high mortality rates such as GC. Notwithstanding, the fact the color and texture descriptors studied by Coimbra and Riaz are very relevant for a fundamental understanding of the characterization of UGI tissue, and their observation of the importance of multi-scale self-similarity nature of the *lumen* is pivotal as inspiration for the upcoming work in the classical computer vision line as well as the methodologies that will be proposed in the modern deep learning context of this thesis.

Myaki *et al.* [301] developed a Scale Invariant Feature Transform (SIFT) Bag of features framework with a Support Vector Machine (SVM) classifier. They collected Magnifying Endoscopy Blue Laser Images (ME-BLI; a modality similar to NBI) images from 100 consecutive EGCs from 95 patients in the Hiroshima University Hospital. All lesions underwent submucosal dissection or surgical resection, and histopathological results served as the gold standard. The train set was built under the same criteria for a different patient group. They generated a higher sample volume by cutting each non-magnified frame into blocks of healthy and lesion tissue. The frames collected dataset contained EGC, flat or depressed small reddened lesions and normal tissue. Scale Invariant Feature Transform (SIFT) descriptors at two different scales were computed and then clustered using K-Means into 128 dimensional feature, called a *bag of visual words*. These embeddings are fed into a logistic SVM for binary classification (cancerous *versus* non-cancerous). They found that the SVM output value was highly correlated with cancer tissue. However, actionable performance metrics were not measured. Later, in 2018, Kanesaka *et al.* [14] developed the first Machine Learning algorithm able to delineate irregular microvessels on ME-NBI by applying an SVM on variation vectors of the grey level co-occurrence matrix (GLCM) [?]. An endoscopist with more than 1000 cases of diagnosis EGC under Magnifying Endoscopy with NBI (ME-NBI) retrospectively collected, reviewed and delineated cancerous regions from 127 consecutive cases from the Osaka International Cancer Institute database. Images were included if EGC was of the depressed type (smaller than 10 mm) and the pathology was substantiated with histopathological findings of endoscopic submucosal dissection (ESD). A clear ME-NBI frame was selected from each patient. A train set comprised of 66 EGC frames and 60 non-cancer frames and a test set with 61 EGC and 20 non-cancer frames were built. Histogram equalisation and Gaussian filtering were

applied to enhance contrast and mitigate noise in each frame. Images were portioned in 40x40 pixels, and the 8 GLCM features were extracted from each block. Afterwards, the variation vectors were computed from the mean and variance of each GLCM feature within each block. Two distinct SVMs were optimised to i) perform global binary classification (cancerous vs non-cancerous) and, should this global prediction indicate a cancer, ii) identify cancerous blocks in a finer-grained classification scheme, respectively. The first SVM measured an accuracy of 96.30%, a PPV of 98.30%, a sensitivity of 96.70% and a specificity of 95.00%. Concerning the second model, using the area-concordance at the block level, it scored an accuracy of 73.80%, PPV of 75.30% , sensitivity of 65.00% , and specificity of 80.80%. The authors point out some shortcomings, such as the risk of selection bias in the data, because they did not consider flat or elevated EGCs. Moreover, the character of the non-EGC class was not very detailed, and there is no mention of Atrophic Gastritis (AG) and Gastric Intestinal Metaplasia (GIM) being collected in this class, since they are both precursor conditions as well as highly challenging to differentiate from EGC [51, 59]. In the same year, Xiaoqi Liu *et al.* [302] proposed an unsupervised region-based active contour model for unsupervised segmentation of abnormal tissue. Rooted on the Chan-Vese framework [303], they leveraged the Hue and Saturation channels of the Hue-Saturation-Intensity (HSI) colour space to extract global colour and local texture features, that were later embedded in a same energy function minimized under a level set reformulation strategy [304]. A total of 770 ME-NBI images from 49 patients were collected in the Xinqiao Hospital during standard clinical practice. The pathologies present in the images spanned chronic atrophic AG, GIM, low-grade neoplasia, high-grade neoplasia and EGC. Each image is converted to the HSI colour space. In the hue component, a Gaussian sooth filter is used to mitigate noise. Then, based on the Chan-Vese formulations, an energy function tailored to maximise the distance between each pixel and the mean pixel values inside and outside the contours is defined. In order to capture microvascular patterns, the authors employ a pixel inhomogeneity factor (PIF) [305] with an adaptive threshold texture descriptor, where the square spatial neighbourhood of each pixel is considered for feature extraction. Small PIF values indicate that the neighbourhood differs smoothly from the centre pixel, while high ones indicate abrupt transitions such as edges. The authors use this property to describe the microstructures of the mucosa. Moreover, the pixel intensities are captured in the Saturation channel to enhance edge detection. The energy component for the local colour descriptor is calculated using

the same strategy for colour. They maximise the difference between the PIF mean value inside and outside the contour with each pixel. The global colour and local texture functional components are combined in a joint energy function regularised by the length of the contour, which is optimised using a level set reformulation approach. It should be noted that each component of this function is scaled in an *ad-hoc* fashion using fixed parameters. The authors devised 3 experiments where they measured the F1 measure and the False Positive Ratio (FPR). In the first experiment, the authors fixed the model parameters, 12 images proposed model scored an average of 0.61 F1 measure and 16.00% FPR, whereas the classical Chan-Vese model score 0.52 and 32.00%, respectively. In a second experiment, they ablated different shape priors for the initialisation of the contour and noted that their approach converged to more credible results. In the third experiment, they calculated the accuracy given different precision ratio thresholds. The proposed model was able to detect, for the lowest threshold (0.5) with an accuracy of 56.70%, while the Chan-Vese model scored 49.60%. Better results could be attained if the parameters were manually set depending on subjective visual cues in the image. Although an argument can be made that the parameter setting may be viable given sufficient operator expertise, it is far from desirable, as it suffers from the problems CADs are designed to address: the mitigation of operator variability. The type of model validation is also outdated compared to today's standards, with accuracy not being the most reliable metric, as it is sensitive to sample class imbalance. Notwithstanding, it is an interesting example of how domain knowledge can improve a well-established model when UGI data is used for the downstream task.

Appendix B

Multifractal recalibration

B.1 Additional results for Multifractal formalism

Remark B.1. The Hausdorff measures relate by (2.4 of ??):

$$\mathcal{H}^n(\mathbb{F}) = V_n(1)^{-1} \text{vol}^n(\mathbb{F}), \quad (\text{B.1})$$

where $V_n(r)$ is the volume of a n -dimensional hypersphere with radius r :

$$V_n(r) = \frac{\pi^{1/2}}{\Gamma(\frac{n}{2} + 1)} r^n$$

B.2 Properties of dimension operator

Definition B.2 (Dimension operator). Let $\dim$ define an arbitrary dimension operator over some set. We want the following properties to hold

1. Monotocity: $\mathbb{E} \subset \mathbb{F} \implies \dim \mathbb{E} \leq \dim \mathbb{F}$.
2. Stability: $\dim(\mathbb{E} \cup \mathbb{F}) = \max(\dim \mathbb{E}, \dim \mathbb{F})$.
3. Countable stability: $\cup_{i=1}^{\infty} (\dim \mathbb{F}_i) = \sup_i \dim \mathbb{F}_i$.
4. Geometric invariance: if f is an affine linear transformation, then $\dim f(\mathbb{F}) = \dim \mathbb{F}$.
5. Bi-Lipschitz invariance: if f is a bi-Lipschitz function, then $\dim f(\mathbb{F}) = \dim \mathbb{F}$.
6. Countable sets: $\dim \mathbb{F} = 0$ if $\mathbb{F}$ finite or countable.
7. Open sets: $\dim \mathbb{F} = n$ if $\mathbb{F}$ is an open subset of $\mathbb{R}^n$.

8. Smooth manifolds: $\dim \mathbb{F} = n$ if $\mathbb{F}$ is a smooth m -dimensional manifold.

Definition B.3 (Self-similar sets). A set $\mathbb{F}$ is called d -regular [306, Def. 8.8] or *self-similar* if there exist constants $c_1, c_2 > 0$ such that for any diadic cube centered on $x \in \mathbb{F}$, $\mathbb{B}_\delta(x)$, and all $0 < \delta < \text{diam}(\mathbb{F})$,

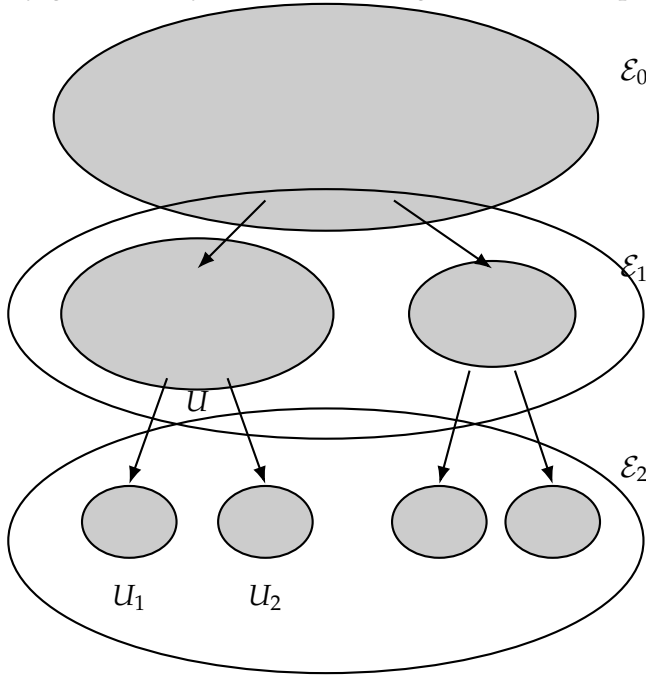
$$c_1 \delta^d \leq \mathcal{H}^d(\mathbb{F} \cap \mathbb{B}_\delta(x)) \leq c_2 \delta^d,$$

where $\mathcal{H}^d$ denotes the d -dimensional Hausdorff measure.

B.3 Repeated iterated subdivision of measures

The Lebesgue measure is deeply tied to our intuitive notion of dimension, and defines the so-called *topological dimension*. However, as we shall see later in this section, there are intricate objects that surprisingly have Lebesgue measure 0.

These troublesome objects are usually iterated in an iterative way. We begin by defining a very general way of characterizing measurable spaces in this fashion.



Definition B.4 (Repeated iterated subdivision of measure). Let $(\mu, \mathbb{I}, \mathcal{B})$ be a measurable space where $0 < \mu(\mathbb{I}) < \infty$. $\mathcal{B}$ will be a Borel σ -algebra built in such a way so that:

1. $\mathbb{B}_0 = \mathbb{I}$.
2. $\mathbb{B} \in \mathbb{B}_k = \{\mathbb{B}^1, \dots, \mathbb{B}^m\}$ so that $\mu(\mathbb{B}) = \sum_{i=1}^m \mu(\mathbb{B}^i)$.

3. Inductively, we let $\mathcal{B}_k$ to be a union of disjoint Borel sets so that each $\mathbb{B} \in \mathcal{B}_k$ is contained in one of the sets of $\mathcal{B}_{k-1}$, and contains a finite ammount of sets in $\mathcal{B}_{k+1}$.

Under these conditions we can derive a more compact notation. Let $\mathbb{I} \equiv \mathbb{B}^0$. Then denote $\mathbb{B} \in \mathcal{B}_k$ as $\mathbb{B}^w := \cup_j \{\mathbb{B}^{w+j} \in \mathcal{B}_{k+1}\}$ using a number string $r := (w_i)_{i=0}^k$, $j \in \{0, \dots, 9\}$, and $r + j := (w_1, w_2, \dots, w_k, j)$, where $+$ is a string concatenation operator.

We slightly abuse notation and denote $\mathbb{B}_k^r$ so that $\mu(\mathbb{B}_k^r) = \sum_{j=1}^{\min(r,k)} \mathbb{B}^{(w_1, \dots, w_j)}$, which loosely translates to *set $\mathbb{B}^r$ measured at resolution k* .

A representation of a possible measurable space built according to this definition may be found on Figure B.3.

Theorem B.5 (Nested set theorem). *Let $\{\mathbb{E}_k\}$ be an increasing Borel set sequence such that $\mathbb{E}_k \subset \mathbb{E}_{k+1}$, then:*

$$\mu \left(\bigcup_{k=1}^{\infty} \mathbb{E}_k \right) = \lim_{k \rightarrow \infty} \mu(\mathbb{E}_k).$$

Proof. Note that

$$\mathbb{E}_k \subset \mathbb{E}_{k+1} \implies \mathbb{E}_k = \mathbb{E}_{k+1} \cup (\mathbb{E}_{k+1} \setminus \mathbb{E}_k),$$

so:

$$\mu \left(\bigcup_{k=1}^{\infty} \mathbb{E}_k \right) = \mu(\mathbb{E}_1) + \lim_{k \rightarrow \infty} \sum_{i=1}^k (\mu(\mathbb{E}_{i+1}) - \mu(\mathbb{E}_i)) = \lim_{k \rightarrow \infty} \mu(\mathbb{E}_k).$$

□

Remark B.6. Suppose that we partition set $[0, 1]$ in diadic intervalles of length 2^{-k} , so that

$$\mathbb{B}_k^r = [r2^{-k}, (r+1)2^{-k}) \text{ s.t. } 0 \leq r \leq 2^k - 1, k = 1, 2, \dots$$

If $\mu(\mathbb{B}_k^r) = 2^{-k}$ for all r , then μ is the Lebesgue measure on $[0, 1]$.

Appendix C

Markov-based neural networks

In this appendix we will detail the experimental framework, specially the pre-processing and feature extraction steps required, for the experimental section of Chapter 4.

C.1 Experimental pipeline

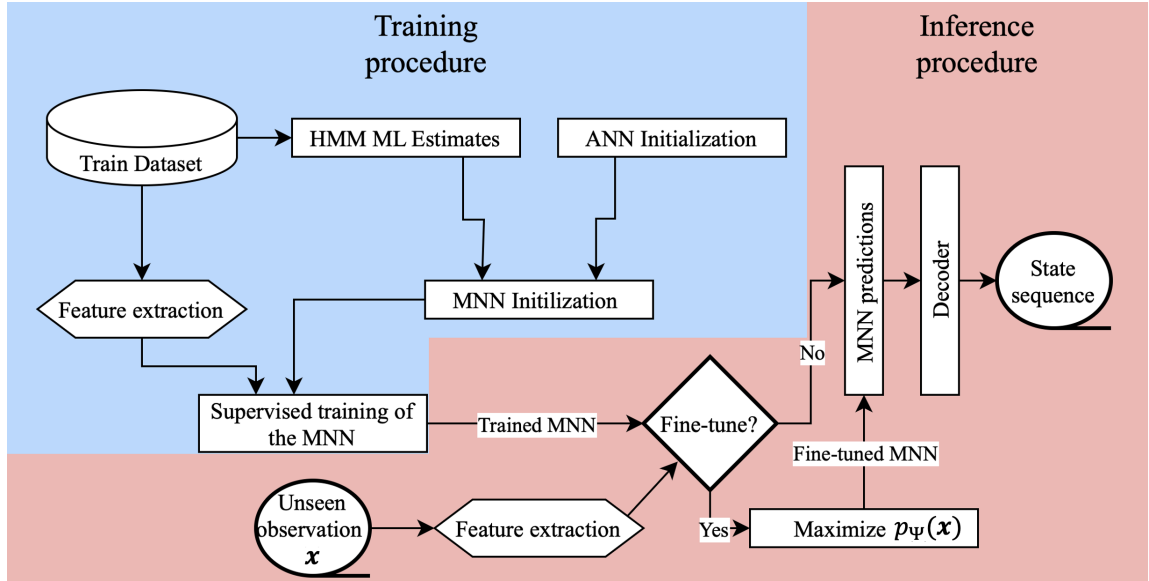


FIGURE C.1: A schematic of the experimental pipeline of Markov-based neural networks.

C.2 Pre-processing

Signals are first band-passed using a second order Butterworth filter, between [25-400] Hz (illustrated on Figure C.2). Before feature extraction, one needs also to perform additional denoising, as described in the next section.

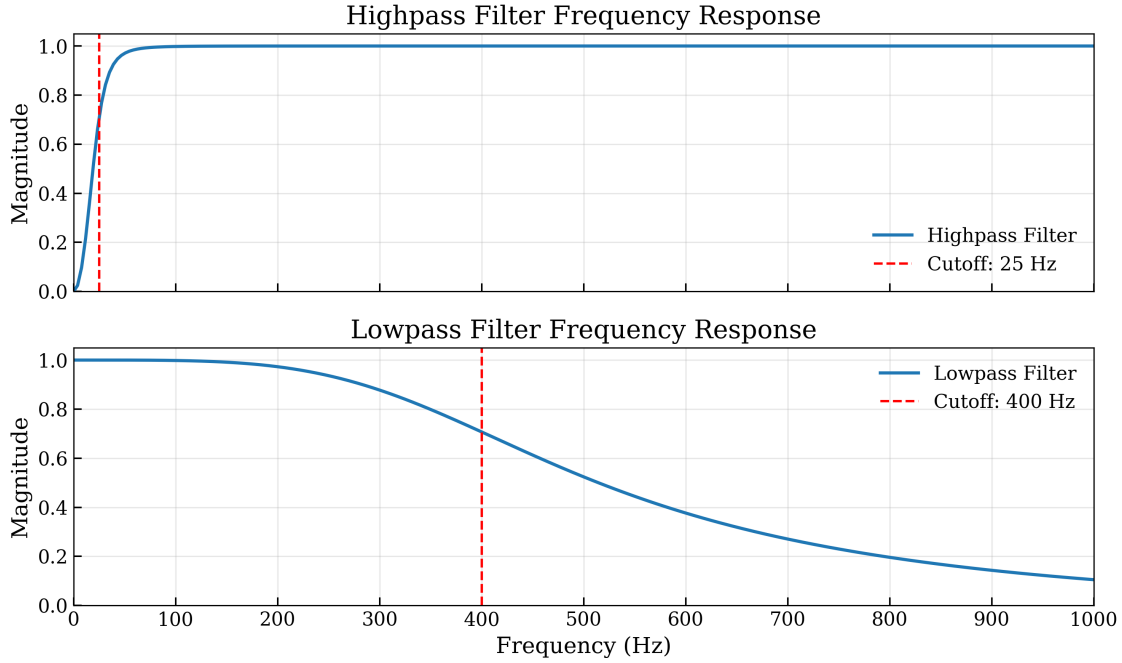


FIGURE C.2: The bandpass filter used comprised of two second-order Butterworth filters.

C.2.1 Denoising

Suppose a Source s is corrupted additively by i.i.d. Gaussian noise ϵ_t . We observe x and not s in our measurements such that:

$$x_t = s_t + \epsilon_t, t = 1, \dots, T \quad (\text{C.1})$$

If one computes its variance:

$$\text{Var} \left(z + \frac{1}{T} \sum_{t=1}^T \epsilon_j \right) = \text{Var}(s) + \text{Var} \left(\frac{1}{T} \sum_{t=1}^T \epsilon_t \right) = \text{Var}(s) + \frac{T\sigma^2}{T^2} \quad (\text{C.2})$$

one observes that the *standard deviation* of the random terms will shrink as $\frac{\sigma}{\sqrt{T}}$.

Due to the periodic nature of heart sounds and their stationarity (at least in a “short” period of time), we can think of s to be the expected waveform of the heart cycle in a recording, i.e. the “characteristic heart cycle” of a patient [307].

In practice, the analysis of a characteristic heartbeat for most downstream applications is not very useful. Even under (quasi)-stationary assumptions, there may be other phenomena present in the signal such as murmurs that may occur in all states of the heart cycles and across several frequency bands. These phenomena may also be “transient” which immediately defeats the purpose of a characteristic heartbeat.

For convenience, in the remainder of this appendix we will use the signal-processing convention $x(t)$ instead of our notation $\mathbf{x}$ to denote a stochastic process that unfolds over time.

Wavelet denoising The Wavelet decomposition will allow us to filter the original signal in an adaptive way. It allows one to make a trade-off between frequency and time resolution as a function of scale. The Discrete Wavelet Transform is given by:

$$\text{DWT}_x^\psi(m, n) = \sum_t x(t) \psi_{m,n}(t) = \sum_t x(t) \psi^* \left(\frac{t - n2^m}{2^m} \right) \quad (\text{C.3})$$

using a precision of 2^{-m} , or m bits. ψ^* is the so-called *mother* Wavelet function.

The multiresolution decomposition of $x(t)$ can be carried out by selecting two mutually orthogonal basis-functions ϕ, ψ so that the decomposition with resolution 2^{-k} is such that:

$$x_k(t) = \underbrace{\sum_n \text{DWT}_x^\psi(k, n)}_{\text{Approximation } A_k \text{ at level } k} + \underbrace{\sum_{m=1}^k \sum_n \text{DWT}_x^\psi(m, n)}_{\text{Details } D_m \text{ from levels 1 to } k} \quad (\text{C.4})$$

In this heart sound data the Haar Wavelet decomposition with $k = 5$ levels (i.e., a floating point precision of 5) is typically used. Figure C.3 as representative example of one of a heart recording decomposition.

Universal thresholding Since the DWT provides a multiresolution decomposition of the signal, we can use our prior knowledge that most of the information is concentrated in the low frequencies to mitigate noise in an adaptive fashion.

Suppose that the detail coefficients at the finest scale are distributed according to a standard Gaussian scaled by σ . Then, using extreme value theory [308], the largest detail coefficient is:

$$\max_{i=1, \dots, k} |D_i| \approx O(\sigma \sqrt{2 \log k}) \quad (\text{C.5})$$

The higher the frequency, the more sensitive to small perturbations our estimates will be, hence we are looking for a robust estimator of σ . Typically one uses the median absolute deviation estimate of the Gaussian:

$$\sigma \approx \frac{\text{median}(|D|)}{0.6745}. \quad (\text{C.6})$$

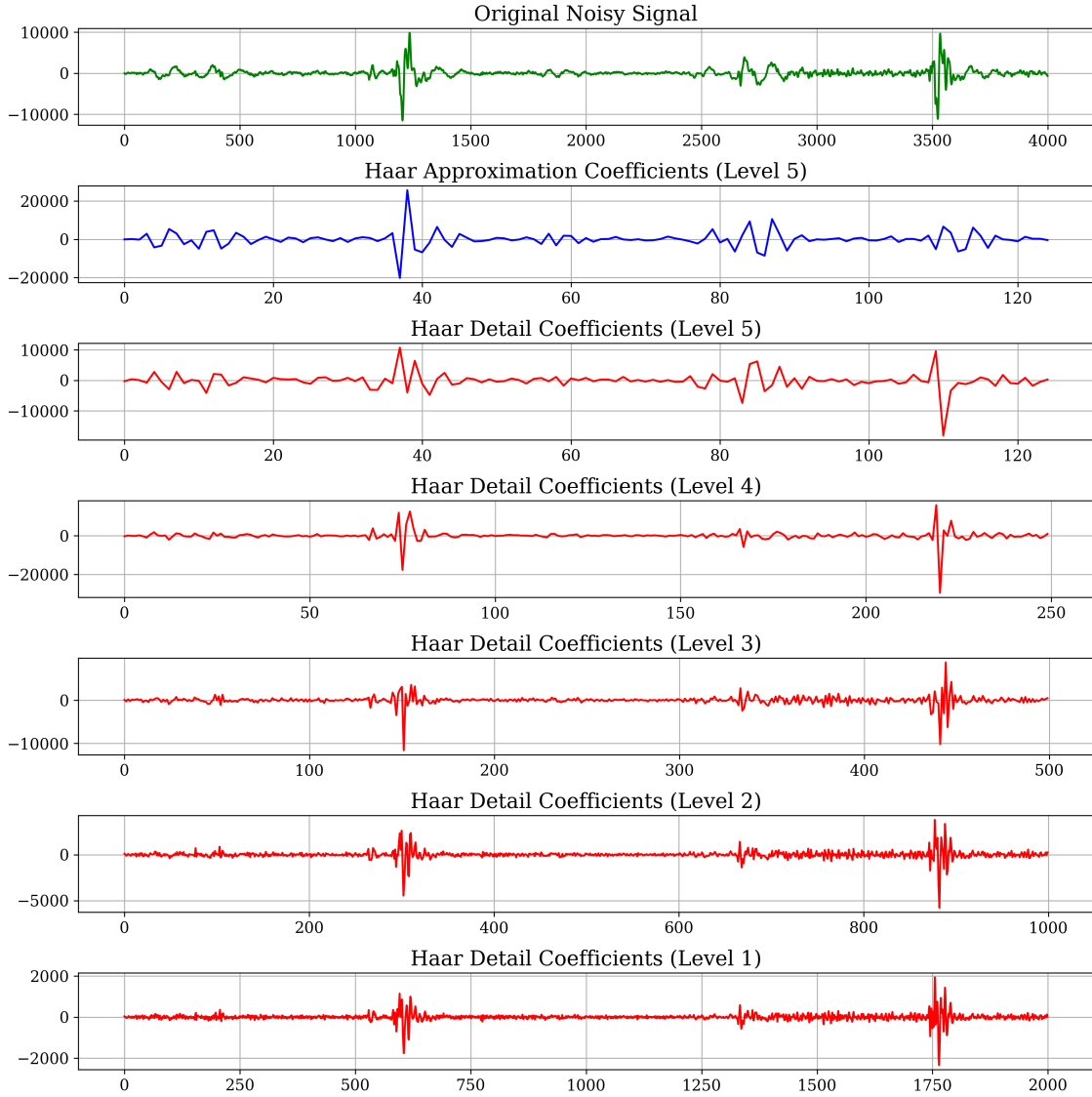


FIGURE C.3: A slice of a recording of the Circor'22 dataset [212] and its decomposition using 5 levels using the Haar mother wavelet.

Now, we only need to define the thresholding function. We will implement the following soft-thresholding:

$$\hat{D}_j = \text{sign}(D_j) \cdot \max(|D_j| - \lambda, 0), \quad (\text{C.7})$$

where $\lambda = \sigma \sqrt{2 \log k}$. This effectively zeroes out coefficients smaller or equal to λ and shifts the remaining D_j s towards 0 by λ . An example of the action of the thresholding at different detail resolutions can be found for a real heart recording in Figure C.4.

An example of a recording before and after denoising may be found on Figure C.5

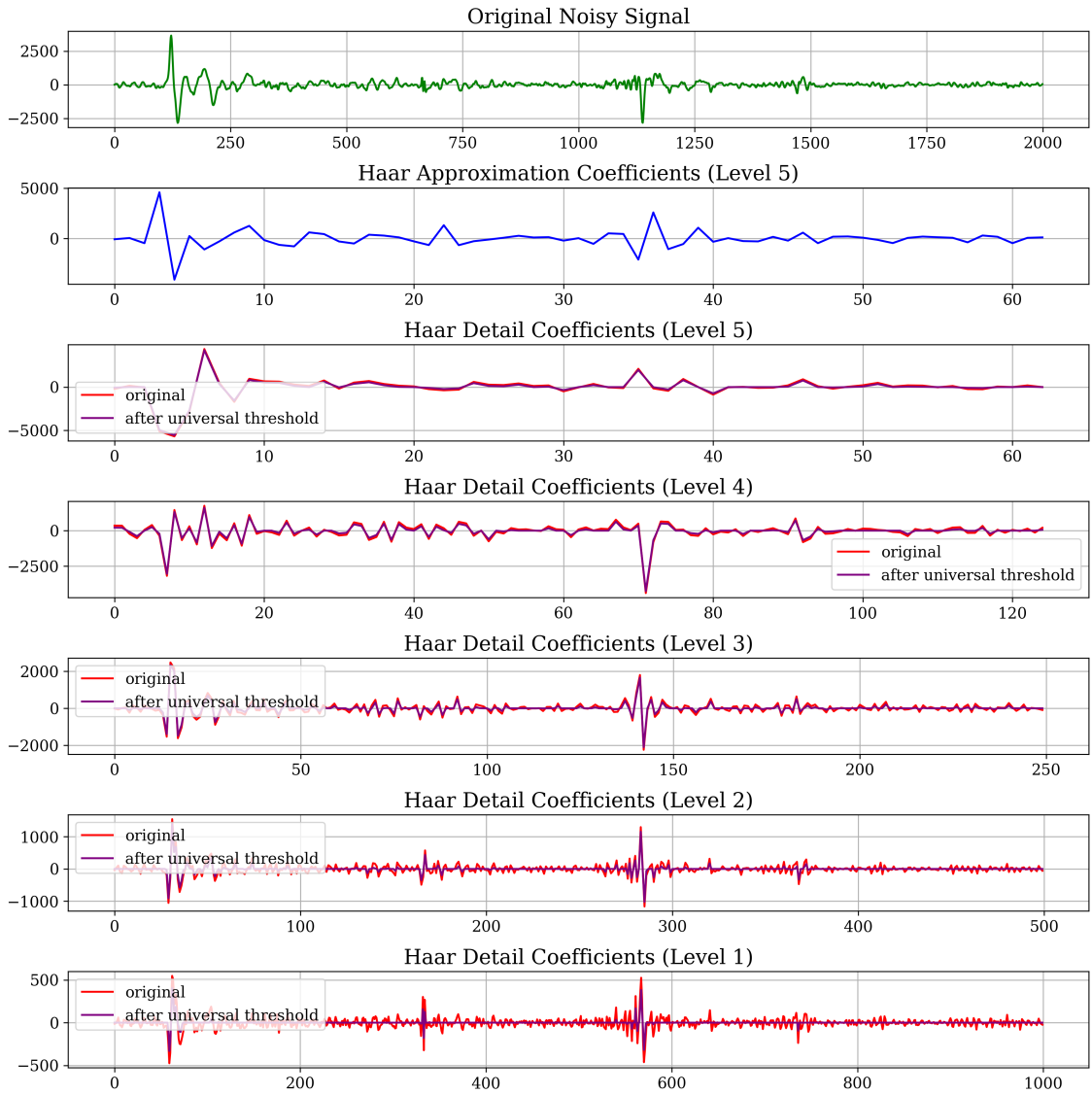


FIGURE C.4: A slice of a *noisy* recording of the Circor'22 dataset [212] and its decomposition using 5 levels using the Haar mother wavelet. Note how most noise concentrates at lower detail levels, and is effectively mitigated suppressed through adaptive threshold.

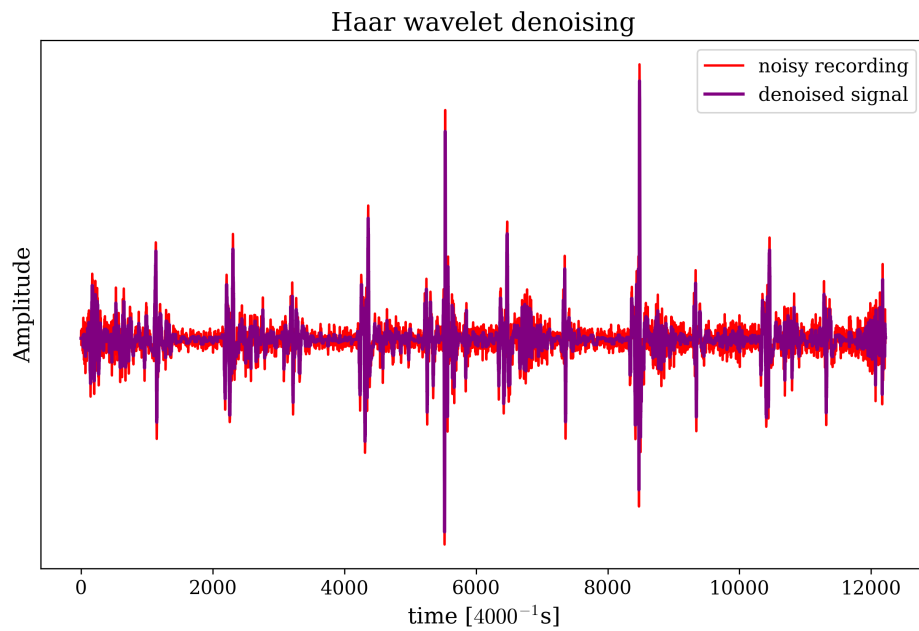


FIGURE C.5: An example of a *noisy* recording of the Circor'22 dataset [212] and the output of the 5 level Haar wavelet denoiser.

C.3 Feature extraction

This section elaborates on the feature extraction prescribed by Springer et al. [206] and Renna et al. [207].

Our downstream task pertains to segmenting to “loud” events (S1 and S2 sounds) that are then followed by more “quiet” onsets (systole and diastole). If we consider our Markov-based neural networks, given that our Markovian prior imposes an order throughout time, our emission distribution is mostly concerned with the energy of the signal at a given instant.

For this purpose, instead of feeding the denoised signal directly, one instead characterizes the energy of the signal. For heart sound segmentation, the feature extraction step will produce a set of (correlated) envelopograms for each signal with the goal of making S1/S2 onsets more salient in comparison to the ensuing systolic/diastolic intervals.

C.3.1 Hilbert Envelopogram

The inner product between sin and cos over $[0, 2\pi]$ is:

$$\int_0^{2\pi} \sin(\omega t) \cos(\omega t) dt = 0$$

These functions are mutually orthogonal, so their inner product with $x(t)$ will capture different information. Specifically, note that $\sin(\omega t) = \cos(\omega t + \frac{\pi}{2})$, so this phase shift operation allows inspection of the variance of signal across two orthogonal directions.

The Hilbert transform is given by:

$$x(t)^* = \frac{1}{\pi} \int x(\tau) \cdot \frac{1}{t - \tau} d\tau$$

which is equivalent to the convolution $x * h(t)$, where $h(t) = \frac{1}{\pi t}$ is the *Hilbert kernel*. The above function can be seen as an inner product in the frequency domain due to the Convolution theorem – the Fourier transform of $h(t)$ is $-j \operatorname{sgn} \omega$, $j \in \mathbb{C}$, so this kernel effectively shifts the phase by $\pm 90^\circ$ of each frequency component of x .

The *analytic* signal is thus given by:

$$\hat{x}(t) = x(t) + jx^*(t)$$

where x and x^* can be seen as two basis functions, and $\hat{x}$ the linear combination $[1, j]^\top$ of them. The Hilbert envelope can be computed as the following Euclidean norm in the complex plane:

$$H(t) = |\hat{x}(t)| := \sqrt{x(t)^2 + x^*(t)^2}.$$

This is also called the *magnitude* or *energy* of the analytic signal.

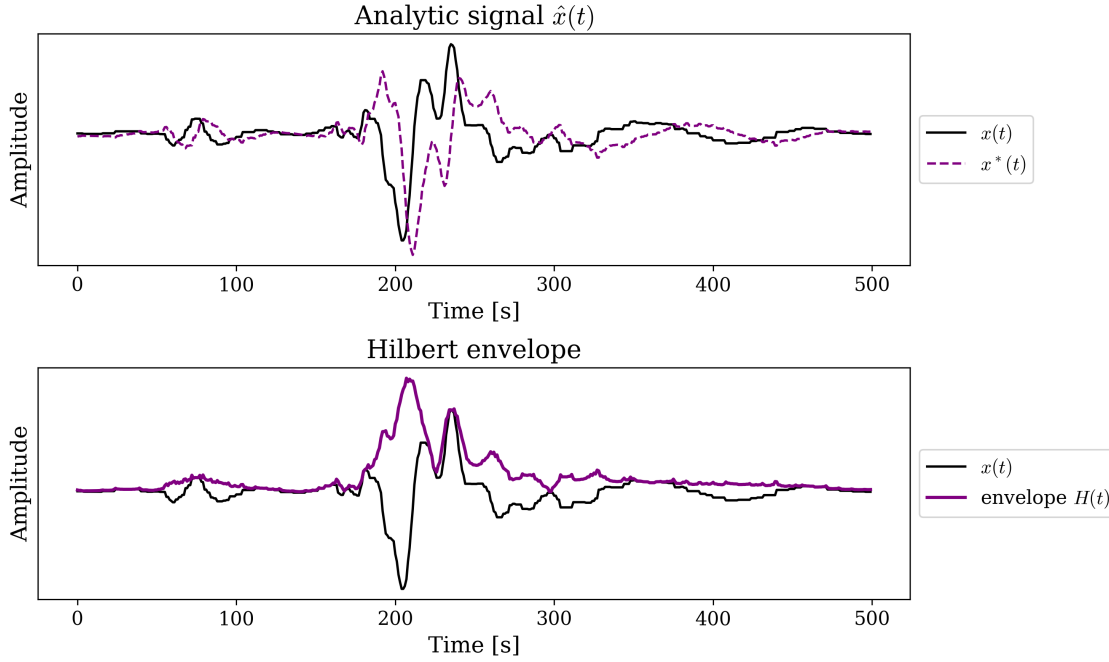


FIGURE C.6: (top) A portion of a *noisy* recording of the Circor'22 dataset [212] and its representation as the analytical signal, comprised by the real $x(t)$ and imaginary $x^*(t)$ parts. (b) The resulting envelopogram $H(t)$.

The Hilbert envelope might be very irregular due to high frequency oscillations. Suppose you assume that the analytic signal can be described as

$$\hat{x}(t) = A(t) \cdot c(t),$$

i.e., a slowly varying amplitude function $A(t)$ and a *carrier* high frequency signal $c(t)$. The log-transform makes this relationship additive, so one can effectively use low-pass filters (LPFs) to remove or mitigate high-frequency noise in the carriers.

$$H_{\text{Homo}}(t) = \exp \left(\text{LPF} \{ \log H(t) \} \right).$$

$H_{\text{Homo}}(t)$ is thus a smoothed version of the original Hilbert envelopogram (see Figure C.7).

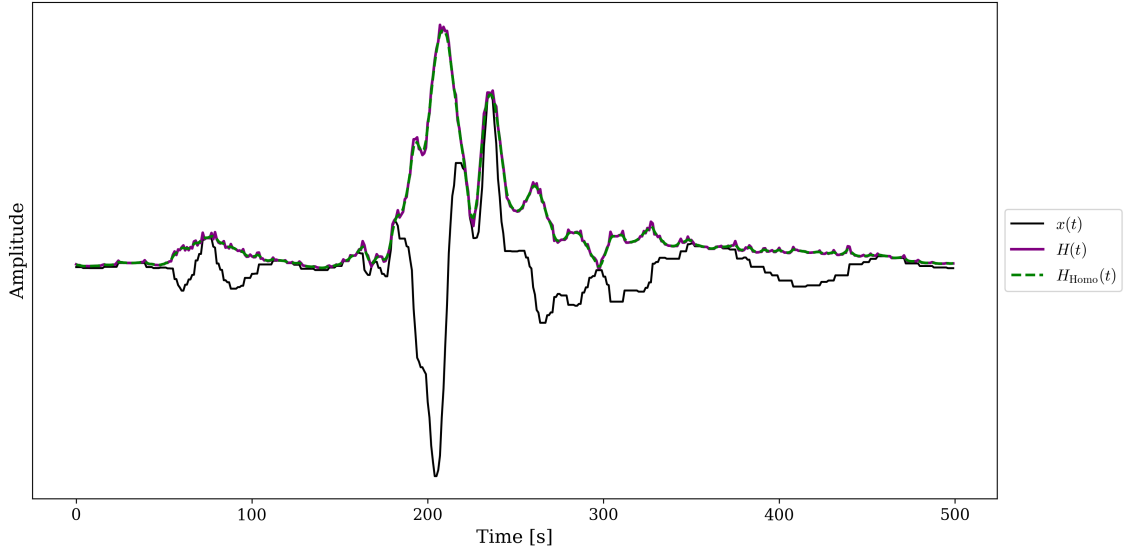


FIGURE C.7: The homomorphic envelopment is a smooth version of the Hilbert envelope, attained by applying a low pass filter to $H(t)$. In this case, an order 2 low-pass Butterworth filter at 640 Hz as prescribed by Springer et al. [206].

C.3.2 Frequency space envelopes

Suppose we compute the discrete-time Fourier transform (DTFT) as:

$$x_{\text{freq}}(\omega, t, t') = \sum_{t=0}^{L-1} x(t' + t)w(t) \exp\{-j\omega t\},$$

where $w(t)$ is a binary mask of size T with support on a subset of $\mathbb{N}^+$ of size L . The power spectral density (PSD) envelopment is given by:

$$H_{\text{PSD}}(t) = \frac{1}{L} |x_{\text{freq}}(\omega, t, t')|^2.$$

Naturally the spectrogram assumes a stationary frequency distribution across windows. This may be adequate for stationary phenomena, but if one wants a more general envelopment that accounts for transient phenomena, or even more complicated signals, one can describe the energy using the DWT instead.

Suppose we compute the DWT subject to universal thresholding as in C.2.1, then the energy of the signal can be taken as the sum of the norm of the k detail coefficients:

$$H_{\text{DWT}} = \sum_k |D_j|.$$

Note that one can use the analytic signal using the Hilbert transform $\hat{x}(t)$ as input instead of the original recording.

C.3.3 Full set of envelopes

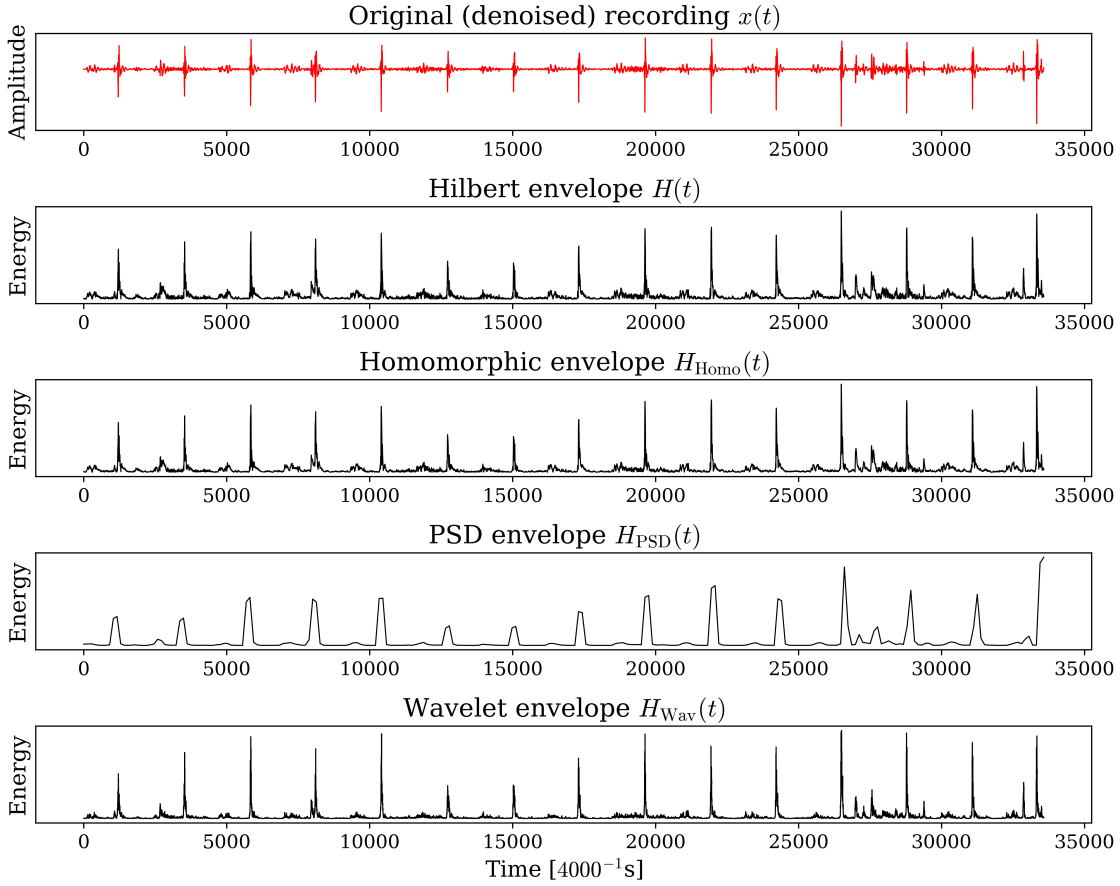


FIGURE C.8: (Top row) A denoised recording from the CirCor’22 dataset. (Remaining rows) The extracted envelopes used as input features for our models. Note how the envelopes simplify the morphology of the signal for the purposes of our task: one expects the energy to be more concentrated during the onset of the two fundamental sounds (S1/S2), which become exaggerated in the envelopes.

In our experiments we collect the envelopes for each of the recordings, defining the following tensor:

$$\mathbf{x}_{\text{env}} = [H^{(i)}(t), H_{\text{Homo}}^{(i)}(t), H_{\text{PSD}}^{(i)}(t), H_{\text{Wav}}^{(i)}(t)]_{i=1}^{|\mathcal{D}|}.$$

In Figure C.8 an example of this feature set is given for a real phonocardiogram.

C.4 Markov likelihood computation limitations

The forward-backward algorithm requires computation of the most likely sequences for each $t = 1, \dots, T^i$. This means that this computation is repeated iteratively (i.e., sequentially), as illustrated below in our implementation.


```

1  ...
   def _complete_likelihood(self, y_true, y_pred):
3      T = tf.shape(y_true)[0]
       sequence = tf.argmax(y_true, axis=1)
5      emissions = tf.reduce_max(y_true * y_pred, axis=1)
       joint = (self.p_states[sequence[0]] * emissions[0]) / self.p_states[
sequence[0]]
7      c_t = 1 / tf.reduce_sum(joint)
       joint = joint * c_t
9      C_t = tf.math.log(c_t)
       for t in range(1, T):
11         joint = joint * self.trans_mat[sequence[t - 1], sequence[t]] * (
emissions[t] / self.p_states[sequence[t]])
         c_t = 1 / tf.reduce_sum(joint)
13         joint = joint * c_t
         C_t += tf.math.log(c_t)
15     return -C_t

17     def _forward_backward(self, y_pred):
        """
19         Calculates the likelihood P(0) using a scaled forward algorithm.
         Bayes theorem is used given DNN output = P(S_t|O_t) to model the
         emissions.
21         That is:  $p(O_t|S_t) = [P(S_t|O_t)P(O_t)]/P(S)$ 

23         Scaled implementation based on section 5.A of:
         Rabiner, L.R., 1989. A tutorial on hidden Markov models and selected
         applications in speech recognition.
25         Proceedings of the IEEE, 77(2), pp.257-286.

27         Parameters
         -----
29         y_pred : tf.Tensor
           The tensor of the output layer of a DNN

31         Returns
         -----
33         tf.Variable

```

```
35         The likelihood  $P(0)$  given the markov chain and neural network output
        """
37         T = tf.shape(y_pred)[0]
        y_pred_ = y_pred / self.p_states
39         alpha = self.p_states * y_pred_[0, :]
        c_t = 1 / tf.reduce_sum(alpha)
41         alpha *= c_t
        C = tf.math.log(c_t)
43         for t in range(1, T):
            alpha_stay = alpha * tf.linalg.diag_part(self.trans_mat)
45             alpha_go = tf.roll(alpha, shift=1, axis=0) * tf.linalg.diag_part(tf.
roll(self.trans_mat, shift=1, axis=0))
            alpha = (alpha_stay + alpha_go) * y_pred_[t, :]
47             c_t = 1 / tf.reduce_sum(alpha)
            alpha *= c_t
49             C += tf.math.log(c_t)

        return -C
```

Appendix D

No free lunches for self-supervised learning

D.1 Additional visualizations

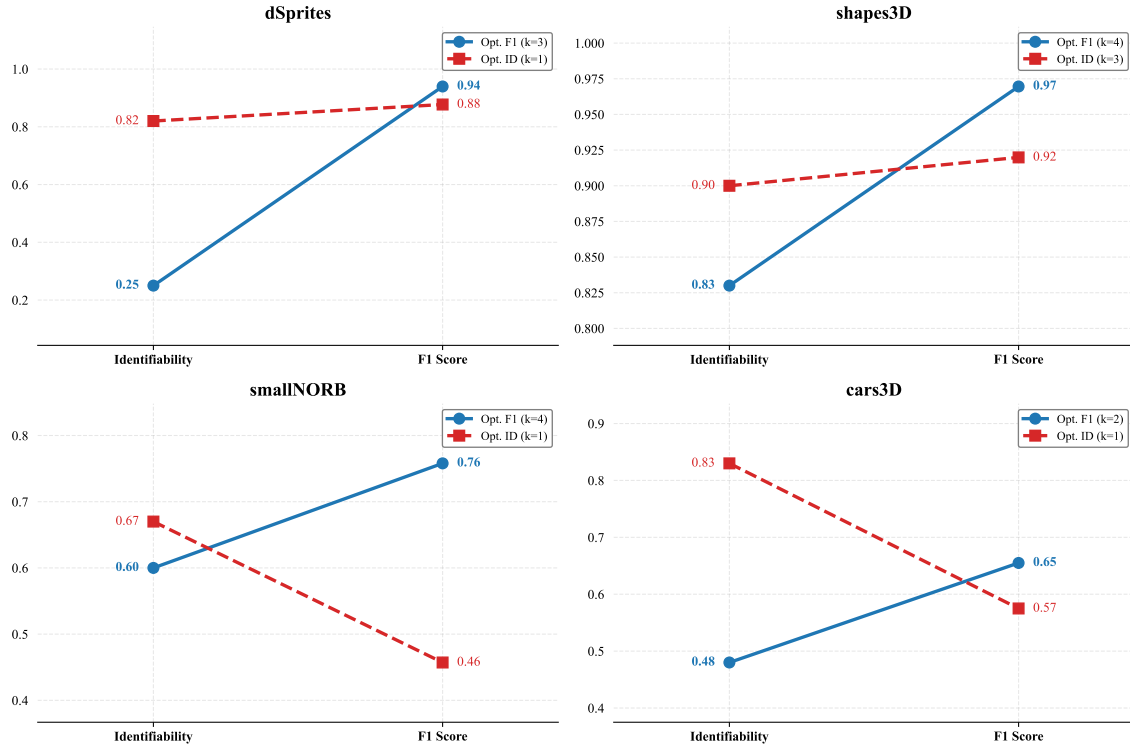


FIGURE D.1: Identifiability-stability (downstream) trade-off cross plots. As predicted by Theorem 5.9, at the extremes, unconstrained CL learns solutions that manifest the trade-off, as a consequence of the interplay of $\mathcal{T}$ on Z and X . F1 measure is computed as the harmonic mean of the predictions on the clean with those of the noisy dataset.

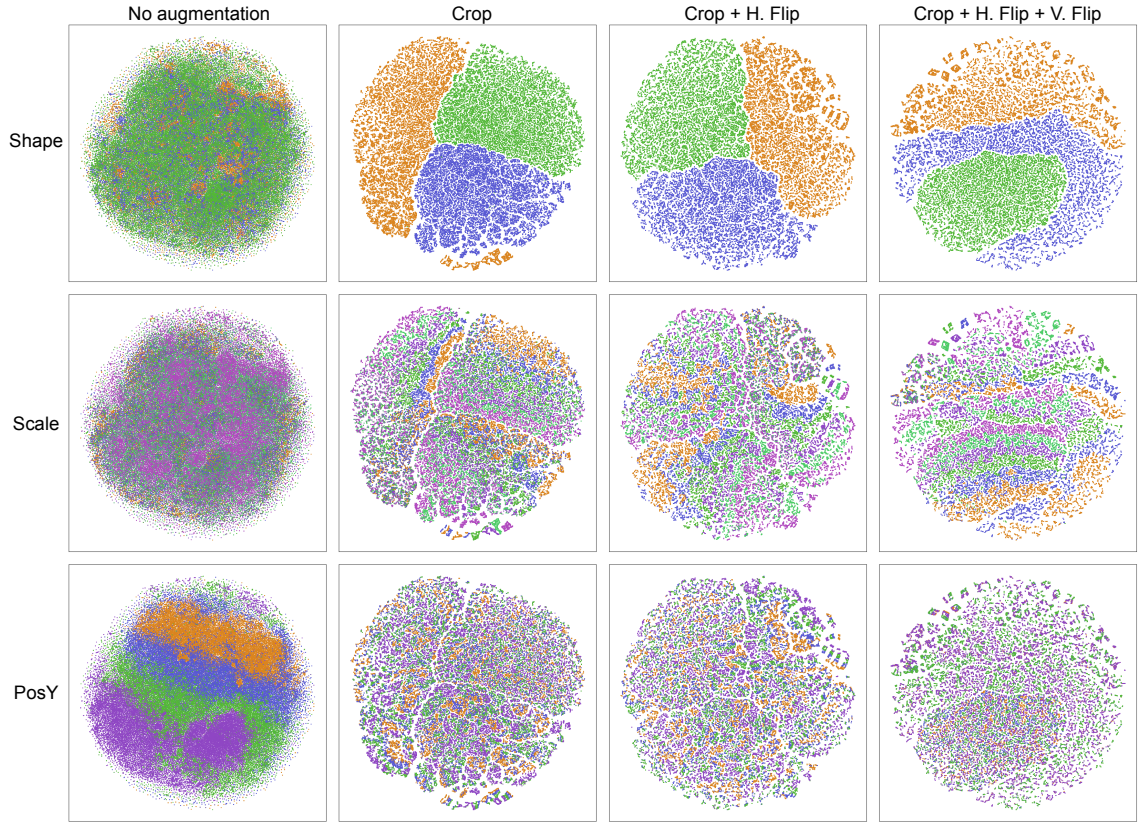


FIGURE D.2: t-SNE [309] for the validation set of one of the seeds for **dSprites** with increasingly stronger augmentations. The augmentations decorrelate the samples with respect to their semantics (shape) into $\mathcal{T}$ -connected clusters.

Bibliography

- [1] I. Goodfellow, “Deep learning,” 2016. [Cited on page [4](#).]
- [2] D. H. Wolpert and W. G. Macready, “No free lunch theorems for optimization,” *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997. [Cited on page [11](#).]
- [3] V. Vapnik, “Principles of risk minimization for learning theory,” *Advances in neural information processing systems*, vol. 4, 1991. [Cited on page [11](#).]
- [4] Y. C. Ho and D. L. Pepyne, “Simple explanation of the no-free-lunch theorem and its implications,” *Journal of Optimization Theory and Applications*, vol. 115, no. 3, pp. 549–570, 2002. [Cited on page [11](#).]
- [5] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, “Geometric deep learning: Grids, groups, graphs, geodesics, and gauges,” *arXiv preprint arXiv:2104.13478*, 2021. [Cited on page [12](#).]
- [6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. [Cited on page [12](#).]
- [7] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner *et al.*, “Relational inductive biases, deep learning, and graph networks,” *arXiv preprint arXiv:1806.01261*, 2018. [Cited on page [12](#).]
- [8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008. [Cited on page [12](#).]

- [9] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013. [Cited on page [13](#).]
- [10] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025. [Cited on page [13](#).]
- [11] W. G. Park, N. J. Shaheen, J. Cohen, I. M. Pike, D. G. Adler, J. M. Inadomi, L. A. Laine, J. G. Lieb, M. K. Rizk, M. S. Sawhney *et al.*, "Quality indicators for egd," *Gastrointestinal endoscopy*, vol. 81, no. 1, pp. 17–30, 2015. [Cited on page [15](#).]
- [12] A. L. Faulx, J. R. Lightdale, R. D. Acosta, D. Agrawal, D. H. Bruining, V. Chandrasekhara, M. A. Eloubeidi, S. R. Gurudu, L. Kelsey, M. A. Khashab *et al.*, "Guidelines for privileging, credentialing, and proctoring to perform gi endoscopy," *Gastrointestinal endoscopy*, vol. 85, no. 2, pp. 273–281, 2017.
- [13] H. Nakashima, H. Kawahira, H. Kawachi, and N. Sakaki, "Artificial intelligence diagnosis of helicobacter pylori infection using blue laser imaging-bright and linked color imaging: a single-center prospective study," *Annals of gastroenterology*, vol. 31, no. 4, p. 462, 2018.
- [14] T. Kanesaka, T.-C. Lee, N. Uedo, K.-P. Lin, H.-Z. Chen, J.-Y. Lee, H.-P. Wang, and H.-T. Chang, "Computer-aided diagnosis for identifying and delineating early gastric cancers in magnifying narrow-band imaging," *Gastrointestinal endoscopy*, vol. 87, no. 5, pp. 1339–1344, 2018. [Cited on pages [16](#) and [129](#).]
- [15] L. A. de Souza Jr, C. Palm, R. Mendel, C. Hook, A. Ebigbo, A. Probst, H. Messmann, S. Weber, and J. P. Papa, "A survey on barrett's esophagus analysis using machine learning," *Computers in biology and medicine*, vol. 96, pp. 203–213, 2018.
- [16] A. J. Trindade, M. J. McKinley, C. Fan, C. L. Leggett, A. Kahn, and D. K. Pleskow, "Endoscopic surveillance of barrett's esophagus using volumetric laser endomicroscopy with artificial intelligence image enhancement," *Gastroenterology*, vol. 157, no. 2, pp. 303–305, 2019. [Cited on page [15](#).]
- [17] H. Takiyama, T. Ozawa, S. Ishihara, M. Fujishiro, S. Shichijo, S. Nomura, M. Miura, and T. Tada, "Automatic anatomical classification of esophagogastroduodenoscopy

- images using deep convolutional neural networks," *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018. [Cited on page 15.]
- [18] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen *et al.*, "Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Scientific data*, vol. 7, no. 1, pp. 1–14, 2020. [Cited on page 15.]
- [19] F. Emura, Y. Saito, and H. Ikematsu, "Narrow-band imaging optical chromocolonoscopy: advantages and limitations," *World journal of gastroenterology: WJG*, vol. 14, no. 31, p. 4867, 2008. [Cited on page 15.]
- [20] H. Kurumi, K. Nonaka, Y. Ikebuchi, A. Yoshida, K. Kawaguchi, K. Yashima, and H. Isomoto, "Fundamentals, diagnostic capabilities, and perspective of narrow band imaging for early gastric cancer," *Journal of Clinical Medicine*, vol. 10, no. 13, p. 2918, 2021. [Cited on pages 15, 16, 18, and 26.]
- [21] F. H. A. Morita, W. M. Bernardo, E. Ide, R. S. P. Rocha, J. C. M. Aquino, M. K. Minata, K. Yamazaki, S. B. Marques, P. Sakai, and E. G. H. de Moura, "Narrow band imaging versus lugol chromoendoscopy to diagnose squamous cell carcinoma of the esophagus: a systematic review and meta-analysis," *BMC cancer*, vol. 17, no. 1, pp. 1–14, 2017. [Cited on page 16.]
- [22] M. Gruner, A. Denis, C. Masliah, M. Amil, E. Metivier-Cesbron, D. Luet, M. Kaasis, E. Coron, M. Le Rhun, S. Lecleire *et al.*, "Narrow-band imaging versus lugol chromoendoscopy for esophageal squamous cell cancer screening in normal endoscopic practice: randomized controlled trial," *Endoscopy*, vol. 53, no. 07, pp. 674–682, 2021. [Cited on page 16.]
- [23] B. W. Eom, J. H. Lee, I. J. Choi, M. C. Kook, B.-H. Nam, K. W. Ryu, and Y.-W. Kim, "Pretreatment risk factors for multiple gastric cancer and missed lesions," *Journal of surgical oncology*, vol. 105, no. 8, pp. 813–817, 2012. [Cited on page 16.]
- [24] M. Dinis-Ribeiro and E. J. Kuipers, "Identification of gastric atrophic changes: from histopathology to endoscopy," *Endoscopy*, vol. 47, no. 06, pp. 533–537, 2015.

- [25] P. Pimentel-Nunes, M. Dinis-Ribeiro, J. Soares, R. Marcos-Pinto, C. Santos, C. Rolanda, R. Bastos, M. Areia, L. Afonso, J. Bergman *et al.*, “A multicenter validation of an endoscopic classification with narrow band imaging for gastric pre-cancerous and cancerous lesions,” *Endoscopy*, vol. 44, no. 03, pp. 236–246, 2012.
- [26] K. Yao, “The endoscopic diagnosis of early gastric cancer,” *Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology*, vol. 26, no. 1, p. 11, 2013. [Cited on page 16.]
- [27] K. Nonaka, S. Arai, S. Ban, H. Kitada, M. Namoto, K. Nagata, Y. Ochiai, O. Togawa, M. Nakao, M. Nishimura *et al.*, “Prospective study of the evaluation of the usefulness of tumor typing by narrow band imaging for the differential diagnosis of gastric adenoma and well-differentiated adenocarcinoma,” *Digestive Endoscopy*, vol. 23, no. 2, pp. 146–152, 2011. [Cited on page 16.]
- [28] X. Zhang, M. Li, S. Chen, J. Hu, Q. Guo, R. Liu, H. Zheng, Z. Jin, Y. Yuan, Y. Xi *et al.*, “Endoscopic screening in asian countries is associated with reduced gastric cancer mortality: a meta-analysis and systematic review,” *Gastroenterology*, vol. 155, no. 2, pp. 347–354, 2018. [Cited on page 16.]
- [29] P. Correa, “Human gastric carcinogenesis: a multistep and multifactorial process—first american cancer society award lecture on cancer epidemiology and prevention,” *Cancer research*, vol. 52, no. 24, pp. 6735–6740, 1992. [Cited on pages 16 and 17.]
- [30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, 2012. [Cited on page 16.]
- [31] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ;0.5mb model size,” 2016. [Online]. Available: <https://arxiv.org/abs/1602.07360> [Cited on page 16.]
- [32] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015. [Online]. Available: <https://arxiv.org/abs/1409.1556> [Cited on page 16.]

- [33] P. Guimarães, A. Keller, T. Fehlmann, F. Lammert, and M. Casper, "Deep-learning based detection of gastric precancerous conditions," *Gut*, vol. 69, no. 1, pp. 4–6, 2020. [Cited on page 16.]
- [34] Y. Zhang, F. Li, F. Yuan, K. Zhang, L. Huo, Z. Dong, Y. Lang, Y. Zhang, M. Wang, Z. Gao *et al.*, "Diagnosing chronic atrophic gastritis by gastroscopy using artificial intelligence," *Digestive and Liver Disease*, vol. 52, no. 5, pp. 566–572, 2020. [Cited on page 16.]
- [35] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018. [Online]. Available: <https://arxiv.org/abs/1608.06993> [Cited on page 16.]
- [36] N. Lin, T. Yu, W. Zheng, H. Hu, L. Xiang, G. Ye, X. Zhong, B. Ye, R. Wang, W. Deng *et al.*, "Simultaneous recognition of atrophic gastritis and intestinal metaplasia on white light endoscopic images based on convolutional neural networks: a multi-center study," *Clinical and translational gastroenterology*, vol. 12, no. 8, 2021. [Cited on page 16.]
- [37] Y. Sakai, S. Takemoto, K. Hori, M. Nishimura, H. Ikematsu, T. Yano, and H. Yokota, "Automatic detection of early gastric cancer in endoscopic images using a transferring convolutional neural network," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 4138–4141. [Cited on page 17.]
- [38] L. Li, Y. Chen, Z. Shen, X. Zhang, J. Sang, Y. Ding, X. Yang, J. Li, M. Chen, C. Jin *et al.*, "Convolutional neural network for the diagnosis of early gastric cancer based on magnifying narrow band imaging," *Gastric Cancer*, vol. 23, no. 1, pp. 126–132, 2020. [Cited on page 17.]
- [39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2015. [Online]. Available: <https://arxiv.org/abs/1512.00567> [Cited on page 17.]
- [40] Y. Zhu, Q.-C. Wang, M.-D. Xu, Z. Zhang, J. Cheng, Y.-S. Zhong, Y.-Q. Zhang, W.-F. Chen, L.-Q. Yao, P.-H. Zhou *et al.*, "Application of convolutional neural network

- in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy," *Gastrointestinal endoscopy*, vol. 89, no. 4, pp. 806–815, 2019. [Cited on page 17.]
- [41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: <https://arxiv.org/abs/1512.03385> [Cited on page 17.]
- [42] Y. Horiuchi, T. Hirasawa, N. Ishizuka, Y. Tokai, K. Namikawa, S. Yoshimizu, A. Ishiyama, T. Yoshio, T. Tsuchida, J. Fujisaki *et al.*, "Performance of a computer-aided diagnosis system in diagnosing early gastric cancer using magnifying endoscopy videos with narrow-band imaging (with videos)," *Gastrointestinal Endoscopy*, vol. 92, no. 4, pp. 856–865, 2020. [Cited on page 17.]
- [43] H. Hu, L. Gong, D. Dong, L. Zhu, M. Wang, J. He, L. Shu, Y. Cai, S. Cai, W. Su *et al.*, "Identifying early gastric cancer under magnifying narrow-band images with deep learning: a multicenter study," *Gastrointestinal Endoscopy*, vol. 93, no. 6, pp. 1333–1341, 2021. [Cited on page 17.]
- [44] L. Wu, Y. Zhang, W. Zhou *et al.*, "Real-time quality control system for upper gastrointestinal endoscopy: a randomized controlled trial (with video)," *Endoscopy*, vol. 55, no. 02, pp. 123–132, 2023. [Cited on page 17.]
- [45] A. Subedi, N. Kharel, P. Basnet, and D. Jha, "Classification of endoscopy and video capsule images using cnn-transformer model," *arXiv preprint arXiv:2408.10733*, 2024, proposed a dual-branch hybrid architecture combining DenseNet201 for local features and Swin Transformer for global dependencies. [Cited on page 17.]
- [46] C. Kusters, "Design of a robust and high-performance cnn-transformer framework for gi endoscopic imaging," PhD Thesis, Eindhoven University of Technology, November 2025, demonstrated that hybrid CNN-Transformer architectures outperform standalone models by up to 16.3% in robustness tests.
- [47] Y. Oukdach, A. Garbaz, Z. Kerkaou, M. El Ansari, L. Koutti, N. Papachrysos, A. F. El Ouafdi, T. de Lange, and C. Distant, "Vision transformer distillation for enhanced gastrointestinal abnormality recognition in wireless capsule endoscopy images," *Journal of Medical Imaging*, vol. 12, no. 1, p. 014502, 2025. [Cited on page 17.]

- [48] M. Muto, K. Yao, M. Kaise, M. Kato, N. Uedo, K. Yagi, and H. Tajiri, "Magnifying endoscopy simple diagnostic algorithm for early gastric cancer (mesda-g)," *Digestive endoscopy*, vol. 28, no. 4, pp. 379–393, 2016. [Cited on page 17.]
- [49] H. Mutoh, Y. Hakamata, K. Sato, A. Eda, I. Yanaka, S. Honda, H. Osawa, Y. Kaneko, and K. Sugano, "Conversion of gastric mucosa to intestinal metaplasia in cdx2-expressing transgenic mice," *Biochemical and biophysical research communications*, vol. 294, no. 2, pp. 470–479, 2002.
- [50] R. A. Busuttil and A. Boussioutas, "Intestinal metaplasia: a premalignant lesion involved in gastric carcinogenesis," *Journal of gastroenterology and hepatology*, vol. 24, no. 2, pp. 193–201, 2009. [Cited on page 17.]
- [51] P. Correa and M. B. Piazuelo, "The gastric precancerous cascade," *Journal of digestive diseases*, vol. 13, no. 1, pp. 2–9, 2012. [Cited on pages 18, 19, 26, and 130.]
- [52] D. S. Jencks, J. D. Adam, M. L. Borum, J. M. Koh, S. Stephen, and D. B. Doman, "Overview of current concepts in gastric intestinal metaplasia and gastric cancer," *Gastroenterology & hepatology*, vol. 14, no. 2, p. 92, 2018. [Cited on page 18.]
- [53] T. Yan, P. K. Wong, I. C. Choi, C. M. Vong, and H. H. Yu, "Intelligent diagnosis of gastric intestinal metaplasia based on convolutional neural network and limited number of endoscopic images," *Computers in Biology and Medicine*, vol. 126, p. 104026, 2020. [Cited on page 18.]
- [54] W. Leung and J. Sung, "Intestinal metaplasia and gastric carcinogenesis," *Alimentary pharmacology & therapeutics*, vol. 16, no. 7, pp. 1209–1216, 2002. [Cited on page 18.]
- [55] M. Rodríguez-Carrasco, G. Esposito, D. Libânio, P. Pimentel-Nunes, and M. Dinis-Ribeiro, "Image-enhanced endoscopy for gastric preneoplastic conditions and neoplastic lesions: A systematic review and meta-analysis," *Endoscopy*, vol. 52, no. 12, pp. 1048–1065, 2020. [Cited on page 18.]
- [56] N. Uedo, R. Ishihara, H. Iishi, S. Yamamoto, T. Yamada, K. Imanaka, Y. Takeuchi, K. Higashino, S. Ishiguro, and M. Tatsuta, "A new method of diagnosing gastric intestinal metaplasia: narrow-band imaging with magnifying endoscopy," *Endoscopy*, vol. 38, no. 08, pp. 819–824, 2006. [Cited on page 18.]

- [57] L. G. Capelle, J. Haringsma, A. C. de Vries, E. W. Steyerberg, K. Biermann, H. van Dekken, and E. J. Kuipers, “Narrow band imaging for the detection of gastric intestinal metaplasia and dysplasia during surveillance endoscopy,” *Digestive diseases and sciences*, vol. 55, no. 12, pp. 3442–3448, 2010. [Cited on page 18.]
- [58] T. Sauerbruch, M. Schreiber, P. Schüssler, and W. Permanetter, “Endoscopy in the diagnosis of gastritis,” *Endoscopy*, vol. 16, no. 03, pp. 101–104, 1984. [Cited on page 18.]
- [59] P. Pimentel-Nunes, D. Libânio, R. Marcos-Pinto, M. Areia, M. Leja, G. Esposito, M. Garrido, I. Kikuste, F. Megraud, T. Matysiak-Budnik *et al.*, “Management of epithelial precancerous conditions and lesions in the stomach (maps ii): European society of gastrointestinal endoscopy (esge), european helicobacter and microbiota study group (ehmsg), european society of pathology (esp), and sociedade portuguesa de endoscopia digestiva (sped) guideline update 2019,” *Endoscopy*, vol. 51, no. 04, pp. 365–388, 2019. [Cited on pages 18 and 130.]
- [60] G. Esposito, P. Pimentel-Nunes, S. Angeletti, R. Castro, D. Libânio, G. Galli, E. Lahner, E. Di Giulio, B. Annibale, and M. Dinis-Ribeiro, “Endoscopic grading of gastric intestinal metaplasia (eggim): a multicenter validation study,” *Endoscopy*, vol. 51, no. 06, pp. 515–521, 2019. [Cited on pages xv, 18, 19, 20, 24, 25, and 26.]
- [61] A. R. Pimenta-Melo, M. Monteiro-Soares, D. Libânio, and M. Dinis-Ribeiro, “Missing rate for gastric cancer during upper gastrointestinal endoscopy: a systematic review and meta-analysis,” *European journal of gastroenterology & hepatology*, vol. 28, no. 9, pp. 1041–1049, 2016. [Cited on page 18.]
- [62] M. L. Martins, M. Pedroso, D. Libânio, M. Dinis-Ribeiro, M. Coimbra, and F. Renna, “Diagnostic performance of deep learning models for gastric intestinal metaplasia detection in narrow-band images,” in *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2023, pp. 1–4. [Cited on pages 18, 20, 21, 24, 25, and 27.]
- [63] A. L. Karperien, H. F. Jelinek, and H. Ahammer, “Multifractal formalism in image and time series analysis,” *Brain structure and function*, vol. 18, no. 21, p. 24, 2016. [Cited on pages 19, 36, 45, 46, 47, 48, 51, and 52.]
- [64] C. J. Evertsz and B. B. Mandelbrot, “Multifractal measures,” *Chaos and fractals*, vol. 1992, pp. 921–953, 1992. [Cited on pages 19, 32, 37, 39, 43, 47, 48, 49, and 54.]

- [65] G. A. Losa, "The fractal geometry of life." in *Biology Forum/Rivista di Biologia*, vol. 102, no. 1, 2009. [Cited on pages 19, 51, and 52.]
- [66] Y. Xu, F. Li, Z. Chen, J. Liang, and Y. Quan, "Encoding spatial distribution of convolutional features for texture representation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 732–22 744, 2021. [Cited on pages 19, 39, 43, 44, 51, 54, 55, and 66.]
- [67] M. Banks, D. Graham, M. Jansen, T. Gotoda, S. Coda, M. Di Pietro, N. Uedo, P. Bhandari, D. M. Pritchard, E. J. Kuipers *et al.*, "British society of gastroenterology guidelines on the diagnosis and management of patients at risk of gastric adenocarcinoma," *Gut*, vol. 68, no. 9, pp. 1545–1575, 2019. [Cited on page 21.]
- [68] P. Marza, L. Fillioux, S. Boutaj, K. Mahatha, C. Desrosiers, P. Piantanida, J. Dolz, S. Christodoulidis, and M. Vakalopoulou, "Thunder: Tile-level histopathology image understanding benchmark," *arXiv preprint arXiv:2507.07860*, 2025. [Cited on page 26.]
- [69] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He, "A survey of visual transformers," *IEEE transactions on neural networks and learning systems*, vol. 35, no. 6, pp. 7478–7498, 2023. [Cited on page 27.]
- [70] M. Pedroso, M. L. Martins, D. Libânio, M. Dinis-Ribeiro, M. Coimbra, and F. Renna, "Fractal bilinear deep neural network models for gastric intestinal metaplasia detection," in *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2023, pp. 1–5. [Cited on pages 27, 43, 51, 53, and 66.]
- [71] M. L. Martins, R. Delas, E. Almeida, D. Marques, D. Libânio, M. Dinis-Ribeiro, F. Renna, and M. T. Coimbra, "Predicting endoscopic grading of gastric intestinal metaplasia using small patches," in *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2025, pp. 1–4. [Cited on pages 27 and 66.]
- [72] F. Renna, M. Martins, A. Neto, A. Cunha, D. Libânio, M. Dinis-Ribeiro, and M. Coimbra, "Artificial intelligence for upper gastrointestinal endoscopy: A roadmap from technology development to clinical practice," *Diagnostics*, vol. 12, no. 5, p. 1278, 2022. [Cited on page 27.]

- [73] E. Almeida, M. L. Martins, D. Marques, R. Delas, T. Almeida, J. Chaves, D. Libânio, F. Renna, M. T. Coimbra, and M. Dinis-Ribeiro, “Artificial intelligence for endoscopic grading of gastric intestinal metaplasia: advancing risk stratification for gastric cancer,” *Endoscopy*, vol. 57, no. 11, pp. 1254–1260, 2025. [Cited on page 27.]
- [74] E. Almeida, M. L. Martins, D. Marques, R. Delas, T. Almeida, J. Chaves, D. Libânio, F. Renna, M. T. Coimbra, and M. Dinis-Ribeiro, “Artificial intelligence for endoscopic grading of gastric intestinal metaplasia: Another step to a greener approach on selecting individuals at risk for gastric cancer,” in *Proceedings of the 30th Portuguese Conference on Pattern Recognition (RECPAD 2024)*. Covilhã, Portugal: APRP, October 2024. [Cited on page 27.]
- [75] SuperAnnotate AI, Inc., *SuperAnnotate*, San Francisco, CA, 2024, computer software. [Online]. Available: <https://www.superannotate.com/> [Cited on page 27.]
- [76] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223. [Cited on page 29.]
- [77] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Scene parsing through ade20k dataset,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641. [Cited on page 29.]
- [78] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755. [Cited on page 29.]
- [79] H. E. Stanley and P. Meakin, “Multifractal phenomena in physics and chemistry,” *Nature*, vol. 335, no. 6189, pp. 405–409, 1988. [Cited on page 30.]
- [80] K. Falconer, *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004. [Cited on pages 31, 35, 36, and 43.]
- [81] K. J. Falconer and K. Falconer, *Techniques in fractal geometry*. Wiley Chichester, 1997, vol. 3. [Cited on page 31.]

- [82] M. F. Barnsley, R. L. Devaney, B. B. Mandelbrot, H.-O. Peitgen, D. Saupe, R. F. Voss, Y. Fisher, and M. McGuire, *The science of fractal images*. Springer, 1988, vol. 1. [Cited on pages 31, 42, and 43.]
- [83] P. Abry, S. Jaffard, and H. Wendt, “A bridge between geometric measure theory and signal processing: Multifractal analysis,” in *Operator-Related Function Theory and Time-Frequency Analysis: The Abel Symposium 2012*. Springer, 2015, pp. 1–56. [Cited on pages 32 and 66.]
- [84] S. Jaffard, S. Seuret, H. Wendt, R. Leonarduzzi, and P. Abry, “Multifractal formalisms for multivariate analysis,” *Proceedings of the Royal Society A*, vol. 475, no. 2229, p. 20190150, 2019. [Cited on page 37.]
- [85] H. Wendt, “Multifractal analysis and univariate and multivariate bayesian estimation for signals and images,” 2022. [Cited on pages 32 and 66.]
- [86] B. B. Mandelbrot and B. B. Mandelbrot, *The fractal geometry of nature*. WH freeman New York, 1982, vol. 1. [Cited on pages 33, 35, and 36.]
- [87] F. Hausdorff, “Dimension und äußeres maß,” *Mathematische Annalen*, vol. 79, no. 1-2, pp. 157–179, 1919. [Cited on page 33.]
- [88] A. B. Chhabra, C. Meneveau, R. V. Jensen, and K. Sreenivasan, “Direct determination of the $f(\alpha)$ singularity spectrum and its application to fully developed turbulence,” *Physical Review A*, vol. 40, no. 9, p. 5284, 1989. [Cited on pages 37 and 54.]
- [89] T. C. Halsey, M. H. Jensen, L. P. Kadanoff, I. Procaccia, and B. I. Shraiman, “Fractal measures and their singularities: The characterization of strange sets,” *Physical review A*, vol. 33, no. 2, p. 1141, 1986. [Cited on pages 37 and 54.]
- [90] Y. Xu, H. Ji, and C. Fermüller, “Viewpoint invariant texture description using fractal analysis,” *International Journal of Computer Vision*, vol. 83, no. 1, pp. 85–100, 2009. [Cited on pages 37, 39, 40, 42, 43, 44, 45, and 54.]
- [91] J. L. Vehel and P. Mignot, “Multifractal segmentation of images,” *Fractals*, vol. 2, no. 03, pp. 371–377, 1994. [Cited on pages xvi, 40, 42, 45, and 60.]
- [92] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141. [Cited on pages xii, xvi, 40, 44, 54, 56, 59, 62, 64, and 65.]

- [93] A. G. Roy, N. Navab, and C. Wachinger, "Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks," *IEEE transactions on medical imaging*, vol. 38, no. 2, pp. 540–549, 2018. [Cited on pages [xii](#), [40](#), [41](#), [45](#), [56](#), [57](#), [59](#), [64](#), [65](#), and [66](#).]
- [94] H. Lee, H.-E. Kim, and H. Nam, "Srm: A style-based recalibration module for convolutional neural networks," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2019, pp. 1854–1862. [Cited on pages [xii](#), [41](#), [45](#), [56](#), [59](#), [64](#), and [65](#).]
- [95] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 783–792. [Cited on pages [xii](#), [41](#), [45](#), [56](#), [59](#), [64](#), [65](#), and [66](#).]
- [96] A. Pentland, "Fractal-based description." in *IJCAI*, 1983, pp. 973–981. [Cited on page [42](#).]
- [97] N. Sarkar and B. B. Chaudhuri, "An efficient differential box-counting approach to compute fractal dimension of image," *IEEE Transactions on systems, man, and cybernetics*, vol. 24, no. 1, pp. 115–120, 1994. [Cited on pages [42](#) and [44](#).]
- [98] B. B. Chaudhuri and N. Sarkar, "Texture segmentation using fractal dimension," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 17, no. 1, pp. 72–77, 1995. [Cited on page [42](#).]
- [99] M. Varma and R. Garg, "Locally invariant fractal features for statistical texture classification," in *2007 IEEE 11th international conference on computer vision*. IEEE, 2007, pp. 1–8. [Cited on pages [42](#), [43](#), and [54](#).]
- [100] Y. Xia, D. Feng, and R. Zhao, "Morphology-based multifractal estimation for texture segmentation," *IEEE Transactions on Image Processing*, vol. 15, no. 3, pp. 614–623, 2006. [Cited on page [42](#).]
- [101] P. Brodatz, "Textures: a photographic album for artists and designers," (*No Title*), 1966. [Cited on page [42](#).]
- [102] M. Ivanovici and N. Richard, "Fractal dimension of color fractal images," *IEEE Transactions on Image Processing*, vol. 20, no. 1, pp. 227–235, 2010. [Cited on page [43](#).]
- [103] M. Ivanovici, N. Richard, and D. Paulus, "Color image segmentation," *advanced color image processing and analysis*, pp. 219–277, 2013. [Cited on page [43](#).]

- [104] M. A. Mohammed, B. Al-Khateeb, A. N. Rashid, D. A. Ibrahim, M. K. Abd Ghani, and S. A. Mostafa, "Neural network and multi-fractal dimension features for breast cancer classification from ultrasound images," *Computers & Electrical Engineering*, vol. 70, pp. 871–882, 2018. [Cited on page 43.]
- [105] G. F. Roberto, A. Lumini, L. A. Neves, and M. Z. do Nascimento, "Fractal neural network: A new ensemble of fractal geometry and convolutional neural networks for the classification of histology images," *Expert Systems with Applications*, vol. 166, p. 114103, 2021. [Cited on pages 51 and 53.]
- [106] G. Xiong, Z. Xiong, L. Jia, and T.-K. Truong, "Spatial multifractal spectrum distribution method for breast ultrasonic image classification," *Chaos, Solitons & Fractals*, vol. 172, p. 113530, 2023. [Cited on page 43.]
- [107] Z. Chen, F. Li, Y. Quan, Y. Xu, and H. Ji, "Deep texture recognition via exploiting cross-layer statistical self-similarity," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 5231–5240. [Cited on pages 44, 51, 55, and 66.]
- [108] F. Zhou and W. Huang, "Fractal few-shot learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2023. [Cited on pages 44, 52, and 55.]
- [109] Y. Chen, A. Xu, L. Yang, J. Li, Y. Jin, and Z. Shi, "Mfen: A multi-layer fractal encoding network based on color distribution for road surface condition recognition," *IEEE Transactions on Intelligent Vehicles*, 2024. [Cited on pages 44, 51, 53, and 66.]
- [110] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," *Computational visual media*, vol. 8, no. 3, pp. 331–368, 2022. [Cited on pages 44, 64, and 65.]
- [111] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2019, pp. 3024–3033. [Cited on page 45.]
- [112] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018. [Cited on page 45.]

- [113] S. Pereira, A. Pinto, J. Amorim, A. Ribeiro, V. Alves, and C. A. Silva, "Adaptive feature recombination and recalibration for semantic segmentation with fully convolutional networks," *IEEE transactions on medical imaging*, vol. 38, no. 12, pp. 2914–2925, 2019. [Cited on pages 45, 57, and 65.]
- [114] A.-M. Rickmann, A. G. Roy, I. Sarasua, and C. Wachinger, "Recalibrating 3d convnets with project & excite," *IEEE transactions on medical imaging*, vol. 39, no. 7, pp. 2461–2471, 2020. [Cited on pages 45 and 57.]
- [115] H. Wang, P. Cao, J. Wang, and O. R. Zaiane, "Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2441–2449. [Cited on page 45.]
- [116] H. Huang, S. Xie, L. Lin, Y. Iwamoto, X. Han, Y.-W. Chen, and R. Tong, "Scaleformer: revisiting the transformer-based backbones from a scale-wise perspective for medical image segmentation," *arXiv preprint arXiv:2207.14552*, 2022.
- [117] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of u-net," *arXiv preprint arXiv:2211.14830*, 2022. [Cited on pages 45, 56, and 64.]
- [118] M. Häfner, T. Tamaki, S. Tanaka, A. Uhl, G. Wimmer, and S. Yoshida, "Local fractal dimension based approaches for colonic polyp classification," *Medical image analysis*, vol. 26, no. 1, pp. 92–107, 2015. [Cited on pages 45, 46, 51, and 53.]
- [119] H. G. E. Hentschel and I. Procaccia, "The infinite number of generalized dimensions of fractals and strange attractors," *Physica D: Nonlinear Phenomena*, vol. 8, no. 3, pp. 435–444, 1983. [Cited on pages 45 and 62.]
- [120] H. Salat, R. Murcio, and E. Arcaute, "Multifractal methodology," *Physica A: Statistical Mechanics and its Applications*, vol. 473, pp. 467–487, 2017. [Cited on pages 46 and 48.]
- [121] P. Bak and K. Chen, "Self-organized criticality," *Scientific American*, vol. 264, no. 1, pp. 46–53, 1991. [Cited on page 51.]

- [122] T. Smith Jr, G. Lange, and W. B. Marks, "Fractal methods and results in cellular morphology—dimensions, lacunarity and multifractals," *Journal of neuroscience methods*, vol. 69, no. 2, pp. 123–136, 1996. [Cited on page 51.]
- [123] M. Molski and J. Konarski, "Tumor growth in the space–time with temporal fractal dimension," *Chaos, Solitons & Fractals*, vol. 36, no. 4, pp. 811–818, 2008. [Cited on pages 51 and 52.]
- [124] C. Atupelage, H. Nagahashi, M. Yamaguchi, M. Sakamoto, and A. Hashiguchi, "Multifractal feature descriptor for histopathology," *Analytical Cellular Pathology*, vol. 35, no. 2, pp. 123–126, 2012. [Cited on pages 51 and 53.]
- [125] D. Nawn, S. Pratiher, S. Chatteraj, D. Chakraborty, M. Pal, R. R. Paul, S. Dutta, and J. Chatterjee, "Multifractal alterations in oral sub-epithelial connective tissue during progression of pre-cancer and cancer," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 1, pp. 152–162, 2020. [Cited on pages 51 and 53.]
- [126] A. Piantanelli, P. Maponi, L. Scalise, S. Serresi, A. Cialabrini, and A. Basso, "Fractal characterisation of boundary irregularity in skin pigmented lesions," *Medical and Biological Engineering and Computing*, vol. 43, pp. 436–442, 2005. [Cited on pages 51 and 52.]
- [127] S. Chatterjee, D. Dey, and S. Munshi, "Optimal selection of features using wavelet fractal descriptors and automatic correlation bias reduction for classifying skin lesions," *Biomedical signal processing and control*, vol. 40, pp. 252–262, 2018. [Cited on pages 51 and 52.]
- [128] R. Korchiyne, A. Sbihi, S. M. Farssi, R. Touahni, and M. T. Alaoui, "Medical image texture segmentation using multifractal analysis," in *2012 International Conference on Multimedia Computing and Systems*. IEEE, 2012, pp. 422–425. [Cited on pages 51 and 53.]
- [129] A. Islam, S. M. Reza, and K. M. Iftexharuddin, "Multifractal texture estimation for detection and segmentation of brain tumors," *IEEE transactions on biomedical engineering*, vol. 60, no. 11, pp. 3204–3215, 2013.
- [130] S. Lahmiri, "Glioma detection based on multi-fractal features of segmented brain mri by particle swarm optimization techniques," *Biomedical Signal Processing and Control*, vol. 31, pp. 148–155, 2017. [Cited on pages 51 and 53.]

- [131] R. Lopes and N. Betrouni, “Fractal and multifractal analysis: a review,” *Medical image analysis*, vol. 13, no. 4, pp. 634–649, 2009. [Cited on pages 51 and 53.]
- [132] A. Kurakin, “The self-organizing fractal theory as a universal discovery method: the phenomenon of life,” *Theoretical Biology and Medical Modelling*, vol. 8, pp. 1–66, 2011. [Cited on page 52.]
- [133] T. Stojić, I. Reljin, and B. Reljin, “Adaptation of multifractal analysis to segmentation of microcalcifications in digital mammograms,” *Physica A: Statistical Mechanics and its Applications*, vol. 367, pp. 494–508, 2006. [Cited on page 52.]
- [134] M. H. Yap, E. A. Edirisinghe, and H. E. Bez, “A novel algorithm for initial lesion detection in ultrasound breast images,” *Journal of Applied Clinical Medical Physics*, vol. 9, no. 4, pp. 181–199, 2008. [Cited on pages 52 and 59.]
- [135] B. J. West, *Fractal physiology and chaos in medicine*. World Scientific, 2012, vol. 16. [Cited on page 53.]
- [136] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*. pmlr, 2015, pp. 448–456. [Cited on page 53.]
- [137] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, “Context encoding for semantic segmentation,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160. [Cited on pages 54, 55, and 61.]
- [138] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, “Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic),” *arXiv preprint arXiv:1902.03368*, 2019. [Cited on page 56.]
- [139] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, “Kvasir-seg: A segmented polyp dataset,” in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II* 26. Springer, 2020, pp. 451–462. [Cited on page 56.]
- [140] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, “Dataset of breast ultrasound images,” *Data in brief*, vol. 28, p. 104863, 2020. [Cited on page 56.]

- [141] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, and P. F. Jaeger, “nnu-net revisited: A call for rigorous validation in 3d medical image segmentation,” *arXiv preprint arXiv:2404.09556*, 2024. [Cited on pages 56 and 58.]
- [142] Y. Wu and K. He, “Group normalization,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19. [Cited on page 57.]
- [143] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv:1412.6980*, 2014. [Cited on pages 57 and 83.]
- [144] 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: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18. Springer, 2015, pp. 234–241. [Cited on pages 59 and 64.]
- [145] L. Lu, Y. Shin, Y. Su, and G. E. Karniadakis, “Dying relu and initialization: Theory and numerical examples,” *arXiv preprint arXiv:1903.06733*, 2019. [Cited on page 62.]
- [146] K. Park, Y. J. Choe, and V. Veitch, “The linear representation hypothesis and the geometry of large language models,” *arXiv preprint arXiv:2311.03658*, 2023. [Cited on page 63.]
- [147] M. Huh, B. Cheung, T. Wang, and P. Isola, “The platonic representation hypothesis,” *arXiv preprint arXiv:2405.07987*, 2024. [Cited on page 63.]
- [148] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I* 13. Springer, 2014, pp. 818–833. [Cited on page 64.]
- [149] M. Raghu, J. Gilmer, J. Yosinski, and J. Sohl-Dickstein, “Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability,” *Advances in neural information processing systems*, vol. 30, 2017. [Cited on page 64.]
- [150] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19. [Cited on page 65.]

- [151] J. Bruna and S. Mallat, “Invariant scattering convolution networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013. [Cited on page 66.]
- [152] M. L. Martins, M. T. Coimbra, and F. Renna, “Singularity strength re-calibration of fully convolutional neural networks for biomedical image segmentation,” in *Proceedings of the 2024 32nd European Signal Processing Conference (EUSIPCO)*, 2024. [Cited on page 66.]
- [153] —, “Understanding squeeze-and-excitation layers for medical image segmentation,” in *Proceedings of the 33rd European Signal Processing Conference (EUSIPCO 2025)*, Isola delle Femmine (Palermo), Italy, Sep. 2025. [Cited on page 66.]
- [154] —, “Monofractal and multifractal recalibration of fully convolutional networks for medical image segmentation,” *SIAM Journal on Imaging Sciences*, 2025, manuscript submitted. [Cited on page 66.]
- [155] M. L. Martins, F. Renna, and M. T. Coimbra, “Multifractal attention functions for fully convolutional neural networks,” in *Proceedings of the 30th Portuguese Conference on Pattern Recognition (RECPAD 2024)*. Covilhã, Portugal: APRP, October 2024. [Cited on page 66.]
- [156] O. Cappé, É. Moulines, and T. Rydén, *Inference in Hidden Markov Models*, ser. Springer Series in Statistics. New York, NY: Springer, 2005. [Cited on page 67.]
- [157] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4. [Cited on pages 67 and 68.]
- [158] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 2002. [Cited on pages 67, 69, 72, 81, and 82.]
- [159] L. Fritz and D. Burshtein, “Simplified end-to-end mmi training and voting for asr,” *arXiv preprint arXiv:1703.10356*, 2017. [Cited on pages 71 and 76.]
- [160] ENS Paris-Saclay, Department of Mathematics, “Perron–frobenius theory for stochastic matrices,” Lecture notes, 2023, available online. [Online]. Available: <https://www.ens-paris-saclay.fr/> [Cited on page 71.]

- [161] G. Strang, *Introduction to linear algebra*. SIAM, 2022. [Cited on page 72.]
- [162] C. Michelot, "A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$," *Journal of Optimization Theory and Applications*, vol. 50, no. 1, pp. 195–200, 1986. [Cited on pages 73 and 80.]
- [163] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 249–256. [Cited on page 74.]
- [164] G. Hinton, P. Dayan, A. To, and R. Neal, "The helmholtz machine through time," in *ICANN-95*, 1995, pp. 483–490. [Cited on page 76.]
- [165] D. S. Clouse, C. L. Giles, B. G. Horne, and G. W. Cottrell, "Time-delay neural networks: Representation and induction of finite-state machines," *IEEE Transactions on Neural Networks*, vol. 8, no. 5, pp. 1065–1070, 1997. [Cited on page 76.]
- [166] M. Qu, Y. Bengio, and J. Tang, "Gmnn: Graph markov neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 5241–5250. [Cited on page 76.]
- [167] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012. [Cited on page 76.]
- [168] "Cardiovascular diseases - global facts and figures," 2021. [Online]. Available: <https://world-heart-federation.org/resource/cardiovascular-diseases-cvds-global-facts-figures/> [Cited on page 76.]
- [169] S. Mendis, P. Puska, B. Norrving, W. H. Organization *et al.*, *Global atlas on cardiovascular disease prevention and control*. World Health Organization, 2011. [Cited on page 76.]
- [170] S. Mangione, "Cardiac auscultatory skills of physicians-in-training: a comparison of three English-speaking countries," *The American journal of medicine*, vol. 110, no. 3, pp. 210–216, 2001. [Cited on page 76.]

- [171] A. A. Ishmail, S. Wing, J. Ferguson, T. A. Hutchinson, S. Magder, and K. M. Flegel, "Interobserver agreement by auscultation in the presence of a third heart sound in patients with congestive heart failure," *Chest*, vol. 91, no. 6, pp. 870–873, 1987.
- [172] H. Vermarien, *Phonocardiography*, 2nd ed., J. G. Webster, Ed. John Wiley & Sons, Ltd, 2006, vol. 5. [Cited on page 76.]
- [173] I. Bank, H. W. Vliegen, and A. V. Bruschke, "The 200th anniversary of the stethoscope: Can this low-tech device survive in the high-tech 21st century?" *European heart journal*, vol. 37, no. 47, pp. 3536–3543, 2016. [Cited on page 76.]
- [174] S. Leng, R. S. Tan, K. T. C. Chai, C. Wang, D. Ghista, and L. Zhong, "The electronic stethoscope," *Biomedical engineering online*, vol. 14, no. 1, pp. 1–37, 2015. [Cited on page 76.]
- [175] W. Zhang, J. Han, and S. Deng, "Abnormal heart sound detection using temporal quasi-periodic features and long short-term memory without segmentation," *Biomedical Signal Processing and Control*, vol. 53, p. 101560, 2019. [Cited on page 77.]
- [176] P. T. Krishnan, P. Balasubramanian, and S. Umapathy, "Automated heart sound classification system from unsegmented phonocardiogram (PCG) using deep neural network," *Physical and Engineering Sciences in Medicine*, vol. 43, pp. 505–515, 2020.
- [177] B. Xiao, Y. Xu, X. Bi, W. Li, Z. Ma, J. Zhang, and X. Ma, "Follow the sound of children's heart: A deep-learning-based computer-aided pediatric CHDs diagnosis system," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 1994–2004, 2020.
- [178] V. Maknickas and A. Maknickas, "Recognition of normal–abnormal phonocardiographic signals using deep convolutional neural networks and mel-frequency spectral coefficients," *Physiological Measurement*, vol. 38, no. 8, p. 1671, 2017.
- [179] T. Nilanon, J. Yao, J. Hao, S. Purushotham, and Y. Liu, "Normal/abnormal heart sound recordings classification using convolutional neural network," in *Computing in Cardiology Conference (CinC)*, 2016, pp. 585–588. [Cited on page 77.]
- [180] J. Oliveira, D. Nogueira, F. Renna, C. Ferreira, A. M. Jorge, and M. Coimbra, "Do we really need a segmentation step in heart sound classification algorithms?" in *2021*

- 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 286–289. [Cited on page 77.]
- [181] M. A. Reyna, Y. Kiarashi, A. Elola, J. Oliveira, F. Renna, A. Gu, E. A. Perez-Alday, N. Sadr, A. Sharma, S. Mattos *et al.*, “Heart Murmur Detection from Phonocardiogram recordings: The George B. Moody PhysioNet challenge 2022,” *medRxiv*, pp. 2022–08, 2022. [Cited on pages 77 and 84.]
- [182] *Detection of heart murmurs in phonocardiograms with parallel hidden semi-Markov models*. Proceedings of Computing in Cardiology 2022, 2022. [Cited on page 77.]
- [183] H. Liang, S. Lukkarinen, and I. Hartimo, “Heart sound segmentation algorithm based on heart sound envelopogram,” in *Computers in Cardiology*, 1997, pp. 105–108. [Cited on page 78.]
- [184] S. Ari, P. Kumar, and G. Saha, “A robust heart sound segmentation algorithm for commonly occurring heart valve diseases,” *J. Med. Eng. Technol.*, vol. 32, no. 6, pp. 456–465, 2008.
- [185] A. Moukadem, A. Dieterlen, N. Hueber, and C. Brandt, “A robust heart sounds segmentation module based on S-transform,” *Biomedical Signal Processing and Control*, vol. 8, no. 3, pp. 273–281, 2013.
- [186] S. Sun, Z. Jiang, H. Wang, and Y. Fang, “Automatic moment segmentation and peak detection analysis of heart sound pattern via short-time modified Hilbert transform,” *Comput. Methods Programs biomed.*, vol. 114, no. 3, pp. 219–230, 2014.
- [187] L. Huiying, L. Sakari, and H. Iiro, “A heart sound segmentation algorithm using wavelet decomposition and reconstruction,” in *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. ‘Magnificent Milestones and Emerging Opportunities in Medical Engineering’* (Cat. No. 97CH36136), vol. 4. IEEE, 1997, pp. 1630–1633.
- [188] A. Castro, T. T. V. Vinhoza, S. S. Mattos, and M. T. Coimbra, “Heart sound segmentation of pediatric auscultations using wavelet analysis,” in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, July 2013, pp. 3909–3912. [Cited on page 78.]

- [189] H. Naseri and M. Homaeinezhad, "Detection and boundary identification of phonocardiogram sounds using an expert frequency-energy based metric," *Ann. Biomed. Eng.*, vol. 41, no. 2, pp. 279–292, 2013. [Cited on page 78.]
- [190] D. Kumar, P. Carvalho, M. Antunes, J. Henriques, L. Eugenio, R. Schmidt, and J. Habetha, "Detection of S1 and S2 heart sounds by high frequency signatures," in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2006, pp. 1410–1416. [Cited on page 78.]
- [191] J. Vepa, "Classification of heart murmurs using cepstral features and support vector machines," in *IEEE EMBC*, 2009, pp. 2539–2542. [Cited on page 78.]
- [192] C. N. Gupta, R. Palaniappan, S. Swaminathan, and S. M. Krishnan, "Neural network classification of homomorphic segmented heart sounds," *Appl. Soft Comput.*, vol. 7, no. 1, pp. 286–297, 2007. [Cited on page 78.]
- [193] T. Chen, K. Kuan, L. A. Celi, and G. D. Clifford, "Intelligent heartsound diagnostics on a cellphone using a hands-free kit," in *2010 AAAI Spring Symposium Series*, 2010. [Cited on page 78.]
- [194] A. C. Stasis, E. Loukis, S. Pavlopoulos, and D. Koutsouris, "Using decision tree algorithms as a basis for a heart sound diagnosis decision support system," in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2003, pp. 354–357. [Cited on page 78.]
- [195] J. E. Hebden and J. Torry, "Neural network and conventional classifiers to distinguish between first and second heart sounds," in *IEE Colloq. Artif. Intell. Methods Biomed. Data Process.*, 1996, pp. 1–6. [Cited on page 78.]
- [196] A. A. Sepehri, A. Gharehbaghi, T. Dutoit, A. Kocharian, and A. Kiani, "A novel method for pediatric heart sound segmentation without using the ECG," *Computer Methods Programs Biomed.*, vol. 99, no. 1, pp. 43–48, 2010.
- [197] T.-E. Chen, S.-I. Yang, L.-T. Ho, K.-H. Tsai, Y.-H. Chen, Y.-F. Chang, Y.-H. Lai, S.-S. Wang, Y. Tsao, and C.-C. Wu, "S1 and S2 heart sound recognition using deep neural networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 372–380, 2017. [Cited on page 78.]

- [198] L. Gamero and R. Watrous, "Detection of the first and second heart sound using probabilistic models," in *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2003, pp. 2877–2880. [Cited on page 78.]
- [199] D. Gill, N. Gavrieli, and N. Intrator, "Detection and identification of heart sounds using homomorphic envelopogram and self-organizing probabilistic model," in *Computers in Cardiology*, 2005, pp. 957–960. [Cited on page 78.]
- [200] Y.-J. Chung, *Pattern Recognition and Image Analysis, Iberian Conference*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, ch. Classification of Continuous Heart Sound Signals Using the Ergodic Hidden Markov Model, pp. 563–570. [Cited on page 78.]
- [201] S. Schmidt, E. Toft, C. Holst-Hansen, C. Graff, and J. Struijk, "Segmentation of heart sound recordings from an electronic stethoscope by a duration dependent hidden-Markov model," in *Computers in Cardiology*, 2008, pp. 345–348. [Cited on page 78.]
- [202] S. Schmidt, C. Holst-Hansen, C. Graff, E. Toft, and J. J. Struijk, "Segmentation of heart sound recordings by a duration-dependent hidden Markov model," *Physiological measurement*, vol. 31, no. 4, pp. 513–529, 2010. [Cited on pages 78, 81, and 85.]
- [203] J. Oliveira, T. Mantadelis, F. Renna, P. Gomes, and M. Coimbra, "On modifying the temporal modeling of HSMMs for pediatric heart sound segmentation," in *2017 IEEE International Workshop on Signal Processing Systems (SiPS)*, Oct 2017, pp. 1–6. [Cited on page 78.]
- [204] J. Oliveira, F. Renna, and M. Coimbra, "A subject-driven unsupervised hidden semi-Markov model and Gaussian mixture model for heart sound segmentation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 323–331, 2019. [Cited on pages 78 and 85.]
- [205] D. B. Springer, L. Tarassenko, and G. D. Clifford, "Support vector machine hidden semi-Markov model-based heart sound segmentation," in *Computing in Cardiology Conference (CinC)*, Sept 2014, pp. 625–628. [Cited on page 78.]
- [206] —, "Logistic regression-hsmm-based heart sound segmentation," *IEEE transactions on biomedical engineering*, vol. 63, no. 4, pp. 822–832, 2015. [Cited on pages xvii, xviii, 78, 79, 81, 83, 84, 85, 87, 88, 89, 90, 91, 95, 143, and 145.]

- [207] F. Renna, J. H. Oliveira, and M. T. Coimbra, "Deep convolutional neural networks for heart sound segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2435–2445, 2019. [Cited on pages [xvii](#), [79](#), [81](#), [83](#), [85](#), [87](#), [88](#), [89](#), [90](#), [91](#), and [143](#).]
- [208] E. Messner, M. Zöhrer, and F. Pernkopf, "Heart sound segmentation: An event detection approach using deep recurrent neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 9, pp. 1964–1974, 2018. [Cited on page [79](#).]
- [209] T. Fernando, H. Ghaemmaghami, S. Denman, S. Sridharan, N. Hussain, and C. Fookes, "Heart sound segmentation using bidirectional LSTMs with attention," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 6, pp. 1601–1609, 2020. [Cited on pages [xvii](#), [79](#), [81](#), [83](#), [87](#), [88](#), [89](#), [90](#), and [91](#).]
- [210] X. Wang, C. Liu, Y. Li, X. Cheng, J. Li, and G. D. Clifford, "Temporal-framing adaptive network for heart sound segmentation without prior knowledge of state duration," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 2, pp. 650–663, 2021. [Cited on page [79](#).]
- [211] C. Liu *et al.*, "An open access database for the evaluation of heart sound algorithms," *Physiological Measurement*, vol. 37, no. 12, pp. 2181–2213, 2016. [Cited on pages [82](#), [83](#), and [84](#).]
- [212] J. Oliveira, F. Renna, P. D. Costa, M. Nogueira, C. Oliveira, C. Ferreira, A. Jorge, S. Mattos, T. Hatem, T. Tavares *et al.*, "The CirCor DigiScope dataset: from murmur detection to murmur classification," *IEEE journal of biomedical and health informatics*, vol. 26, no. 6, pp. 2524–2535, 2021. [Cited on pages [xviii](#), [82](#), [84](#), [140](#), [141](#), [142](#), and [144](#).]
- [213] J. Bloit and X. Rodet, "Short-time Viterbi for online HMM decoding: Evaluation on a real-time phone recognition task," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 2121–2124. [Cited on page [96](#).]
- [214] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen, "Variational autoencoders and nonlinear ica: A unifying framework," in *International conference on artificial intelligence and statistics*. PMLR, 2020, pp. 2207–2217. [Cited on page [96](#).]

- [215] H. Hälvä and A. Hyvärinen, “Hidden markov nonlinear ICA: Unsupervised learning from nonstationary time series,” in *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, vol. 124, 2020, arXiv preprint arXiv:2006.12107. [Online]. Available: <https://arxiv.org/abs/2006.12107>
- [216] R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel, “Contrastive learning inverts the data generating process,” in *International conference on machine learning*. PMLR, 2021, pp. 12 979–12 990. [Cited on page 96.]
- [217] P. Reizinger, A. Bizeul, A. Juhos, J. E. Vogt, R. Balestrieri, W. Brendel, and D. Klindt, “Cross-entropy is all you need to invert the data generating process,” *arXiv preprint arXiv:2410.21869*, 2024. [Cited on page 96.]
- [218] F. Renna, M. L. Martins, and M. Coimbra, “Joint training of hidden markov model and neural network for heart sound segmentation,” in *Computing in Cardiology (CinC)*, Brno, Czech Republic, September 2021, pp. 1–4. [Cited on page 96.]
- [219] M. L. Martins, M. T. Coimbra, and F. Renna, “Markov-based neural networks for heart sound segmentation: Using domain knowledge in a principled way,” *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 11, pp. 5357–5368, 2023. [Cited on page 96.]
- [220] R. Silva and T. Reis, “Técnica para melhorar auscultação cardíaca vale prémio a docente da fcup,” <https://noticias.up.pt/2021/10/04/tecnica-para-melhorar-auscultacao-cardiaca-vale-premio-a-docente-da-fcup/>, October 4 2021, notícias U.Porto (Universidade do Porto). [Cited on page 96.]
- [221] R. Silva and R. Abreu, “Investigadores da u.porto dominam prémios do recpad 2021,” <https://noticias.up.pt/2021/11/17/investigadores-da-u-porto-dominam-premios-do-recpad-2021/>, November 17 2021, notícias U.Porto (Universidade do Porto). [Cited on page 96.]
- [222] Advanced Processing of Biomedical Signals and Images Consortium, “Lecturers and organizers – a-bip program,” <https://a-bip.pt/#lecturers>, 2025, blended Intensive Program (BIP) official website. [Cited on page 96.]
- [223] XX ENEEB Organizing Committee, “Workshops – xx eneeb,” <https://eneeb.aneeb.pt/workshops-carol/>, 2025, workshops page of the XX ENEEB (Engenharia Biomédica) event.

- [224] Symposium on Bioengineering Organizing Committee, “Workshops and visits descriptions – symposium on bioengineering,” <https://www.symposium.nebfeupicbas.pt/workshops.php>, 2025, workshops and visits information from the Symposium on Bioengineering website.
- [225] IEEE-EMBS International Conference on Body Sensor Networks (BSN) 2025 Organizing Committee, “Workshops & tutorials – ieee-embs international conference on body sensor networks (bsn 2025),” <https://bsn.embs.org/2025/workshops/>, 2025, official workshops and tutorials listing. [Cited on page 96.]
- [226] Hugging Face, “Hugging face datasets api,” <https://huggingface.co/docs/datasets>, 2025, accessed via Hugging Face documentation. [Cited on page 97.]
- [227] M. L. Martins, “Circordigscope dataset now available at huggingface dataset hub!” <https://miguelmartins.github.io/posts/2025/07/circor-digiscope-huggingface/>, July 3 2025, blog post on Miguel L. Martins’s personal website. [Cited on page 97.]
- [228] N. C. Thompson, K. Greenewald, K. Lee, and G. F. Manso, “The computational limits of deep learning,” *arXiv preprint arXiv:2007.05558*, 2020. [Online]. Available: <https://arxiv.org/abs/2007.05558> [Cited on page 99.]
- [229] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann *et al.*, “Scaling language models: Methods, analysis & insights from training gopher,” *arXiv preprint arXiv:2112.11446*, 2021. [Online]. Available: <https://arxiv.org/abs/2112.11446>
- [230] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’21)*. ACM, 2021, pp. 610–623.
- [231] P. Villalobos, J. Sevilla, L. Heim, T. Besiroglu, M. Hobbhahn, and A. Ho, “Will we run out of data? an analysis of the limits of scaling datasets in machine learning,” *arXiv preprint arXiv:2211.04325*, 2022. [Online]. Available: <https://arxiv.org/abs/2211.04325> [Cited on page 99.]
- [232] R. Bommasani, “On the opportunities and risks of foundation models,” *arXiv preprint arXiv:2108.07258*, 2021. [Cited on page 99.]

- [233] P. Reizinger, R. Balestriero, D. Klindt, and W. Brendel, “Position: An empirically grounded identifiability theory will accelerate self supervised learning research,” in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu, Eds., vol. 267. PMLR, 13–19 Jul 2025, pp. 82 072–82 090. [Online]. Available: <https://proceedings.mlr.press/v267/reizinger25a.html> [Cited on pages 99, 107, 108, and 125.]
- [234] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging properties in self-supervised vision transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660. [Cited on page 99.]
- [235] Y. Yu, T. Chu, S. Tong, Z. Wu, D. Pai, S. Buchanan, and Y. Ma, “Emergence of segmentation with minimalistic white-box transformers,” in *Conference on Parsimony and Learning*, ser. Proceedings of Machine Learning Research, Y. Chi, G. K. Dziugaite, Q. Qu, A. W. Wang, and Z. Zhu, Eds., vol. 234. PMLR, 03–06 Jan 2024, pp. 72–93. [Online]. Available: <https://proceedings.mlr.press/v234/yu24a.html>
- [236] F. Charton and J. Kempe, “Emergent properties with repeated examples,” *arXiv preprint arXiv:2410.07041*, 2024.
- [237] 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. [Cited on page 99.]
- [238] A. N. Gorban and I. Y. Tyukin, “Blessing of dimensionality: mathematical foundations of the statistical physics of data,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2118, p. 20170237, 2018. [Cited on page 99.]
- [239] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen *et al.*, “Toy models of superposition,” *arXiv preprint arXiv:2209.10652*, 2022. [Cited on page 99.]

- [240] M. Huh, B. Cheung, T. Wang, and P. Isola, “The platonic representation hypothesis,” *CoRR*, vol. abs/2405.07987, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2405.07987> [Cited on page 99.]
- [241] C. Fefferman, S. Mitter, and H. Narayanan, “Testing the manifold hypothesis,” *Journal of the American Mathematical Society*, vol. 29, no. 4, pp. 983–1049, 2016. [Cited on page 99.]
- [242] B. Kiani, J. Wang, and M. Weber, “Hardness of learning neural networks under the manifold hypothesis,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 5661–5696, 2024. [Cited on page 99.]
- [243] A. Bastounis, A. C. Hansen, and V. Vlačić, “The extended smale’s 9th problem - on computational barriers and paradoxes in estimation, regularisation, computer-assisted proofs, and learning.” [Online]. Available: <http://www.mins.ee.ethz.ch/pubs/p/smales9th> [Cited on page 99.]
- [244] M. J. Colbrook, V. Antun, and A. C. Hansen, “The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and smale’s 18th problem,” *Proceedings of the National Academy of Sciences*, vol. 119, no. 12, p. e2107151119, 2022. [Cited on page 99.]
- [245] D. Klindt, C. O’Neill, P. Reizinger, H. Maurer, and N. Miolane, “From superposition to sparse codes: interpretable representations in neural networks,” *arXiv preprint arXiv:2503.01824*, 2025. [Cited on pages 99, 108, and 125.]
- [246] E. Rusak, P. Reizinger, A. Juhos, O. Bringmann, R. S. Zimmermann, and W. Brendel, “Infonce: Identifying the gap between theory and practice,” in *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. Li, S. Mandt, S. Agrawal, and E. Khan, Eds., vol. 258. PMLR, 03–05 May 2025, pp. 4159–4167. [Online]. Available: <https://proceedings.mlr.press/v258/rusak25a.html> [Cited on pages 100, 102, 103, 107, 108, and 115.]
- [247] S. Matthes, Z. Han, and H. Shen, “Towards a unified framework of contrastive learning for disentangled representations,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 67 459–67 470, 2023.

- [248] P. Reizinger, A. Bizeul, A. Juhos, J. E. Vogt, R. Balestrieri, W. Brendel, and D. Klindt, "Cross-entropy is all you need to invert the data generating process," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: <https://openreview.net/forum?id=hrqNOxpItr> [Cited on pages 102, 103, 107, 108, 111, 112, 114, 115, and 116.]
- [249] R. Balestrieri, "Unsupervised learning on a diet: Datum index as target free of self-supervision, reconstruction, projector head," *arXiv preprint arXiv:2302.10260*, 2023. [Cited on page 102.]
- [250] M. Ibrahim, D. Klindt, and R. Balestrieri, "Occam's razor for self supervised learning: What is sufficient to learn good representations?" 2024. [Online]. Available: <https://arxiv.org/abs/2406.10743> [Cited on pages 104, 107, 109, and 115.]
- [251] B. Rodas, N. Montesino, J. Ambsdorf, D. Klindt, and R. Balestrieri, "Diet-cp: Lightweight and data efficient self supervised continued pretraining," 2025. [Online]. Available: <https://arxiv.org/abs/2509.06990>
- [252] R. Balestrieri and Y. LeCun, "Lejepa: Provable and scalable self-supervised learning without the heuristics," *arXiv preprint arXiv:2511.08544*, 2025. [Cited on pages 102, 107, and 109.]
- [253] V. Antun, F. Renna, C. Poon, B. Adcock, and A. C. Hansen, "On instabilities of deep learning in image reconstruction and the potential costs of ai," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 088–30 095, 2020. [Cited on page 103.]
- [254] M. J. Colbrook, V. Antun, and A. C. Hansen, "The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and smale's 18th problem," *Proceedings of the National Academy of Sciences*, vol. 119, no. 12, p. e2107151119, 2022.
- [255] N. M. Gottschling, V. Antun, A. C. Hansen, and B. Adcock, "The troublesome kernel: On hallucinations, no free lunches, and the accuracy-stability tradeoff in inverse problems," *SIAM Review*, vol. 67, no. 1, pp. 73–104, 2025. [Cited on pages 103 and 104.]
- [256] S. Smale, "Mathematical problems for the next century," *The Mathematical intelligencer*, vol. 20, no. 2, pp. 7–15, 1998. [Cited on page 103.]

- [257] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, “What makes for good views for contrastive learning?” *Advances in neural information processing systems*, vol. 33, pp. 6827–6839, 2020. [Cited on pages 103 and 108.]
- [258] T. Wang and P. Isola, “Understanding contrastive representation learning through alignment and uniformity on the hypersphere,” in *International conference on machine learning*. PMLR, 2020, pp. 9929–9939. [Cited on pages 107 and 108.]
- [259] Y. Wang, Q. Zhang, Y. Wang, J. Yang, and Z. Lin, “Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap,” in *International Conference on Learning Representations (ICLR)*, 2023. [Cited on pages 103, 104, 107, 108, and 109.]
- [260] R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel, “Contrastive learning inverts the data generating process,” in *International conference on machine learning*. PMLR, 2021, pp. 12 979–12 990. [Cited on pages 103, 107, and 108.]
- [261] J. Von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello, “Self-supervised learning with data augmentations provably isolates content from style,” *Advances in neural information processing systems*, vol. 34, pp. 16 451–16 467, 2021. [Cited on pages 103, 107, and 108.]
- [262] Q. Garrido, Y. Chen, A. Bardes, L. Najman, and Y. Lecun, “On the duality between contrastive and non-contrastive self-supervised learning,” in *International Conference on Learning Representations (ICLR)*, 2023. [Cited on page 107.]
- [263] A. Jha, M. B. Blaschko, Y. M. Asano, and T. Tuytelaars, “The common stability mechanism behind most self-supervised learning approaches,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 16 609–16 619. [Cited on pages 107 and 114.]
- [264] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018, accepted at NeurIPS 2018. [Cited on page 107.]
- [265] A. Hyvärinen, I. Khemakhem, and R. Monti, “Identifiability of latent-variable and structural-equation models: from linear to nonlinear,” *Annals of the Institute of Statistical Mathematics*, vol. 76, no. 1, pp. 1–33, 2024. [Cited on page 108.]

- [266] A. Hyvarinen, H. Sasaki, and R. Turner, “Nonlinear ica using auxiliary variables and generalized contrastive learning,” in *The 22nd international conference on artificial intelligence and statistics*. PMLR, 2019, pp. 859–868. [Cited on page 108.]
- [267] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen, “Variational autoencoders and nonlinear ica: A unifying framework,” in *International conference on artificial intelligence and statistics*. PMLR, 2020, pp. 2207–2217. [Cited on page 108.]
- [268] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 1597–1607. [Cited on pages 108 and 115.]
- [269] Y. Dubois, B. Bloem-Reddy, K. Ullrich, and C. J. Maddison, “Lossy compression for lossless prediction,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 014–14 028, 2021.
- [270] R. Shwartz Ziv and Y. LeCun, “To compress or not to compress—self-supervised learning and information theory: A review,” *Entropy*, vol. 26, no. 3, p. 252, 2024. [Cited on page 108.]
- [271] J. Cui, H. Wen, and Y. Wang, “An augmentation-aware theory for self-supervised contrastive learning,” in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: <https://openreview.net/forum?id=cN9GkuBo4w> [Cited on page 109.]
- [272] W. Morningstar, A. Bijamov, C. Duvarney, L. Friedman, N. Kalibhat, L. Liu, P. Mansfield, R. Rojas-Gomez, K. Singhal, B. Green *et al.*, “Augmentations vs algorithms: What works in self-supervised learning,” *arXiv preprint arXiv:2403.05726*, 2024. [Cited on page 109.]
- [273] Y. Zhong, H. Tang, J. Chen, J. Peng, and Y.-X. Wang, “Is self-supervised contrastive learning more robust than supervised learning?” in *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. [Online]. Available: <https://openreview.net/forum?id=fuPVTbOEKuj> [Cited on page 109.]
- [274] I. Bendidi, A. Bardes, E. Cohen, A. Lamiable, G. Bollot, and A. Genovesio, “No free lunch in self supervised representation learning,” *arXiv preprint arXiv:2304.11718*, 2023.

- [275] C. Li, R. Pang, Z. Xi, T. Du, S. Ji, Y. Yao, and T. Wang, “An embarrassingly simple backdoor attack on self-supervised learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4367–4378. [Cited on page 109.]
- [276] H. Van Assel, M. Ibrahim, T. Biancalani, A. Regev, and R. Balestrieri, “Joint embedding vs reconstruction: Provable benefits of latent space prediction for self supervised learning,” *arXiv preprint arXiv:2505.12477*, 2025. [Cited on pages 109 and 116.]
- [277] R. Balestrieri, L. Bottou, and Y. LeCun, “The effects of regularization and data augmentation are class dependent,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 37 878–37 891, 2022. [Cited on page 109.]
- [278] N. Saunshi, J. Ash, S. Goel, D. Misra, C. Zhang, S. Arora, S. Kakade, and A. Krishnamurthy, “Understanding contrastive learning requires incorporating inductive biases,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 19 250–19 286. [Cited on page 109.]
- [279] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999. [Cited on page 112.]
- [280] S. Wright, “Correlation and causation,” *Journal of agricultural research*, vol. 20, no. 7, p. 557, 1921. [Cited on page 112.]
- [281] T. Moutakanni, M. Oquab, M. Szafraniec, M. Vakalopoulou, and P. Bojanowski, “You don’t need data-augmentation in self-supervised learning,” *CoRR*, vol. abs/2406.09294, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2406.09294> [Cited on page 113.]
- [282] A. Damian, T. Ma, and J. D. Lee, “Label noise sgd provably prefers flat global minimizers,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 449–27 461, 2021. [Cited on pages 113 and 114.]
- [283] L. Wu, M. Wang, and W. Su, “The alignment property of sgd noise and how it helps select flat minima: A stability analysis,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 4680–4693, 2022. [Cited on page 113.]
- [284] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, “Challenging common assumptions in the unsupervised learning of disentangled

- representations,” in *international conference on machine learning*. PMLR, 2019, pp. 4114–4124. [Cited on pages 115 and 118.]
- [285] Y. Wang, K. Hu, S. Gupta, Z. Ye, Y. Wang, and S. Jegelka, “Understanding the role of equivariance in self-supervised learning,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 127 483–127 510, 2024. [Cited on pages 117 and 119.]
- [286] M. Coimbra, F. Riaz, M. Areia, F. B. Silva, and M. Dinis-Ribeiro, “Segmentation for classification of gastroenterology images,” in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 4744–4747. [Cited on pages 127 and 128.]
- [287] F. Riaz, M. Areia, F. B. Silva, M. Dinis-Ribeiro, P. P. Nunes, and M. Coimbra, “Gabor textons for classification of gastroenterology images,” in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE, 2011, pp. 117–120. [Cited on page 127.]
- [288] F. Riaz, F. B. Silva, M. D. Ribeiro, and M. T. Coimbra, “Invariant gabor texture descriptors for classification of gastroenterology images,” *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 10, pp. 2893–2904, 2012.
- [289] —, “Impact of visual features on the segmentation of gastroenterology images using normalized cuts,” *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 5, pp. 1191–1201, 2012. [Cited on pages 127 and 128.]
- [290] M. Varma and A. Zisserman, “A statistical approach to texture classification from single images,” *International journal of computer vision*, vol. 62, pp. 61–81, 2005. [Cited on page 127.]
- [291] N. Dioguardi, F. Grizzi, B. Franceschini, P. Bossi, and C. Russo, “Liver fibrosis and tissue architectural change measurement using fractal-rectified metrics and hurst’s exponent,” *World journal of gastroenterology: WJG*, vol. 12, no. 14, p. 2187, 2006. [Cited on page 127.]
- [292] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000. [Cited on page 128.]

- [293] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002. [Cited on page 128.]
- [294] C. Li, C. Xu, C. Gui, and M. D. Fox, "Level set evolution without re-initialization: a new variational formulation," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 430–436. [Cited on page 128.]
- [295] X. Y. Stella and J. Shi, "Multiclass spectral clustering," in *Computer Vision, IEEE International Conference on*, vol. 2. IEEE Computer Society, 2003, pp. 313–313. [Cited on page 128.]
- [296] S. X. Yu, "Segmentation using multiscale cues," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1. IEEE, 2004, pp. I–I. [Cited on page 128.]
- [297] J. Malik, S. Belongie, T. Leung, and J. Shi, "Contour and texture analysis for image segmentation," *International journal of computer vision*, vol. 43, no. 1, pp. 7–27, 2001. [Cited on page 128.]
- [298] S. Vera, D. Gil, A. Lopez, and M. González, "Multilocal creaseness measure," 2012. [Cited on page 128.]
- [299] M. Sarifuddin and R. Missaoui, "A new perceptually uniform color space with associated color similarity measure for content-based image and video retrieval," in *Proc. of ACM SIGIR 2005 workshop on multimedia information retrieval (MMIR 2005)*, 2005, pp. 1–8. [Cited on page 128.]
- [300] W. Tao, H. Jin, and Y. Zhang, "Color image segmentation based on mean shift and normalized cuts," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 5, pp. 1382–1389, 2007. [Cited on page 128.]
- [301] R. Miyaki, S. Yoshida, S. Tanaka, Y. Kominami, Y. Sanomura, T. Matsuo, S. Oka, B. Raytchev, T. Tamaki, T. Koide *et al.*, "A computer system to be used with laser-based endoscopy for quantitative diagnosis of early gastric cancer," *Journal of Clinical Gastroenterology*, vol. 49, no. 2, pp. 108–115, 2015. [Cited on page 129.]

- [302] X. Liu, C. Wang, J. Bai, G. Liao, and Y. Zhao, "Hue-texture-embedded region-based model for magnifying endoscopy with narrow-band imaging image segmentation based on visual features," *Computer methods and programs in biomedicine*, vol. 145, pp. 53–66, 2017. [Cited on page 130.]
- [303] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on image processing*, vol. 10, no. 2, pp. 266–277, 2001. [Cited on page 130.]
- [304] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations," *Journal of computational physics*, vol. 79, no. 1, pp. 12–49, 1988. [Cited on page 130.]
- [305] J. Ding, J. Shen, H. Pang, S. Chen, and J. Yang, "Exploiting intensity inhomogeneity to extract textured objects from natural scenes," in *Asian Conference on Computer Vision*. Springer, 2009, pp. 1–10. [Cited on page 130.]
- [306] P. Mattila, *Geometry of Sets and Measures in Euclidean Spaces*. Cambridge University Press, 1995. [Cited on page 134.]
- [307] S. R. Messer, J. Agzarian, and D. Abbott, "Optimal wavelet denoising for phonocardiograms," *Microelectronics journal*, vol. 32, no. 12, pp. 931–941, 2001. [Cited on page 138.]
- [308] A. Nobel, "Extreme value theory for the gaussian," University of North Carolina, Technical Report, November 2019. [Online]. Available: <https://nobel.web.unc.edu/wp-content/uploads/sites/13591/2019/11/Gaussian-Extremes-1.pdf> [Cited on page 139.]
- [309] G. E. Hinton and S. Roweis, "Stochastic neighbor embedding," *Advances in neural information processing systems*, vol. 15, 2002. [Cited on pages xix and 151.]