Streamlining Clinical Pipelines for Cardiovascular Imaging with Deep Learning

MPhil/PhD in Data Science

Tuan Aqeel Bohoran
(ID: N0954201)

Under the supervision of
Archontis Giannakidis, Ph.D.

Department of Physics and Mathematics
Nottingham Trent University

October 3, 2024

Acknowledgments

I express my deepest gratitude towards my supervisor **Archontis Giannakidis, Ph.D.** for the constant help and encouragement from the start of the thesis work.
I have been fortunate to have a supervisor who gave me the freedom to explore on my own and at the same time helped me plan the future work-plan with timely reviews and suggestions wherever required. Thank you for giving me enough extension of time whenever required and having faith in me.

Additionally, I wish to acknowledge the contributions of my external co-supervisor, **Professor Gerry P. McCann**, from Glenfield Hospital in Leicester, UK and my collaborator, **Polydoros N. Kampaktsis, MD, PhD**, from Columbia University Irving Medical Center (CUIMC), USA. Their expertise and collaborative spirit enriched this project significantly. I would also like to thank my co-supervisors **Jonathan Crofts** and **Laurence Shaw**.

My heartfelt thanks also go to: (i) **my family**, especially my father Faleel Bohoran, my mother Rishani Bohoran, my elder brother Imran Bohoran, my elder sister Mumtaz Bohoran, my younger sister Aneesha Bohoran, my younger brother Ameen Bohoran and my cousin Arshad Kitchilan, (ii) **my teacher** Kamal Perera (iii) **my friend** Thanos Siouras, and (iv) **my colleagues** at Nottingham Trent University (NTU) (David Jenkins, Suliman Almansour) and Glenfield Hospital (Kelly Parke, Alice Cowell). Your unwavering support, discussions, and camaraderie have made this academic journey more meaningful.

Lastly, I express my gratitude to the **EU Horizon programme**, **the Imaging, Materials and Engineering Centre at NTU**, and **the School of Science and Technology at NTU** for their financial support and resources.

This work would not have been possible without the collective efforts of all those mentioned above. Thank you for being part of this remarkable journey.

Nottingham, UK
October 3, 2024

Tuan Aqeel Bohoran

List of Publications

[1] **T. A. Bohoran**, K. S. Parke, M. P. M. Graham-Brown, M. Meisuria, A. Singh, J. Wormleighton, D. Adlam, D. Gopalan, M. J. Davies, B. Williams, M. Brown, G. P. McCann, and A. Giannakidis (2023). Resource efficient aortic distensibility calculation by end to end spatiotemporal learning of aortic lumen from multicentre multivendor multidisease CMR images. Scientific Reports, vol. 13, 21794.

[2] P. N. Kampaktsis , **T. A. Bohoran**, M. Lebehn, L. McLaughlin, J. Leb, Z. Liu, S. Moustakidis, A. Siouras, A. Singh, R. T. Hahn, G. P. McCann, A. Giannakidis (2024). An attention-based deep learning method for right ventricular quantification using 2D echocardiography: Feasibility and accuracy. Echocardiography, Volume 41, Issue 1, e15719. January 2024.

[3] **T. A. Bohoran**, P. N. Kampaktsis, L. McLaughlin, J. Leb, S. Moustakidis, G. P. McCann, A. Giannakidis (2023). Right Ventricular Volume Prediction by Feature Tokenizer Transformer-Based Regression of 2D Echocardiography Small-Scale Tabular Data. In: Bernard, O., Clarysse, P., Duchateau, N., Ohayon, J., Viallon, M. (eds) Functional Imaging and Modeling of the Heart (FIMH 2023), June 19 – 22, Lyon, France. Lecture Notes in Computer Science, vol 13958. Springer, Cham, pp. 292–300.

[4] **T. A. Bohoran**, P. N. Kampaktsis , G. P. McCann, A. Giannakidis (2023). Fast-tracking the deep residual network training for arrhythmia classification by leveraging the power of dynamical systems. In Proceedings of the 17th IEEE International Conference on Signal Image Technology & Internet Systems (IEEE SITIS 2023), November 8 - 10, Bangkok, Thailand.

[5] **T. A. Bohoran**, P. N. Kampaktsis, L. McLaughlin, J. Leb, S. Moustakidis, G. P. McCann, A. Giannakidis (2023). Embracing uncertainty flexibility: Harnessing a supervised tree kernel to empower ensemble modelling for 2D echocardiography-based prediction of right ventricular volume. In the Proceedings of the 16th International Conference of Machine Vision (ICMV 2023), November 15 – 18, Yerevan, Armenia. SPIE, Vol. 13072, 1307214.

[6] P. Kampaktsis, **T. A. Bohoran**, L. McLaughlin, A. Singh, J. Leb, R. T. Hahn, G. McCann, A. Giannakidis (2024). 2D echocardiographic right ventricular quantification using planimetry and deep learning. In the 2024 American College of Cardiology Annual Scientific Session (ACC 2024), Atlanta, USA, April 6 – 8. Journal of the American College of Cardiology, Volume 83, Issue 13, Supplement A, 1522.

[7] P. N. Kampaktsis, **T. A. Bohoran**, M. Lebehn, L. McLaughlin, J. Leb, S. Moustakidis, R. T. Hahn, A. Giannakidis (2024). A 2D echocardiographic method for accurate right ventricular quantification using deep learning. In the 2024 European Society Cardiology Congress (ESC 2024), London, UK, August 30 – September 2.

[8] **T. A. Bohoran**, K. S. Parke, A. Cowley, G. S. Gulsin, J. Yeo, A. Dattani, G. P. McCann, and A. Giannakidis (2024). Segmetron: Sample-efficient Model-agnostic Cardiac Semantic Segmentation with a Trustworthy Reject Option via PQ Learning, IEEE Journal of Biomedical and Health Informatics, (submitted, under review).

Contents

List of Tables

List of Figures

Style Conventions

$\times$	:	Scalar multiplication.
$\odot$	:	Hadamard (element-wise) product.
$*$	:	Convolution operation.
$\frac{\partial}{\partial x}$	:	Partial derivative with respect to x.
$\int$	:	Integral sign, used to denote integration.
∇	:	Gradient.
∞	:	Infinity symbol, represents an unbounded limit.
$\mathbb{1}$	:	Indicator function.
$\lvert . \rvert$	:	Set cardinality.
$\lfloor . \rfloor$	:	Floor function.
$\lvert . \rvert^{\infty}$	:	Infinity norm.
$\square^T$	:	Transpose operation.
$\bigcap$	:	Intersection operation symbol.
$\forall$	:	For every.
$A \leftarrow \vec{0}$	:	Initialise vector A to zeros.
$[CLS]$	:	Classification token used in Transformer models.
$\texttt{Err}_Q$	:	Error rate on test distribution Q.
$\mathcal{H}_0$	:	The null hypothesis.
$\mathcal{H}_1$	:	The alternative hypothesis.
$\inf()$	:	Mathematical function that returns the infimum of a set.
$\mathcal{L}$	:	Commonly used to represent a loss function.
LayerNorm	:	Layer normalisation.
Linear	:	Linear transformation layer.
$\max()$	:	Mathematical function that returns the maximum of its arguments.
$\min()$	:	Mathematical function that returns the minimum of its arguments.
$\widetilde{\mathcal{O}}$	:	$\mathcal{O}$ notation hiding logarithmic factors.
$P(y \mid x)$	:	Conditional probability that y occurs given that x has already occurred.
$\mathbb{R}$	:	Set of all real numbers.
$\mathbb{R}^w$	:	Space of w-dimensional vectors.
$\mathbb{R}^{h \times w}$	:	Space of $h \times w$ matrices.
$\texttt{Reject}_P$	:	Rejection rate on training distribution P.
stack	:	Stacking function to stack values on the given axis.
$\sup()$	:	Mathematical function that returns the supremum of a set.
tanh	:	Hyperbolic tangent activation function.
$\mathbb{X}_i$	:	i^{th} dimensional input feature space.
$\dot{z}$	:	Differentiation of function z with respect to time.
$\in$	:	Belongs to.
π	:	Mathematical constant Pi.
$\sum$	:	Summation operator.

Abbreviations

2DE	: Two-Dimensional Echocardiography.
3DE	: Three-Dimensional Echocardiography.
AAo	: Ascending Aorta.
AAMI	: Association for the Advancement of Medical Instrumentation.
AD	: Aortic Distensibility.
Adam	: Adaptive Moment Estimation.
ANOVA	: Analysis Of Variance.
APE	: Absolute Percentage Error.
BA	: Bland-Altman.
BConvLSTM	: Bi-directional Convolutional Long Short-Term Memory.
BConvLSTM2D	: Two-Dimensional Bi-directional Convolutional Long Short-Term Memory.
BIH	: Beth Israel Hospital.
BMI	: Body Mass Index.
BN	: Batch Normalisation.
CatBoost	: Categorical Boosting.
CBU	: Categorical Boosting with Uncertainty.
CDF	: Cumulative Distribution Function.
CI	: Confidence Interval.
CMR	: Cardiac Magnetic Resonance Imaging.
CNN	: Convolutional Neural Network.
CO$_2$eq	: Carbon Dioxide Equivalent Emissions.
Conv2D	: Two-Dimensional Convolutional.
ConvLSTM2D	: Two-Dimensional Convolutional Long Short-Term Memory.
CRPS	: Continuous Ranked Probability Score.
CUIMC	: Columbia University Irving Medical Center.
CVA	: Cerebrovascular Accident.
CVD	: Cardiovascular Disease.
DAo	: Descending Aorta.
DBP	: Diastolic Blood Pressure.
DCE	: Disagreement Cross Entropy.
DL	: Deep Learning.
DM	: Diabetes Mellitus.
DSC	: Dice Similarity Coefficient.
DTW	: Dynamic Time Warping.
ECG	: Electrocardiogram.
ED	: End Diastole.
EDS	: Enforced Disagreement Segmenter.
EF	: Ejection Fraction.
ES	: End Systole.
FAC	: Fractional Area Change.

FN	False Negative.
FourC	Standard Four Chamber.
GBRT	Gradient Boosted Regression Tree.
IBUG	Instance-Based Uncertainty Quantification for GBRTs.
ICC	Intraclass Correlation Coefficient.
IS	Interval Score.
KDE	Kernel Density Estimation.
KS	Kolmogorov–Smirnov.
LightGBM	Light Gradient-Boosting Machine.
LSTM	Long Short-Term Memory.
MAE	Mean Absolute Error.
MAPE	Mean Absolute Percentage Error.
MIT	Massachusetts Institute of Technology.
ML	Machine Learning.
MLE	Maximum Likelihood Estimation.
MMD	Maximum Mean Discrepancy.
MMD-D	Deep Kernel Maximum Mean Discrepancy.
MSE	Mean Squared Error.
NGBoost	Natural Gradient Boosting.
NLL	Negative Log-Likelihood.
NLP	Natural Language Processing.
NN	Neural Network.
ODE	Ordinary Differential Equation.
OOD	Out-Of-Distribution.
PGBM	Probabilistic Gradient Boosting Machine.
PLAX	Parasternal Long Axis.
PP	Pulse Pressure.
PPE	Parameters Per Epoch.
PSAXAV	Parasternal Short Axis at the Level of the Aortic Valve.
PSAXbase	Parasternal Short Axis at the Base of the Left Ventricle.
PSAXdistal	Parasternal Short Axis at the Apex of the Left Ventricle.
PSAXmid	Parasternal Short Axis at the Mid Left Ventricle.
ReLU	Rectified Linear Unit.
ResNet	Residual Network.
RFI	Relative Feature Importance.
RMD	Relative Mahalanobis Distance.
RMSE	Root Mean Squared Error.
ROI	Region Of Interest.
RV	Right Ventricular.
RVEDV	Right Ventricular End-Diastolic Volume.
RVEF	Right Ventricular Ejection Fraction.
RVESV	Right Ventricular End-Systolic Volume.
RVInflow	Right Ventricular Inflow.
RVS	Maximum Tissue Doppler Velocity in Systole.
SBP	Systolic Blood Pressure.
SC	Selective Classification.
SD	Standard Deviation.
SGD	Stochastic Gradient Descent.
SOTA	State-of-the-art.
SSFP	Steady-State Free Precession.
STFT	Short-Time Fourier Transform.

SubC	:	Subcostal Four Chamber.
TabTransformer	:	Tabular Transformer.
TAPSE	:	Tricuspid Annular Plane Systolic Excursion.
TE	:	Echo Time.
TP	:	True Positive.
TPR	:	True Positive Rate.
TR	:	Repetition Time.
TRE	:	Tricuspid Regurgitation.
TTE	:	Transthoracic Echocardiography.
VC	:	Vapnik–Chervonenkis.
XGBoost	:	Extreme Gradient Boosting.
YOLOv3	:	You Only Look Once Version 3.

Latin Symbols

$\mathcal{A}$	: 1st curve in the definition of Fréchet distance.
$A(x_i, x_{te})$	: Affinity of a training example x_i to a target example x_{te}.
$\widetilde{A}$	: Aortic lumen cross-sectional area, mm^2.
A_{DSC}	: Segmented region in dice similarity coefficient.
A_{HD}	: 1st non-empty subset of M in the Hausdorff distance definition.
A_{ts}	: 1st temporal sequence in the dynamic time wrapping distance definition.
$\widetilde{A}_{\max}$	: Maximum aortic lumen cross-sectional area, mm^2.
$\widetilde{A}_{\min}$	: Minimum aortic lumen cross-sectional area, mm^2.
$A^{(k)}$	: Vector containing the top k affinity scores from the vector A.
a_{HD}	: Starting point in Hausdorff definition.
$\mathcal{B}$	: 2nd curve in the definition of Fréchet distance.
B_{DSC}	: Ground truth region in Dice similarity coefficient.
B_{HD}	: 2nd non-empty subset of M in the Hausdorff distance definition.
B_i	: Bias for the i^{th} feature.
B_{ts}	: 2nd temporal sequence in the dynamic time wrapping distance definition.
$b_{\text{BConvLSTM2D}}$	: Bias term of the BConvLSTM2D output.
b_C	: Bias term in a BConvLSTM2D layer.
b_F	: Bias term in a BConvLSTM2D layer.
b_{HD}	: Closest point to a_{HD} in Hausdorff distance definition.
b_I	: Bias term in a BConvLSTM2D layer.
b_O	: Bias term in a BConvLSTM2D layer.
C	: Number of classes.
C_R	: Number of calibration rounds.
C_t	: Cell state tensor at time t of a BConvLSTM2D layer.
C_{t-1}	: Cell state tensor in the previous time step in a BConvLSTM2D layer.
D	: Dataset.
D_{CRPS}	: Forecasted distribution in the CRPS definition.
D_{dist}	: Target distribution in IBUG.
D_n	: Kolmogorov–Smirnov statistic in one-sample test.
$D_{n,m}$	: Kolmogorov–Smirnov statistic in two-sample test.
D_{ts}	: Matrix where each element $D_{ts}(i,j)$ represents the cumulative distance or cost of aligning two temporal sequences up to the i^{th} and j^{th} elements, respectively.
D_{val}	: Validation subset of the dataset.
$\hat{D}_{te}$	: Predicted conditional distribution for a test instance using IBUG.
$\hat{D}^k_{y_j}$	: Predicted conditional distribution for the j^{th} validation instance using the top k affinity scores.
DPR	: Dropout Rate.
d_{BAi}	: Difference between two individual measurements in Bland-Altman analysis.

d_{FD} : Distance in metric space M between two continuous functions.

d_{HD} : Metric on M in the Hausdorff distance definition.

$d_{ts}(a_i, b_j)$: Euclidean distance between the points a_i and b_j.

$\bar{d}_{BA}$: Mean difference betweenthe two measurements in Bland-Altman analysis.

$\mathcal{E}$: Residual module.

E : Ensemble size.

e_i : One-hot encoded vector for the i^{th} categorical feature.

e_m : Maximum number of epochs.

F : GBRT predictive model.

$F_0(x)$: Commencing base learner of a GBRT.

F_D : CDF of D_{CRPS}.

$F_{m-1}(x)$: GBRT model's prediction on x before adding the m^{th} tree.

$F_n(x)$: Empirical CDF in KS test.

F_t : Forget gate at time t of a BConvLSTM2D layer.

$\acute{F}(x)$: Reference distribution CDF in one-sample KS test.

$\hat{F}(\mathcal{A}, \mathcal{B})$: Fréchet distance between curves $\mathcal{A}$ and $\mathcal{B}$.

f_B : Pre-trained baseline semantic segmentation model.

$f_i(Xf_i)$: Function that transforms the i^{th} feature value into an embedding.

f_P : Ensemble model fine-tuned to disagree with f_B on $\mathbb{P}^{\star}$.

$f_P^{(i)}$: i^{th} segmenter in ensemble f_P.

$f_{PD}(x_i|\theta)$: Probability density or mass function of observation x_i with θ fixed.

f_Q : Ensemble model fine-tuned to disagree with f_B on $\mathbb{Q}$.

$f_Q^{(i)}$: i^{th} segmenter in ensemble f_Q.

$\text{filter}(m_h, n_w)$: Filter weight at (m_h, n_w).

$G_m(x)$: Empirical CDF in two-sample KS test.

g_{0i} : It is 1 if pixel i is aortic vessel and 0 if it is background.

g_{1i} : It is 0 if pixel i is aortic vessel and 1 if it is background.

g_i : Forecast probability for the i^{th} instance in the check score definition.

H_{HD} : Hausdorff distance.

H_{t-1} : Hidden state tensor in the previous time step in an BConvLSTM2D layer.

$\overleftarrow{H}_j$: Backward hidden state tensor at time step j.

$\overrightarrow{H}_j$: Forward hidden state tensor at time step j.

$h_m(x)$: Prediction of the m^{th} tree in a GBRT.

$h|_S$: Selective classifier.

I_t : Input gate at time of a BConvLSTM2D layer.

I_l^t : Set of training instances that fall into leaf l of tree t.

i_h : Output tensor's height index in a Conv2D layer.

i_p : Output tensor's height index in a UpSampling2D layer.

J : Objective function of Adam optimiser.

j_p : Output tensor's width index in a UpSampling2D layer.

j_w : Output tensor's width index in a Conv2D layer.

K_{cand} : List of candidate values for k.

k : Number of the most frequent training samples to estimate uncertainty.

k_p : Output tensor's channel index in a UpSampling2D layer.

$\mathcal{L}$: Likelihood function.

L : Loss function.

$\widetilde{L}$: PQ learner.

L_A : Learning algorithm.

L_{DCE} : Disagreement cross entropy loss.

L_{EDS} : Enforced disagreement segmenter overall loss function.

L_{RN}	: Depth of the residual network.
L_{trad}	: Loss function component to agree on$P$.
LT	: Number of Transformer layers.
M	: Metric space.
M_e	: Evaluation metric.
M_{GBRT}	: Number of trees in GBRTs.
m_*	: 1st moment vector estimates at time * in Adam optimisation variants.
$\hat{m}_*$	: Bias-corrected estimates of m_*.
m_h	: Filter/kernel tensor's height index.
N	: Number of training samples.
N_S	: Number of test pixels.
n	: Number of observations.
n_{BA}	: Number of observations in Bland-Altman analysis.
n_w	: Filter/kernel tensor's width index.
O_t	: Output gate at time t of a BConvLSTMD layer.
o_i	: Binary observation of the i^{th} instance in the check score definition.
$\mathbb{P}$	: Labelled dataset sampled from distribution P.
$\mathbb{P}^\star$	: Data taken from distribution P, unseen by f_B .
$\hat{\mathbb{P}}^*$	: Labelled $\mathbb{P}^*$ using f_B pseudo-labels.
$\mathbb{P}_{train}$	: f_B training data sampled from distribution P.
$\mathbb{P}_{val}$	: f_B validation data sampled from distribution P.
P	: Distribution of input samples in training time.
$P_s(X)$	: Marginal distribution of X in the source domain.
$P_s(Y)$	: Marginal distribution of Y in the source domain.
$P_s(Y\|X)$	: Conditional distribution of Y given X in the source domain.
$P_t(X)$	: Marginal distribution of X in the target domain.
$P_t(Y)$	: Marginal distribution of Y in the target domain.
$P_t(Y\|X)$	: Conditional distribution of Y given X in the target domain.
$\widetilde{P}$	: Predicted probability distribution in the interval score definition.
p_{0i}	: Probability that pixel i belongs to the aorta.
p_{1i}	: Probability of pixel i being in the background class.
$\mathbb{Q}$	: Testing data from distribution Q.
$\hat{\mathbb{Q}}$	: Labelled $\mathbb{Q}$ using f_B pseudo-labels.
Q	: Arbitrary distribution of unlabelled input samples in testing time.
q	: Small positive parameter used as a step size or scaling factor.
R_i	: Rank of the i^{th} positive difference in Wilcoxon signed-rank test with Bonferroni correction.
$R_m(x_i)$	: Leaf within the m^{th} tree to which x_i is allocated.
r	: Pearson's correlation coefficient.
r_{im}	: Pseudo-residual for the i-th observation at the m^{th} iteration in a GBRT.
r_m^j	: j^{th} leave in m^{th} iteration of a GBRT.
r_{tol}	: Predetermined risk tolerance for the Lipschitz constant.
S	: Subset of X where $h\|_S$ predicts on.
$\widetilde{S}$	: Check score.
S_i	: Number of unique categories for the i^{th} categorical feature.
S_j^k	: Probabilistic score for the j^{th} validation instance using the top k affinity scores.
s_d	: Standard deviation of the differences in Bland-Altman analysis.
s_p	: Upsampling factor.
T	: Token matrix resulting from the transformation of input features.
T^+	: Test statistic in Wilcoxon signed-rank test with Bonferroni correction.

T_0	:	Initial token matrix with the [CLS] token appended.
T_{cur}	:	Current epoch.
T_{evol}	:	Evolution time.
T_{grow}	:	Epoch at the last growth occurrence.
T_i	:	Output of the i^{th} Transformer layer.
$T_{i(cat)}$	:	Embedding for the i^{th} categorical feature.
$T_{i(num)}$	:	Embedding for the i^{th} numerical feature.
T_L	:	Output of the cascaded Transformers.
$T_L^{[CLS]}$	:	Final representation of the [CLS] token after all Transformer layers.
T_{tot}	:	Total number of epochs.
t	:	Time step of Adam optimisation variants.
V	:	Validation score in IBUG accelerated tuning of k.
V_i	:	i^{th} Transformer layer.
v_*	:	2nd moment vector estimates at time $*$ in Adam optimisation variants.
$\hat{v}_*$	:	Bias-corrected estimates of v_*.
$W+$	:	Sum of the positive ranks in the Wilcoxon signed-rank test.
$W-$	:	Sum of the negative ranks in the Wilcoxon signed-rank test.
$W_{i(cat)}$	:	Weight matrix for the i^{th} categorical feature.
$W_{i(num)}$	:	Weight vector for the i^{th} numerical feature.
W_{H*}	:	2D convolution kernel corresponding to hidden states.
W_{X*}	:	2D convolution kernel corresponding to input states.
$W_y^{\overleftarrow{H}}$	:	Weight matrix for backward hidden states.
$W_y^{\overrightarrow{H}}$	:	Weight matrix for forward hidden states.
w_j	:	Weights in a residual network associated with the j^{th} layer.
$w(t)$	:	Weights of the network as a function of time t.
X	:	p-dimensional instance space.
X_0	:	Observed data in the likelihood function.
$X_{feature}$	:	Input features to Feature Tokeniser.
Xf_i	:	i^{th} element of $X_{feature}$.
$Xf_{i(cat)}$	:	i^{th} categorical feature.
$Xf_{i(num)}$	:	i^{th} numerical feature.
X_t	:	Input to the BConvLSTM2D layer at time step t.
x	:	Empirical observation in the CRPS definition.
$x1_i$	:	1st sample point in Pearson's correlation coefficient.
$x2_i$	:	2nd sample point in Pearson's correlation coefficient.
x_{BA}	:	1st measurement in Bland-Altman analysis.
x_i	:	Input vector for the i^{th} instance in D.
x_i^j	:	j^{th} feature of x_i.
x_{input}	:	Input to the ReLU function.
x_{te}	:	Test set instance.
$\overline{x1}$	:	Mean of $x1_i$.
$\overline{x2}$	:	Mean of $x2_i$.
Y	:	Target space.
Y_j	:	Output of the BConvLSTM2D at time step j.
Y_{pred}	:	Predicted output in Feature Tokeniser Transformer.
y	:	Observed outcome in the interval score definition.
$\bar{y}$	:	Mean value of the observed data.
$\widetilde{y}$	:	Prediction by the ensemble segmenter.
y_i	:	Corresponding target (ground truth) value for the i^{th} input instance (x_i).
$\hat{y}_i$	:	i^{th} prediction by a model.

y_{BA}	: 2nd measurement in Bland-Altman analysis.
z_0	: Input feature map.
z_j	: Input to the j^{th} layer in a residual network.
z_{j+1}	: Output of the j^{th} layer in a residual network.
$z(t)$	: Feature map as a function of time t.

Greek Symbols

α	: Statistical significance level.
α_{FD}	: 1st continuous function from $[0, 1]$ in the Fréchet distance definition.
α_{FT}	: Variable in Tversky Loss which control the magnitude of the penalties for false positives.
β_1	: Exponential decay rate for the first moment estimates in Adam optimisation variants.
β_2	: Exponential decay rate for the second moment estimates in Adam optimisation variants.
β_{FD}	: 2nd continuous function from $[0, 1]$ in the Fréchet distance definition.
β_{FT}	: Variable in Tversky Loss which control the magnitude of the penalties for false negatives.
γ_{FT}	: Adjustable focussing parameter of Focal Tversky loss.
γ_f	: Tuning parameter used to scale the variance in IBUG.
γ_m	: Scaling factor controlling the contribution of the m^{th} tree in a GBRT.
δ	: Failure probability parameter in PQ learning.
δ_f	: Tuning parameter used to adjust the variance in IBUG.
ϵ	: Tolerance constant.
$\widetilde{\epsilon}$	: Error parameter in PQ learning.
η	: Learning rate.
η_α	: Learning rate or step size of Adam optimiser.
η_{max}	: Maximum learning rate in the learning rate scheduler.
η_{min}	: Minimum learning rate in the learning rate scheduler.
θ	: Parameter set.
θ_*	: Model parameters at time $*$.
θ_m^j	: Parameter associated with the j^{th} leaf at iteration m.
λ	: Tuning parameter in the PQ learning overall objective.
$\mu_{F(x_{te})}$	: Mean of the predicted conditional output distribution for the target instance x_{te}.
ρ_z	: Parameter to account for instances with abnormally low variance.
σ	: Sigmoid activation function.
$\sigma^2_{F(x_{te})}$	: Variance of the predicted conditional output distribution for the target instance x_{te}.
$\widetilde{\phi}(\theta, \lambda)$	: H-Divergence general class of continuous functions.
ϕ_P	: Pixel disagreement rate between f_B and f_P.
ϕ_Q	: Pixel disagreement rate between f_B and f_Q.

Abstract

Cardiovascular diseases continue to be the leading cause of morbidity and mortality worldwide, emphasising the need for accurate diagnosis, efficient monitoring, and timely intervention in managing these conditions. This thesis seeks to address key challenges in the cardiovascular field by proposing novel, resource-efficient computational techniques aimed at improving the accuracy and reliability of cardiovascular assessments through the application of cutting-edge machine learning (ML) and deep learning (DL) approaches.

The first theme introduces a new DL model that automatically segments the aortic lumen from cine cardiovascular magnetic resonance (CMR) images. The model, based on bi-directional ConvLSTM (BConvLSTM) U-Net with densely connected convolutions, provides a fresh perspective on measuring aortic distensibility (AD). It addresses significant challenges in existing methods by using a hierarchical learning framework that efficiently processes spatio-temporal aspects of video inputs. The model combines encoder and decoder feature maps through non-linear functions, and manages the high class imbalance in the data through using an appropriate loss function. The study succeeded in applying the model to a multi-centre, multi-vendor dataset with diverse patient demographics. The results show that the proposed model exceeds the current state-of-the-art methods in accuracy. Moreover, it achieves this with significantly less environmental impact, consuming approximately ∼3.9 times less fuel and generating ∼2.8 times fewer carbon emissions. This model has great potential as a tool for exploring genome-wide associations between AD, aortic areas, and cognitive performance in expansive biomedical databases. The study's focus on energy usage and carbon emissions demonstrates the commitment to sustainable deep learning practices. This research not only enhances the capabilities of CMR-derived aortic stiffness evaluation but also sets the stage for more widely applicable and systematic deep learning-powered methodologies in medical imaging, balancing accuracy with environmental consciousness.

In the ongoing endeavour to predict right ventricular (RV) volume from 2D echocardiography images, there is a lack of accurate methods that take advantage of planimetry data. The second theme delves into this intricate task. It analyses 100 RV volumes, and tabular input information encompasses planimetry data from eight standard echocardiography views, as well as age, gender, and cardiac phase information. We present two ML regression methods herein. The first method utilises a gradient-boosted regression tree (GBRT)-based method, enhanced by a supervised tree kernel, in order to not only forecast RV volume but also estimate uncertainty in predictions. The second method employs a multi-head attention-based transformer (called the feature tokeniser transformer), which tokenises tabular data by distinguishing between numerical and categorical inputs. Our findings indicate that while the initial GBRT-based method shows promise, it exhibits limitations in prediction accuracy. There is a significant improvement in RV volume prediction accuracy with the transformer-based model, surpassing the initial GBRT-based approach. This research highlights the importance of incorporating attention mechanisms and feature tokenising in methodological strategies. Additionally, we conduct a "gain" explainability analysis using the GBRTs, facilitating the development of more clinically viable pipelines.

This will ultimately lead to more refined and accurate predictive models in RV volume analysis based on echocardiography.

Arrhythmia, characterised by irregular heartbeats, poses significant health risks. Residual networks, a subset of DL architectures, have emerged as a potent tool to detect abnormalities in electrocardiogram signal anomalies. However, the enhanced accuracy and capabilities afforded by increasing network depth in these models come at the cost of heightened computational demands, which poses a considerable challenge to their practical applicability. Addressing this critical bottleneck, the third theme presents a methodology for the resource-economical development of ML-enabled systems for arrhythmia detection. The proposed methodology, grounded in the dynamical system perspective of residual networks, initiates the training process with a shallow network and then progressively increases its depth. We rigorously validate the method on the PhysioNet MIT-BIH arrhythmia data set using heartbeat spectrograms as training input. The results show that the proposed training requires a minimum of 39.47% fewer parameters per epoch compared to conventional vanilla training, a feat achieved without sacrificing and potentially improving overall performance. Our findings suggest the methodology not only drastically reduces training time but also promises significant savings in energy consumption and environmental costs, offering a glimpse into a future of more sustainable and resource-efficient machine learning developments in arrhythmia detection.

Semantic segmentation enables a higher level of understanding of visual scenes. It also forms an essential capability towards delivering a plethora of life-changing technologies. However, the discrepancy between the distribution of the input samples used to train the model and the input distribution encountered during testing (commonly known as covariate shift), as well as the lack of trustworthy methods for detecting it, hinder the practical application of semantic segmentation. In the fourth theme, we develop a reliable, sample-efficient, distribution-free and model-agnostic hypothesis test, named the Segmetron, to detect detrimental covariate shift in semantic segmentation. To deal with the intractability of the above problem and deliver strong performance guarantees on unknown arbitrary test distributions, we build upon recent advancements in the PQ learning setting of selective classification, and extend it to a different discriminative model (i.e. segmenters). To assess an unlabelled target domain, Segmetron leverages an existing (but random) pre-trained semantic segmentation model and the labelled samples used to train it. To train the enforced disagreement segmenters to learn the same generalisation region as the pre-trained semantic segmentation model, we propose loss functions (to agree) which are more apropos to the semantic segmentation task. Our approach stands apart from previous studies on semantic segmentation robustness which relied on synthetic domain shifts. Instead, we analyse two real-world covariate shifts from the cardiovascular magnetic resonance imaging field, concerned with binary (aorta, background) and multi-class (left ventricle, right ventricle, myocardium, background) semantic segmentation tasks. Our experiments demonstrate that the Segmetron hypothesis test outperforms other state-of-the-art techniques in terms of statistical power on both semantic segmentation tasks, given access to only a 3D dataset from one patient. This work holds considerable value because it aligns with “Responsible AI” principles and it happens at a time when the machine learning community is striving to increase public trust and acceptance of AI technologies. Moreover, Segmetron has the potential to support the successful deployment of a plethora of semantic segmentation-based transformative AI solutions.

Chapter 1

Introduction

Cardiovascular diseases (CVDs) remain the leading cause of morbidity and mortality worldwide, underscoring the critical importance of accurate diagnosis, effective monitoring, and timely intervention in managing these conditions. The global burden of CVDs necessitates the development of advanced diagnostic tools that are not only precise but also accessible and resource-efficient. This thesis aims to address some of the most pressing challenges in the cardiovascular domain by introducing novel, resource-efficient computational techniques designed to enhance the accuracy and reliability of cardiovascular assessments through the use of state-of-the-art machine learning (ML) and deep learning (DL) methodologies.

The aorta is the body's main artery, carrying oxygen-laden blood to peripheral organs and tissues. The assessment of aortic stiffness, a key indicator of vascular health, has gained prominence as a predictor of cardiovascular events. Aortic distensibility (AD), a measure of aortic stiffness, is typically evaluated using cardiovascular magnetic resonance imaging (CMR), which, despite its accuracy, is limited by its high cost. Moreover, existing image processing-focussed methods for estimating AD from aortic cine CMR images have proven to be suboptimal in the context of modern clinical practice.

Right ventricular (RV) volume assessment plays a pivotal role in evaluating RV size and function, which are crucial for diagnosing and managing a wide spectrum of CVDs. CMR has emerged as the gold standard for quantifying RV volumes. However, access to CMR scanners is severely limited, and the high cost associated with CMR examinations can be a barrier for some healthcare systems and patients. Additionally, CMR scan times can be lengthy. Two-dimensional echocardiography (2DE) emerges as the primary alternative imaging modality for RV evaluation, despite challenges in accurately depicting the complex three-dimensional geometry of the RV. The uncertainty quantification of each ML/DL model for RV volume prediction is crucial for clinicians, as it provides insights into the confidence level they can place in the model's estimates and the associated uncertainty.

The increasing prevalence of arrhythmias, particularly in ageing populations, has highlighted the need for automated and efficient methods for arrhythmia detection. Many people have irregular heartbeats, which can be fatal in some cases. Electrocardiogram (ECG) signals, widely used in clinical practice, provide a non-invasive means of monitoring heart rhythm, but their manual interpretation is labour-intensive and prone to error. Existing literature underscores the pivotal role of network depth in enhancing the arrhythmia detection accuracy and capabilities of residual models. However, the computational demands during training, accompanying an increased depth, pose significant hurdles, limiting such models' practical applicability.

Another critical concern is the presence of covariate shift as well as its detection, particularly in cardiovascular semantic segmentation tasks, which are essential for accurately identifying and understanding anatomical structures. Covariate shift, or the disparity between the input samples used during model training and those encountered during testing or deployment, can significantly degrade model performance. This issue is especially prevalent in medical imaging due to diverse imaging protocols, patient population heterogeneity, and varying conditions across different hospitals and devices.

This thesis aims to address the aforementioned challenges by developing resource-efficient computational models that enhance the accuracy and reliability of cardiovascular assessments. The investigations focus on four critical areas: The calculation of AD using spatio-temporal DL models, the prediction of RV volume (as well as the uncertainty level associated with this prediction) from 2DE data using supervised learning and transformer-based approaches, the acceleration of deep residual network training for arrhythmia classification through dynamical systems theory, and the detection of covariate shift in real-world cardiovascular semantic segmentation tasks Collectively, this thesis contributes to the ongoing efforts to improve cardiovascular diagnostics, enhance patient outcomes, and reduce the environmental impact of computational methods in healthcare. The research presented herein is structured around four primary objectives:

- **Objective 1:** To improve the calculation of AD by developing a novel, resource-efficient spatio-temporal DL model that automates the segmentation of the aortic lumen from CMR images across the entire cardiac cycle. This model addresses the limitations of current semi-automated techniques, which are time-consuming, prone to observer variability, and resource-intensive.
- **Objective 2:** To develop and evaluate ensemble- and transformer-based models for the prediction of RV volumes from 2DE data. By leveraging gradient-boosted regression trees (GBRTs) and tabular feature tokeniser transformers, this research aims to enhance the accuracy of RV volume predictions while also providing uncertainty quantification, thereby improving the clinical utility of 2DE as an alternative to CMR.
- **Objective 3:** To accelerate the training of deep residual networks for arrhythmia classification by employing a dynamical systems approach that dynamically adjusts the network's depth during training. This approach seeks to reduce the computational and environmental costs associated with DL while maintaining high classification accuracy, thereby making automated ECG analysis more feasible for clinical use.
- **Objective 4:** To develop a reliable, sample-efficient, distribution-free, and model-agnostic hypothesis test, named Segmetron, to detect detrimental covariate shift in semantic segmentation tasks from cardiovascular imaging. This research aims to increase the reliability of DL models in real-world clinical settings and align with "Responsible AI" initiatives to improve the trustworthiness of AI solutions in healthcare.

The structure of this thesis reflects the logical progression of research from problem identification and problem formulation to the development and evaluation of novel computational solutions in the cardiovascular domain. The thesis is organised into four main chapters, each aiming to meet one of the above objectives.

Chapter 2

Resource-efficient Aortic Distensibility Calculation by End-to-end Spatio-temporal Learning of Aortic Lumen from Multi-centre Multi-vendor Multi-disease CMR Images

2.1 Introduction

2.1.1 Clinical Background

The aorta is the body's main artery, carrying oxygen-laden blood to peripheral organs and tissues. The efficient operation of the cardiovascular system is highly dependent on the elastic buffer capacity of the aortic wall, which facilitates the conversion of blood flow from a pulsatile form (originating from left ventricular contraction) to a constant form, as required by the periphery [1]. However, it is well-established that the aortic wall's suppleness gradually declines as a natural consequence of ageing [2]. This process can be exacerbated by factors such as hypertension [3], diabetes [4], connective tissue disorders [5], genetic variations in proteins [6], and congenital heart abnormalities [7, 8]. Increased aortic rigidity has been identified as an early indicator of vascular ageing [9] and as a potent standalone predictor of unfavourable cardiovascular events and mortality across diverse populations [10, 11, 12, 13].

Aortic distensibility (AD) serves as a direct measure of aortic stiffness and is defined [14, 15] as the maximum relative change in the aortic lumen cross-sectional area ($\widetilde{A}$) during the cardiac cycle for a given pressure step at a constant vessel length:

$$\text{AD } (10^{-3}\text{mmHg}^{-1}) = \frac{\widetilde{A}_{\text{max}} - \widetilde{A}_{\text{min}}}{\widetilde{A}_{\text{min}} \times \text{PP}} \tag{2.1}$$

In this equation, PP represents the pulse pressure (equal to the systolic blood pressure minus the diastolic blood pressure), while $\widetilde{A}_{\text{max}}$ and $\widetilde{A}_{\text{min}}$ denote the maximum and minimum areas, respectively. Importantly, AD is inversely proportional to the square of the pulse wave velocity [3]. Numerous studies [12, 16, 17, 18] have emphasised the value of AD as an indicative measure of aortic stiffness.

Cardiovascular magnetic resonance imaging (CMR) is widely acknowledged as the premier non-invasive modality for determining ventricular volumes and mass. The application of CMR has expanded to include the calculation of AD from electrocardiogram-gated steady-state free precession (SSFP) cine images acquired in the plane orthogonal to the thoracic aorta at the level of the pulmonary artery bifurcation. Compared to alternative techniques for evaluating aortic rigidity, CMR offers superior high-resolution aorta imaging in both spatial and temporal dimensions. Furthermore, CMR consistently positions the imaging plane orthogonal to the vessel and permits [3] local stiffness analysis at multiple aortic segments within the same examination. This is particularly important when considering the heterogeneous nature of aortic stiffness, as different regions of the aorta may exhibit varying degrees of stiffness due to factors such as age, genetics, or comorbidities. The capacity to accurately measure local stiffness at multiple aortic segments during a single examination offers valuable insights into the spatial distribution of aortic stiffness and, consequently, the underlying mechanisms driving this phenomenon. The validity of aortic stiffness assessment using CMR has been established [19] through comparison with invasive intra-aortic pressure measurements.

Given the importance of aortic stiffness as an early marker of vascular ageing and predictor of adverse cardiovascular outcomes, the development of accurate and non-invasive methods for assessing aortic stiffness has become a crucial area of research. In this context, CMR has emerged as a leading modality, allowing for the detailed examination of AD and its associated factors.

Furthermore, the non-invasive nature of CMR allows for the serial assessment of aortic stiffness over time, providing a unique opportunity to monitor the progression of vascular ageing and evaluate the impact of various interventions on AD. This could prove instrumental in developing novel therapeutic strategies aimed at mitigating or delaying the decline in aortic wall suppleness, ultimately improving cardiovascular health and reducing the risk of adverse events.

The continued refinement and optimisation of CMR techniques and protocols are expected to yield even more accurate and reliable measures of aortic stiffness. Advanced computational approaches, such as machine learning and artificial intelligence, could also be integrated into CMR-based assessments to facilitate the identification of subtle changes in aortic distensibility and improve the prediction of individual risk profiles.

2.1.2 The Challenge

However, employing image processing- methods for estimating AD from aortic cine CMR images has proven to be suboptimal in the context of modern clinical practice [20, 21, 22]. Subsequent to data acquisition, the images are transferred to a processing workstation where the available software performs this calculation by making use of semi-automated methods that rely on CMR experts to manually trace the vessel border of both the ascending (AAo) and descending aorta (DAo) in every 6th frame of the multi-phase CMR examination. Next, the manual contours are propagated by the software across all time frames of the cardiac cycle, a step that also requires user input for the setting of several algorithmic parameters. The manual correction and/or visual validation of the propagated regions of interest (ROIs) is also an indispensable part of the standard operating procedure. Finally, the areas of the aortic lumen masks are calculated, and the minimum and maximum areas, in conjunction with the pulse pressure measurements, are used to calculate the AD, as described above. Another related paper [22] employed a deformable model-based approach for segmenting the aortic lumen which also requires, as part of the initialisation process, the handcrafted definition of an ROI and the centre of the aorta. In addition to being a time-consuming process, the image interpretation stage within these workflows is susceptible to inconsistencies stemming from variations within the evaluations of single observers and discrepancies between multiple observers. Furthermore, obtaining the necessary CMR expertise carries a substantial financial burden. To improve the clinical applicability of CMR-derived

AD measurements and, as a result, facilitate the efficient management of patients who exhibit increased aortic wall stiffness, it is essential to develop swift, fully automated approaches that simultaneously enhance the precision and reliability of aortic lumen area quantification during the cardiac cycle. However, incorporating automation into the cine CMR image interpretation procedure for AD calculation poses significant challenges [23]. These challenges encompass (i) the distinct variations in cross-sectional aortic shape throughout the cardiac cycle among various patients and spanning numerous pathologies and (ii) the fluctuations in aorta brightness due to blood flow. Moreover, the considerable differences in CMR acquisition protocols among diverse studies and institutions further exacerbate the complexity of this endeavour.

2.1.3 Related Work

Deep Learning (DL) [24], or hierarchical representation learning, is a quickly expanding sector of machine learning where models acquire intricate raw-input-data representations tailored to a particular task. Implementing DL models has been transformative [25] in computer vision (as well as numerous other industries), enabling or even exceeding human-level performance in multiple visual tasks, such as image classification, object detection, and semantic segmentation. Recent investigations [26] have underscored the exceptional abilities of DL models for analysing CMR images. Two recent publications [27, 28] utilised DL-based methods to entirely automate aortic lumen segmentation from cine CMR images. However, the two aorta DL research projects viewed the task as a sparse annotation issue, assessing their approach on a minimal number of cardiac cycle time-frames (specifically, end-diastole (ED) and end-systole (ES)) with available ground truth labels. The objectives for remaining time-frames (also employed for input data mapping during training) could not be visually verified and were obtained through pipelines susceptible to registration errors or active contour poor convergence and initialisation failures. Furthermore, prior studies either entirely disregarded [28] the temporal continuity inherent in cine image sequences by addressing the image segmentation task as static or emulated [27] time usage by stacking the recurrent component after the convolutional layers. Nonetheless, correlated spatio-temporal features cannot be learned when spatial and temporal features are explicitly determined in separate network regions [29]. Additionally, prior publications merged feature maps from the encoder and decoder using basic concatenation. However, it has been contended [30] that such an approach yields less accurate segmentation than employing non-linear functions for this task. Subsequently, the loss function utilised by previous research overlooked the fact that the ROI class is considerably smaller than the background class. Furthermore, prior publications examined datasets procured by adhering to a single data acquisition protocol on a relatively healthy cohort. Lastly, prior efforts entirely neglected the resource efficiency aspect of the suggested pipelines. Nonetheless, this is a crucial [30] concern as the significant computation and energy requirements of DL models coincide with considerable environmental and financial costs. Enhancing algorithm efficiency should be a top priority [30] in DL research, along with accuracy.

2.1.4 Our Contribution

The contributions of this study are:

- We propose to enhance the AD calculation by performing aortic lumen segmentation throughout the cardiac cycle by making use of a novel resource-efficient spatio-temporal DL model, inspired by the bi-directional ConvLSTM (BConvLSTM) U-Net with densely connected convolutions [31].

- This is the first work to perform end-to-end (i.e., over the entire cardiac cycle) hierarchical learning and testing of the aortic lumen area from cine CMR images.
- Our approach joins the temporal with the spatial processing of the video input by merging the encoder and decoder feature maps through a BConvLSTM [32] (non-linear) unit.
- We employ the focal Tversky loss [33] during training which is better suited for problems with a high class imbalance in the data.
- We use multi-centre multi-vendor data from a highly heterogeneous patient cohort which significantly adds to the generalisation power of the proposed aortic lumen segmentation algorithm.
- We show that the proposed network outperforms SOTA methods in terms of segmentation accuracy.
- The network we propose in this study is resource-efficient in helping promote environmentally friendly and more inclusive DL research and practices.
- To examine the impact brought by each contributing factor, we perform ablation studies.

2.2 Materials and Methods

2.2.1 Study Population and Image Dataset

The study population comprises participants from four clinical studies analysed at the University Hospitals of Leicester NHS Trust MRI core lab, including AD assessment. These included participants with spontaneous coronary artery dissection [34], asymptomatic type 2 diabetes (from Lydia [35] and DIASTOLIC [36] trials), hypertension (the Pathway 2 study) and healthy volunteers (recruited in studies [34] and [35]). In total, we analysed 424 aortic MRI datasets taken from 376 patients. The number of datasets is greater than the number of patients because patients' re-evaluations were also included. The participants' demographic, anthropometric and clinical characteristics are presented in Table 2.1. The UK National Research and Ethics Service approved each study, and written informed consent was obtained from all subjects prior to participation. All methods were performed in accordance with the relevant guidelines and regulations.

2.2.2 Image Acquisition

Scans were performed in three centres employing four distinct MRI scanners: Leicester (Siemens Aera, 1.5T and Skyra, 3T), Cambridge (GE Signa, 1.5T), and Dundee (Siemens TrioTim, 3T). SSFP cine images of the AAo and DAo in a plane perpendicular to the thoracic aorta at the pulmonary artery bifurcation level were reconstructed into 30-40 phases, as previously detailed [20, 37]. The typical image matrix was 256×186 to 256 pixels. The in-plane pixel height and width varied between 1.093mm and 1.914mm. Concurrently, brachial blood pressure was measured to ascertain pulse pressure.

2.2.3 Data Pre-processing and Annotation

The end-to-end data annotation was carried out semi-automatically by three experts from the Glenfield Hospital in Leicester using the Java Image Manipulation Software Version 6 (Xinapse

Table 2.1: Patient characteristics for the aortic distensibility study.

	(n=376)
Age, mean (±SD), y	48 (±8)
Male, No. (%)	150 (40)
Female, No. (%)	226 (60)
BMI, mean (±SD), kg/m^2	31.74 (±7.41)
SBP, mean (±SD), mmHg	127.33 (±17.44)
DBP, mean (±SD), mmHg	79 (±12.59)
Hypertension, No. (%)	197 (52)
Smoking, No. (%)	192 (51)
History of CVA, No. (%)	2 (1)
History of DM, No. (%)	208 (55)
Renal impairment, No. (%)	78 (21)

BMI: Body Mass Index, SBP: Systolic Blood Pressure, DBP: Diastolic Blood Pressure, CVA: Cerebrovascular Accident, DM: Diabetes Mellitus.

Systems Ltd, Essex, UK) [38] as previously described [20, 37]. The experts were blinded to the patients' details. All MRI images and masked images (annotations) were zero-padded to 256×256 pixels so that their dimensions match.

2.2.4 Neural Network Architecture

Before we discuss the neural network (NN) architecture, we shall present its elements and functions.

Conv2D Layer

The two-dimensional convolution (Conv2D) layer is a fundamental building block in many state-of-the-art (SOTA) convolutional neural network (CNN) architectures. This layer applies a two-dimensional convolution operation to the input data, which involves sliding a filter (or kernel) over the input data and computing a dot product between the filter weights and the corresponding input values.

Mathematically, the operation of a Conv2D layer can be expressed as follows [24]:

$$\text{output}(i_h, j_w) = \sum_{m_h} \sum_{n_w} \text{input}(i_h + m_h, j_w + n_w) \odot \text{filter}(m_h, n_w), \tag{2.2}$$

where:

- i_h, j_w are the height and width indices for the output tensor, respectivly, and m_h, n_w are the corresponding indices for the filter tensor;

- input$(i_h + m_h, j_w + n_w)$ is the input value at the $(i_h + m_h, j_w + n_w)^{th}$ position;

- filter(m_h, n_h) is the filter weight at the $(m_h, n_w)^{th}$ position;

- $\odot$ denotes Hamard (element-wise) multiplication.

The output of the Conv2D layer is a feature map, which represents a higher-level abstraction of the input image. This feature map can then be fed to another Conv2D layer or a fully connected layer for further processing.

Investigations show that the Conv2D layer is a powerful tool for improving the accuracy of DL models. For example, in a study on object detection, using a Conv2D layer in a You Only Look Once Version 3 (YOLOv3) model resulted in improved detection performance compared to using a fully connected layer [39]. In another study, using a Conv2D layer in a DL model for detecting diabetic retinopathy from retinal images resulted in improved classification accuracy [40].

MaxPooling2D Layer

The MaxPooling2D layer is primarily utilised for image tensor down-sampling operations while retaining important features of tensors. Down-sampling is done by striding a pooling window from the top right to the bottom left in an image tensor. In each stride, the pooling window selects the maximum value within the pooling window and projects it as a new image tensor with reduced dimensions. The new reduced dimensions can be calculated as below:

$$\text{output height} = \frac{\text{height} - \text{pooling size} + 2\ (\text{padding size})}{\text{stride}} + 1, \tag{2.3}$$

$$\text{output width} = \frac{\text{width} - \text{pooling size} + 2\ (\text{padding size})}{\text{stride}} + 1. \tag{2.4}$$

ReLU Layer

A ReLU (Rectified Linear Unit) layer [41] is an activation function commonly used in artificial NNs, particularly in DL models. It introduces non-linearity into the network, enabling the model to learn complex patterns and representations from the input data.

The operation of a ReLU layer can be defined mathematically as follows:

$$f(x_{\text{input}})_{\text{RL}} = \max(0, x_{\text{input}}). \tag{2.5}$$

The ReLU layer takes an input (a scalar value, vector, or tensor) and returns an output using the function f(x). This function sets any negative input values to 0 and keeps positive values unchanged. The ReLU layer processes each input element independently in an element-wise implementation. As a result, the output has the same shape as the input but with negative values replaced by zeros.

This simple operation makes ReLU computationally efficient compared to other activation functions, such as the sigmoid or the hyperbolic tangent (tanh) functions. ReLU helps mitigate the vanishing gradient problem, which occurs when training deep NNs with gradient-based optimisation methods.

Dropout Layer

The Dropout layer [42] is a regularisation technique commonly used in deep NNs to prevent overfitting. This layer randomly drops out (or sets to zero) a fraction of the neurons in the previous layer during each training iteration. This forces the network to learn more robust and generalised features as different subsets of neurons are used for each iteration.

Mathematically, the Dropout layer can be expressed as follows:

$$\text{output} = \frac{1}{1 - \text{DPR}} \times \text{input} \odot \text{mask}, \tag{2.6}$$

where:

- input is the input tensor to the Dropout layer;
- DPR is the dropout rate, which is the probability of dropping out a neuron;
- mask is a binary mask tensor, with the same shape as the input tensor, where each element is either 0 or 1 with a probability of DPR and $1 - \text{DPR}$, respectively;
- output is the output tensor of the Dropout layer.

During training, the binary mask tensor is randomly generated for each batch, and the input tensor is multiplied element-wise by this mask. Scaling is applied to the weights to compensate for the fact that dropout is not used at the test time.

Studies have shown that the Dropout layer is a powerful tool for preventing overfitting in deep neural networks. For example, in a study on image classification, using a Dropout layer in a deep NN improved classification accuracy compared to a network without Dropout [43]. In another study, using a Dropout layer in a recurrent neural network improved the model's performance in speech recognition [44].

BatchNormalisation Layer

By normalising activations, Batch Normalisation (BN) layers [45] help stabilise the distributions of internal activations as the model trains. BN also makes it possible to use significantly higher learning rates, reducing initialisation sensitivity and the sensitivity to initialisation. These effects help accelerate the training, sometimes dramatically [25, 45].

Concatenation Layer

The concatenation layer combines the outputs of multiple layers into a single layer. This layer essentially concatenates the outputs along a specified dimension.

For instance, in image processing, the output of a CNN is usually a feature map. To use these features for further processing or classification, concatenating them with features from another CNN layer, a fully connected layer, or any other type of layer is often necessary.

Research has shown that the concatenation layer is a powerful tool for improving the accuracy of DL models. For example, in a study on facial expression recognition, the concatenation of multiple feature maps from different CNN layers significantly increased accuracy compared to using a single feature map alone [46]. In another study, the concatenation of multiple modalities of MRI data using a 3D CNN architecture improved classification accuracy for Alzheimer's disease detection [47].

UpSampling2D Layer

The UpSampling2D layer is commonly used in deep NNs for image processing tasks such as image segmentation and object detection. This layer increases the resolution of the input feature maps by upsampling them in each dimension, usually by a factor of 2 or 3.

Mathematically, the UpSampling2D layer can be expressed as follows:

$$\mathrm{output}(i_\mathrm{p}, j_\mathrm{p}, k_\mathrm{p}) = \mathrm{input}(\lfloor i_\mathrm{p}/s_\mathrm{p} \rfloor, \lfloor j_\mathrm{p}/s_\mathrm{p} \rfloor, k_\mathrm{p}), \tag{2.7}$$

where:

- input is the input tensor to the UpSampling2D layer;
- s_p is the upsampling factor, usually set to 2 or 3;
- output is the output tensor of the UpSampling2D layer;
- $i_\mathrm{p}, j_\mathrm{p}, k_\mathrm{p}$ represent the indices of the output along the first spatial dimension, second spatial dimension, and channel dimension, respectively;
- $\lfloor . \rfloor$ denotes the floor function.

During upsampling, each pixel in the input tensor is duplicated s_p times in each dimension to create a tensor of increased spatial resolution. Then, each pixel in the output tensor is set to the value of the nearest pixel in the input tensor.

In image processing tasks, the UpSampling2D layer is an effective tool for improving the accuracy in image processing tasks. For example, in an experiment comparing semantic segmentation using a U-Net model with a transposed convolution layer, a layer of UpSampling2D yielded greater segmentation accuracy [48]. In another super resolution study, using an UpSampling2D layer improved image quality [49].

BConvLSTM2D Layer

The two-dimensional long short-term memory (ConvLSTM2D) layer is a type of layer that combines the capabilities of CNNs and long short-term memory (LSTM) networks. It helps process spatio-temporal data, such as videos and sequences of images, where both spatial and temporal features are essential. The bidirectional version of this layer, two-dimensional bi-directional long short-term memory (BConvLSTM2D), has the ability to process the input sequence both forward and backwards, allowing the model to capture information from both past and future frames.

Mathematically, the overall BConvLSTM2D output Y_j at the time step j can be calculated as

$$Y_j = \tanh(W_y^{\overrightarrow{H}} * \overrightarrow{H}_j + W_y^{\overleftarrow{H}} * \overleftarrow{H}_j + b_{\text{BConvLSTM2D}}), \tag{2.8}$$

where $\overrightarrow{H}$ and $\overleftarrow{H}$ denote the forward and backwards hidden state tensors, respectively, $W_y^{\overrightarrow{H}}$ and $W_y^{\overleftarrow{H}}$ denote the forward and backward convolution kernels corresponding to the hidden states, $b_{\text{BConvLSTM2D}}$ represents the bias term, $*$ operator symbolises the convolution operation and the hyperbolic tangent was employed to combine the outputs of the forward and backward paths in a non-linear way. The mathematical equations for obtaining each of the two hidden state tensors are as follows:

$$I_t = \sigma(W_{XI} * X_t + W_{HI} * H_{t-1} + b_I) \tag{2.9}$$

$$F_t = \sigma(W_{XF} * X_t + W_{HF} * H_{t-1} + b_F) \tag{2.10}$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \tanh(W_{XC} * X_t + W_{HC} * H_{t-1} + b_C) \tag{2.11}$$

$$O_t = \sigma(W_{XO} * X_t + W_{HO} * H_{t-1} + b_O) \tag{2.12}$$

$$H_t = O_t \odot \tanh(C_t) \tag{2.13}$$

Here, I_t, F_t, and O_t denote the input, forget, and output gates, respectively, at time t, while C_t and σ represent the cell state at time t and the sigmoid function, respectively, X_t refers to the input at time t (Figure 2.1), $\odot$ represents the Hadamard product. C_{t-1}, H_{t-1} are the cell and hidden state tensors in the previous time step, W_{X*} and W_{H*} are the 2D convolution kernels corresponding to input and hidden states, respectively, and b_I, b_F, b_O, b_C are the bias terms. The input gate governs the information that is retained in the cell state. The forget gate regulates the information that is discarded from the cell state. The output gate controls the information utilised to compute the output of the BConvLSTM2D. The cell state maintains the internal state of the BConvLSTM2D.

During training, the gradients of the loss function with respect to the parameters of the BConvLSTM2D layer are calculated and used to update the weights and biases using an optimiser, such as stochastic gradient descent (SGD) [50] or Adam [51].

Proposed Network

Our proposed network structure (Figure 2.1) takes inspiration from the bi-directional convolutional long short-term memory (BConvLSTM) U-Net with densely connected convolutions [31]. It comprises: (i) a contracting pathway (or encoder) to capture context within the image by converting it into a high-level feature representation and (ii) a symmetrical expanding pathway (or decoder) for interpreting feature maps, facilitating accurate localisation (image location) and generating a full-resolution segmentation map. The encoder contains four down-sampling layers, while the decoder has four up-sampling layers.

Each encoding step consists of a sequence of two convolutional layers (3×3 filters and a ReLU non-linear activation [41]) followed by a BN layer [45] and a 2×2 max-pooling. BN was utilised to hasten the convergence of the optimiser during network training while max-pooling was employed to reduce the dimensions of the feature maps significantly. To reduce over-fitting, dropout regularisation [42] was employed in the last two steps of the contracting path before the BN layer. The number of filters each encoding layer computes over the input doubles at each step ([16, 32, 64, 128, 256]).

Each decoding layer consists of a sequence of two convolutional layers (3×3 filters and a ReLU non-linear activation) apart from the final decoding step, which has five convolutional layers. In each step of the expanding path, the output of the previous layer is passed onto an up-conv layer (i.e. up-sampling function followed by a 2×2 convolution; this process doubles the size of the feature map) and then combined with the corresponding (same-resolution-level) representation in the contracting path using skip connections. Combining these two types of feature maps is a channel-wise concatenation in all steps except for the second up-sampling layer, where we propose to merge them more complexly using a BConvLSTM [32] building block that outputs information about all hidden temporal states. This is to account for the spatio-temporal composition of the input. The number of channels reduces in every step of the expanding path ([256, 128, 64, 32, 16, 1]), whereas the size of the feature maps progressively increases to reach the input size after the final layer.

Our approach deviates from the methodology presented in [31] that inspired this study in the following ways: Since the dataset was time distributed, we utilise a BConvLSTM building block that outputs a sequence of feature maps for all time steps. Moreover, for the purpose of a more efficient architecture, we: (i) decrease the number of filters in the convolutional layers to one-fourth compared to [31], (ii) incorporate BConvLSTM in a single step only, (iii) include only one densely packed convolutional block in the final encoding stage.

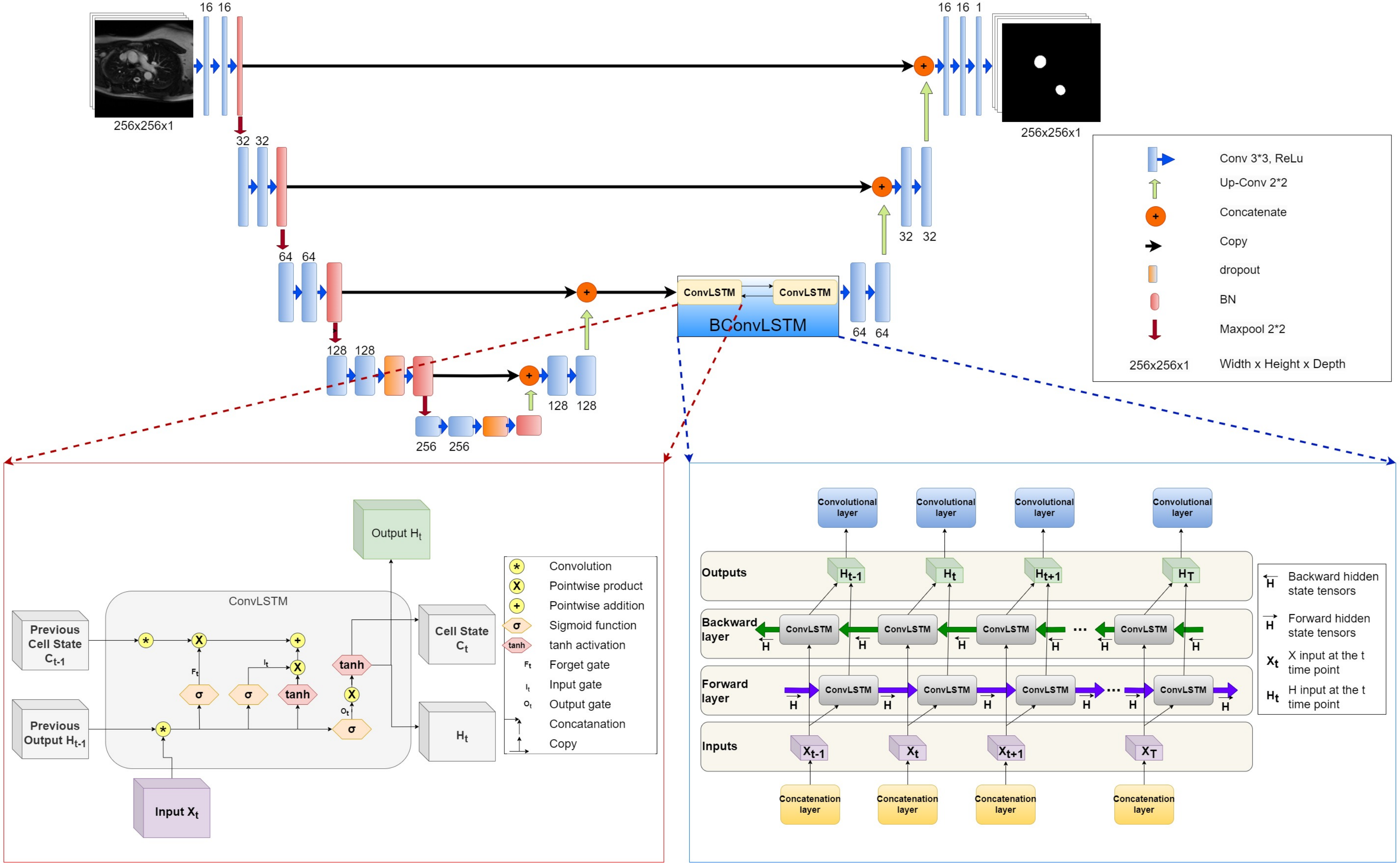


Figure 2.1: The proposed deep learning model, including two diagrams that illustrate how the temporal dimension is handled.

2.2.5 Adam Optimisation

Adam (short for Adaptive Moment Estimation) [51] is a popular optimisation algorithm commonly used in machine learning (ML) and DL. It is an extension of SGD and incorporates adaptive learning rates for each parameter, which can improve convergence speed and robustness.

The update rule for Adam is given by:

$$\begin{aligned}
m_t &= \beta_1 \cdot m_{t-1} + (1-\beta_1)\nabla_\theta J(\theta_{t-1}), \\
v_t &= \beta_2 \cdot v_{t-1} + (1-\beta_2)[\nabla_\theta J(\theta_{t-1})]^2, \\
\hat{m}_t &= \frac{m_t}{1-\beta_1^t}, \\
\hat{v}_t &= \frac{v_t}{1-\beta_2^t}, \\
\theta_t &= \theta_{t-1} - \eta_\alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}.
\end{aligned}$$

where:

- t is the time step;
- θ_* represents the model parameters at time $*$;
- m_* and v_* are the first and second moment vector estimates at time $*$, respectively;
- $\hat{m}_*$ and $\hat{v}_*$ are the bias-corrected estimates of m_* and v_* ;
- J is the objective function;
- η_α is the learning rate or the step size;
- $\nabla_\theta J(\theta_{t-1})$ is the gradient of the objective function with respect to the parameters, assigned the parameter values at the previous time step;
- β_1 and β_2 are the exponential decay rates for the first and second moment estimates, respectively, typically close to 1 but less than 1;
- ϵ is a small constant (usually 10^{-8}) added to prevent division by zero;

Adam combines the advantages of both momentum [52] and RMSProp [53]. It adapts the learning rates for each parameter based on the first and second moments of the gradients, allowing it to converge efficiently and effectively in a wide range of deep learning tasks.

2.2.6 Implementation and Training

The dataset was randomly split into a training set (272 datasets), a validation set (68 datasets) and a testing set (84 datasets). For training, we used the Adam optimiser [51] for 250 epochs with a constant learning rate of 0.001, a decay of 0.0005 and a batch size of 120 (approximately four patients). The dropout value that we used was 0.5 in dropout layers. The initialiser of the network was He Normal [54]. Online data augmentation techniques were used to improve the proposed model's generalisation ability. The augmented data were obtained from the original images by applying random rotations (by a degree between -30° and +30°) and random translations along the x- and/or y-axis in either direction (by up to 20 pixels). All hyper-parameters were tuned using grid-search based on the validation accuracy. To address the severe class imbalance between pixel values 0 and 1 in each frame, we utilised the Focal Tversky loss function [33].

Focal Tversky loss [33] is an extension of the Tversky loss [55], which is a generalisation of the Dice loss used in image segmentation tasks. Focal Tversky loss can be defined as

$$\text{Focal Tversky Loss} = (1 - \text{Tversky Loss})^{\gamma_{FT}} \tag{2.14}$$

where γ_{FT} is the adjustable focussing parameter, and Tversky Loss is the Tversky index given by

$$\text{Tversky Loss} = \frac{\sum_i^N p_{0i} g_{0i}}{\sum_i^N p_{0i} g_{0i} + \alpha_{FT} \sum_i^N p_{0i} g_{1i} + \beta_{FT} \sum_i^N p_{1i} g_{0i}} \tag{2.15}$$

where p_{0i} is the probability that pixel i belongs to the aorta and p_{1i} is the probability of pixel i being in the background class. In addition, g_{0i} is 1 for the aortic vessel segmentation area and 0 for the background, and the opposite is true for g_{1i}. Finally, α_{FT} and β_{FT} are variables in Tversky Loss which control the magnitude of the penalties for false positives and false negatives, respectively. To improve model convergence and the recall rate, we trained our model with $\alpha_{FT} = 0.8$, $\beta_{FT} = 0.8$ and $\gamma_{FT} = 1$.

All the methods were trained and implemented using the TensorFlow framework.

2.2.7 Model Evaluation and Statistical Analysis

Evaluation Methods

In this section, we concisely provide an overview of the evaluation metrics utilised to implement and thoroughly analyse the findings and outcomes of this research effectively.

Bland-Altman analysis: Bland-Altman (BA) analysis [56], also known as a Tukey mean-difference plot, is a statistical method used to visually assess the agreement between two measurements of the same quantity. The method is beneficial when the two measurements are obtained using different methods or instruments, and their agreement needs to be evaluated.

The BA plot shows the difference between the two measurements on the y-axis and the average of the two measurements on the x-axis. The plot also includes three horizontal lines: one representing the mean difference between the two measurements and the other representing the limits of agreement, defined as the mean difference plus or minus two standard deviations of the differences.

The BA analysis allows for the identification of systematic bias between the two methods and the presence of any outliers or trends in the differences. It also provides a measure of the overall variability of the differences.

The mathematics behind the BA analysis is straight-forward. Let the two measurements be denoted as x_{BA} and y_{BA}, and let their mean difference be denoted as $\bar{d}_{BA}$. $\bar{d}_{BA}$ can be calculated as follows:

$$\bar{d}_{BA} = \frac{1}{n_{BA}} \sum_{i=1}^{n_{BA}} d_{BAi}, \tag{2.16}$$

where n_{BA} is the number of observations.

The limits of the agreement are given by:

$$\bar{d}_{BA} \pm 1.96 s_d, \tag{2.17}$$

where s_d is the standard deviation of the differences, which can be calculated as:

$$s_d = \sqrt{\frac{1}{n_{BA} - 1} \sum_{i=1}^{n_{BA}} (d_{BAi} - \bar{d}_{BA})^2}. \tag{2.18}$$

Wilcoxon signed-rank test with Bonferroni correction: The Wilcoxon signed-rank test with Bonferroni correction [57, 58] is a non-parametric statistical hypothesis test used to compare two related samples where the data do not follow a normal distribution or the assumptions of the paired t-test are not met. The test is performed by ranking the absolute differences between the two paired samples and calculating the test statistic, which is the sum of the ranks for the positive differences.

The Bonferroni correction is used to adjust the significance level of the test when multiple tests are performed simultaneously. It is a conservative correction that helps to reduce the likelihood of making a Type I error.

The test is carried out as follows:

1. Calculate the absolute difference between each pair of observations.

2. Rank the absolute differences from smallest to largest, ignoring the sign.

3. Calculate the test statistic, which is the sum of the ranks of the positive differences.

4. Calculate the critical value of the test statistic using a Wilcoxon signed-rank table or a statistical software package.

5. Compare the calculated test statistic with the critical value. If the calculated test statistic exceeds the critical value, reject the null hypothesis. Otherwise, fail to reject the null hypothesis.

6. To adjust for multiple comparisons, divide the significance level (typically 0.05) by the number of tests being performed.

The null hypothesis for the Wilcoxon signed-rank test is that there is no difference between the two related samples. The alternative hypothesis is that there is a difference between the two related samples.

The test statistic can be calculated as follows:

$$T^+ = \sum_{i=1}^{n} R_i, \tag{2.19}$$

where R_i is the rank of the i^{th} positive difference.

The critical value for the test statistic can be obtained from a Wilcoxon signed-rank table or a statistical software package. The Bonferroni correction involves dividing the significance level by the number of tests. If three tests are performed, then the adjusted significance level would be $0.05/3 = 0.0167$. If the calculated test statistic is greater than 0.05, it can be concluded that there is no significant difference between the two related samples. Else, if the test statistic is less than 0.05, it can be concluded that there is a significant difference between the two related samples.

Dice Similarity Coefficient (DSC): The DSC, also known as the Sørensen-Dice coefficient, is a metric used to measure the similarity or overlap between two sets. In medical imaging, it is commonly used to evaluate the performance of image segmentation algorithms by comparing the segmented region to a ground truth region.

The DSC coefficient is defined as:

$$DSC = \frac{2|A_{DSC} \cap B_{DSC}|}{|A_{DSC}| + |B_{DSC}|} \tag{2.20}$$

where A_{DSC} and B_{DSC} are two sets, $|A_{DSC}|$ and $|B_{DSC}|$ denote the cardinality of sets A_{DSC} and B_{DSC}, and $|A_{DSC} \cap B_{DSC}|$ denotes the cardinality of the intersection of sets A_{DSC} and B_{DSC}.

DSC ranges from 0 to 1, where 0 indicates no overlap between the two sets and 1 indicates complete overlap. The closer the Dice coefficient is to 1, the better the segmentation performance.

For example, in medical imaging, if A_{DSC} represents the segmented region and B_{DSC} represents the ground truth region, then the DSC can be used to evaluate the similarity between the two regions. If the DSC is high, the segmentation algorithm accurately identifies the ROI. Conversely, if the DSC is low, the algorithm has failed to accurately identify the ROI.

Fréchet Distance: The Fréchet distance between two curves in a metric space measures their similarity. Informally, one can visualise this distance as the shortest possible leash that connects a dog and its owner if they traverse their respective paths from start to finish without backtracking. Mathematically, let $\mathcal{A}$ and $\mathcal{B}$ be two curves in a metric space M. The Fréchet distance between $\mathcal{A}$ and $\mathcal{B}$, denoted by $\hat{F}(\mathcal{A}, \mathcal{B})$, is defined as:

$$\hat{F}(\mathcal{A}, \mathcal{B}) = \inf_{\alpha,\beta} \max_{t\in[0,1]} d_{FD}(\mathcal{A}(\alpha_{FD}(t)), \mathcal{B}(\beta_{FD}(t))), \tag{2.21}$$

where α_{FD} and β_{FD} are continuous functions from $[0, 1]$ to $[0, 1]$, d_{FD} denotes the distance in M [59], and $\inf()$ denotes the infimum of a set.

Hausdorff Distance: The Hausdorff distance provides a measure of the distance between two point sets within a metric space. Given two non-empty subsets A_{HD} and B_{HD} of a metric space M, the Hausdorff distance H_{HD} is defined as:

$$H_{HD}(A_{HD}, B_{HD}) = \max \left\{ \sup_{a_{HD} \in A_{HD}} \inf_{b \in B_{HD}} d_{HD}(a_{HD}, b_{HD}), \sup_{b_{HD} \in B_{HD}} \inf_{a \in A_{HD}} d_{HD}(a_{HD}, b_{HD}) \right\}, \tag{2.22}$$

where d_{HD} is the metric on M and sup() denotes the supremum of a set. In simpler terms, the Hausdorff distance quantifies the greatest distance from a point a_{HD} in one set to the closest point b_{HD} in the other set [60].

Dynamic Time Warping (DTW) Distance: Dynamic Time Warping (DTW) is a method utilised to measure the similarity between two temporal sequences which might be out of sync, stretched, or compressed. Given two sequences $A_{ts} = \{a_1, a_2, \ldots, a_n\}$ and $B_{ts} = \{b_1, b_2, \ldots, b_m\}$, DTW aims to align them in such a way that the cumulative distance between them is minimised.

The DTW distance is computed using a dynamic programming approach, where the cost at each point (i, j) in a matrix is:

$$D_{ts}(i, j) = d_{ts}(a_i, b_j) + \min\{D_{ts}(i-1, j), D_{ts}(i, j-1), D_{ts}(i-1, j-1)\} \tag{2.23}$$

where $d_{ts}(a_i, b_j)$ represents the distance between the points a_i and b_j, often computed using the Euclidean distance. D_{ts} is a matrix where each element $D_{ts}(i, j)$ represents the cumulative distance or cost of aligning two temporal sequences up to the i^{th} and j^{th} elements, respectively [61].

Evaluation Process

To evaluate the accuracy of our automated segmentation masks against the ground truth, we used the DSC and calculated the absolute area error (in mm^2) and absolute AD error (in mmHg^{-1}). Moreover, we employed BA analysis [62] to assess the agreement between the maximum and minimum aorta areas and AD values for both the AAo and DAo. All analysis was conducted on the test set. The temporal fidelity of the segmentation performance across a cardiac cycle was assessed both qualitatively and quantitatively using the Fréchet, Hausdorff and DTW distances for a representative case.

To examine the impact brought by each contributing factor, we performed ablation studies.

We delved into a series of ablation experiments, meticulously designed to dissect and appreciate the individual contributions of various features within our proposed framework. These experiments are crucial for shedding light on the intricate workings of our model, offering a clear lens through which we can understand the impact of each component on the overall performance.

Our approach dissected the framework into several key variants, each omitting a specific feature or strategy. The variants are as follows:

No full labels: This model variant was trained using labels derived from ES and ED frames, propagated through the cardiac cycle using non-rigid image registration. This approach tests the resilience of the model against partially labelled data.

No focal Tversky: In this iteration, we employed the Dice coefficient loss function, deliberately overlooking the severe class imbalance inherent in our dataset. This allows us to scrutinise the effectiveness of the focal Tversky loss in managing class imbalance.

No non-linearities: Here, the model was trained by simply linearly concatenating the encoder and decoder feature maps, eschewing any non-linear processing. This experiment assesses the impact of non-linear integration of features on the model's performance.

No dense pruning: This version was trained with a more complex structure, incorporating three densely packed convolutional blocks in the final encoding step, as opposed to the simplified structure of the proposed model.

No filter pruning: This model variant utilises a larger number of filters in the convolutional layers, specifically four times more than our proposed model, to evaluate the effects of model complexity on performance.

No BConvLSTM pruning: In contrast to our original design, which uses a single BConvLSTM unit, this model incorporates the BConvLSTM unit at three different stages of the network.

We also measured the resource efficiency of our approach by calculating the carbon dioxide equivalent (CO_2eq) emissions (in g) and energy consumption (in kWh) during training using the Carbontracker method [63]. Carbontracker, a Python-based open-source tool, tracks and predicts training deep learning models' energy consumption and carbon emissions on a given GPU. In addition, we reported the equivalent distance travelled by car [64] that would generate the same emission volume to put the carbon footprint produced during model training into perspective. We also calculated the training and inference times.

Furthermore, we compared our method's evaluation metrics to those of SOTA [27] and unpruned [31] methods, also trained on the same multi-centre, multi-vendor multi-disease CMR dataset following similar training and hyperparamater tuning procedures as described above. To test if there is a statistically significant difference (at level 0.05) between the performances of the proposed and SOTA methods, we used the Wilcoxon signed-rank test with Bonferroni correction.

2.3 Results

2.3.1 Model Accuracy

The aortic area absolute errors (expressed in mm^2) and AD (represented in $mmHg^{-1}$), as well as the Dice coefficients for both AAo and DAo, are presented in Table 2.2. The proposed method attained lower absolute area and AD errors and higher Dice coefficient values in comparison to both the SOTA and unpruned methods.

Figures 2.2, 2.3, and 2.4 elucidate the BA analysis plots for the maximum and minimum area and AD values, contrasting our method, the SOTA, and the unpruned strategies, respectively, against the gold standard. From this visual representation, it is discernible that the aortic metrics projected by our method exhibit around $\sim$3.9 times reduced variability compared to both the SOTA and unpruned methodologies. Figure 2.5 conducts a side-by-side qualitative assessment between our proposal and the SOTA model across three distinctive cases to provide a more granular perspective. The first case illuminates a scenario where the SOTA model significantly undervalues the pixel count assigned to the aorta. The subsequent case profiles a patient for whom the SOTA technique amplifies the AAo segment. The final case underscores a unique occurrence where the DAo's positioning deviates from its typical alignment relative to the AAo, leading the SOTA model to overlook the AAo entirely. In stark contrast, our method meticulously segments both AAo and DAo across these diverse scenarios.

Moreover, Table 2.3 catalogues the qualitative and quantitative juxtapositions of AAo and DAo time-area curves spanning a singular cardiac cycle for a prototypical instance. Upon examination, it is evident that the phase synchronisation of the proposed method's cardiac cycle and the benchmarked curves align more harmoniously than the representations from the SOTA, for both AAo and DAo.

2.3.2 Model Resource Efficiency

Table 2.4 presents the anticipated carbon dioxide equivalent (CO_2eq) emissions (measured in g), the energy expenditure (captured in kWh), and the corresponding distance (in km) a car might traverse during the culminating training phase consisting of 250 epochs for our method, the prevailing SOTA, and the unpruned models. Further detailed in Table 2.4 are the duration of training (represented in hours : minutes : seconds) and the mean inference times (expressed in milliseconds).

Contrasting the environmental impact, our proposed technique emerges as approximately ∼2.8 times more eco-friendly, utilising around ∼3.9 times reduced energy compared to the SOTA model. Furthermore, it boasts efficiency, demanding roughly ∼5.2 times shorter training duration and showcasing inference speeds that are approximately ∼2.7 times brisker. Compared to the unpruned model, our approach displays a marked reduction in environmental footprint, being roughly ∼5.2 times cleaner and conserving about ∼6.6 times the energy. Regarding efficiency metrics, our method necessitates nearly ∼9.3 times shorter training phases and delivers inference outcomes that are approximately ∼4.4 times swifter than its unpruned counterpart.

2.3.3 Ablation Study Results

Each model was rigorously evaluated using the same clinical images, hyperparameters, and evaluation metrics as outlined in Table 2.2. The purpose of these experiments is to unravel the complex tapestry of our model's architecture, examining how each thread contributes to the overall pattern of performance.

Table 2.5 encapsulates the outcomes of these ablation studies, offering a comprehensive view of the repercussions of each modification. Notably, the exclusion of non-linearities led to the most pronounced decline in performance, particularly in the accurate calculation of Aortic Distensibility (AD). This finding underscores the paramount importance of non-linear processing in capturing the nuanced relationships within our data.

Furthermore, the augmentation in model size, as seen in the 'No dense pruning' and 'No filter pruning' scenarios, did not translate into enhanced accuracy. This intriguing result hints at a potential overfitting issue, suggesting that a leaner model architecture might be more efficacious for our specific application.

The absence of full labels and the focal Tversky loss significantly impacted the accuracy of AAo AD calculations. This phenomenon could be attributed to an increase in false positives, particularly from surrounding structures, highlighting the critical role these components play in refining our model's focus and precision.

In examining the results across the board, we observe a consistent and substantial drop in performance, both in terms of the absolute error in the area and the Dice coefficient for all variants. However, the absolute error in AD calculations exhibited more variability, likely due to its increased sensitivity to segmentation artefacts. This variability further emphasises the importance of robust and accurate segmentation throughout the cardiac cycle, a challenge our proposed model adeptly addresses.

In conclusion, these ablation studies not only affirm the judicious design choices made in our proposed model but also offer invaluable insights into the complexities of DL architectures for medical image analysis. Through this meticulous dissection, we gain a deeper understanding of each component's role, allowing us to refine our approach further and pave the way for more nuanced and effective models in the future.

Table 2.2: Quantitative performance of the proposed method using absolute errors in the aortic area, aortic distensibility (AD), and the Dice coefficient. Also provided are the p values obtained from the Wilcoxon-signed rank test with Bonferroni correction ($\alpha = 0.05$). The mean ground truth area (averaged over time and patients) was 678.826 mm^2 (SD: 146.329 mm^2) for the AAo and 370.610 mm^2 (SD: 85.109 mm^2) for the DAo. Boldface indicates best performance.

Model	Absolute error in area (mm^2)		Absolute error in AD (10^{-3} mmHg^{-1})		Dice coefficient	
	AAo	DAo	AAo	DAo	AAo	DAo
Proposed, mean (±SD)	**7.346 (±2.257)**	**4.749 (±2.567)**	**0.394 (±0.401)**	**0.544 (±0.908)**	**0.989 (±0.003)**	**0.991 (±0.004)**
SOTA, mean (±SD)	32.323 (±25.420)	11.809 (±10.204)	1.088 (±1.395)	0.942 (±2.013)	0.965 (±0.0161)	0.978 (±0.015)
Unpruned, mean (±SD)	30.739 (±21.110)	12.389 (±7.209)	2.490 (±2.218)	1.404 (±1.759)	0.980 (±0.009)	0.970 (±0.012)
p values (Proposed vs SOTA)	1.710×10^{-15}	1.539×10^{-14}	5.291×10^{-08}	9.022×10^{-03}	1.710×10^{-15}	1.710×10^{-15}
p values (Proposed vs Unpruned)	1.711×10^{-15}	3.503×10^{-15}	7.427×10^{-13}	4.888×10^{-07}	1.651×10^{-14}	1.779×10^{-15}

AAo: Ascending Aorta, DAo: Descending Aorta, SD: Standard Deviation, SOTA: State-Of-The-Art.

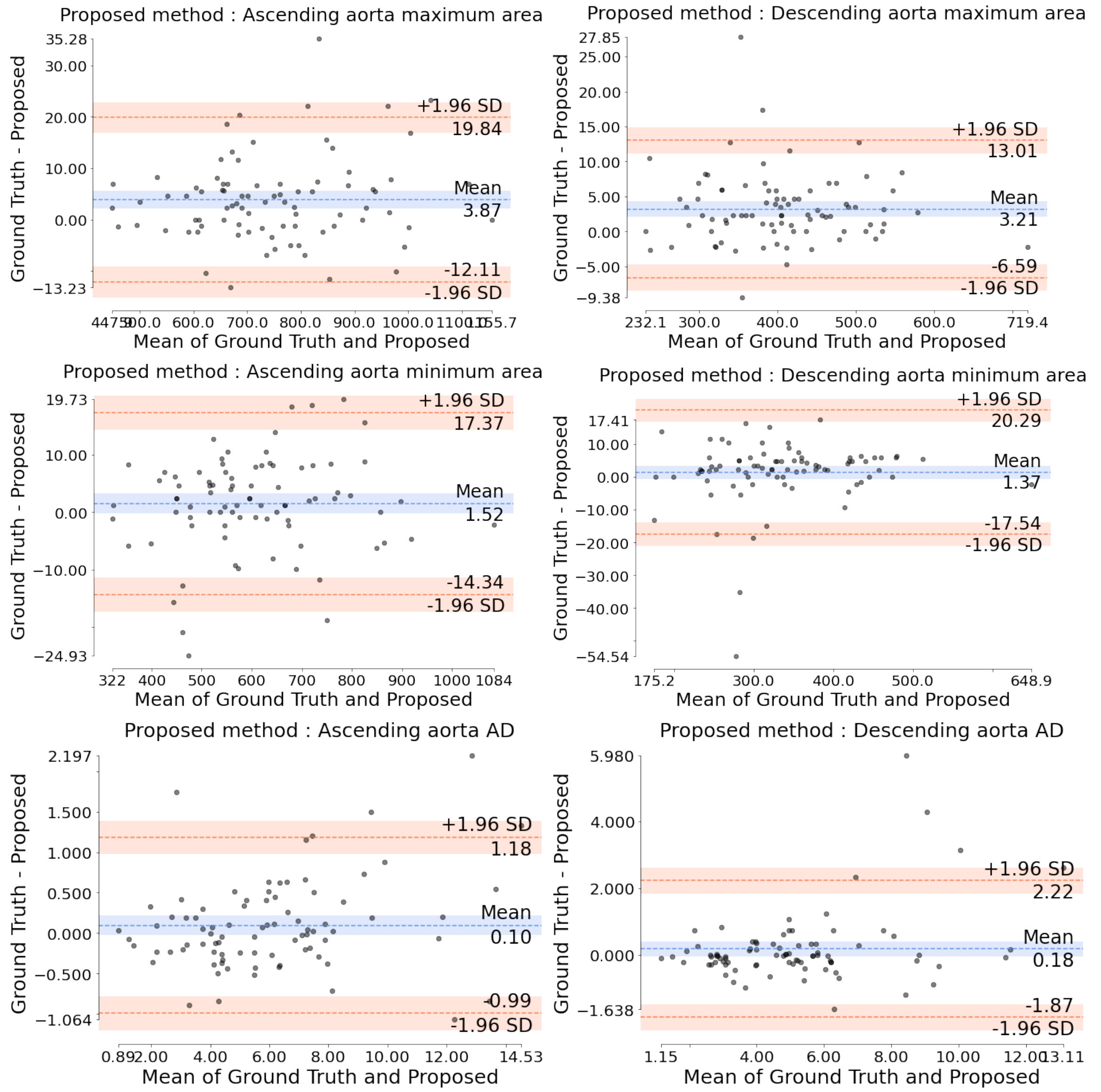


Figure 2.2: Bland-Altman analysis for graphically comparing the proposed method to the ground truth with respect to aorta maximum areas, aorta minimum areas and aortic distensibility (AD) values. Y-axis gives the difference between the two methods, whereas X-axis represents their mean. The area is measured in mm^2. AD is measured in 10^{-3} mmHg^{-1}. SD is the standard deviation.

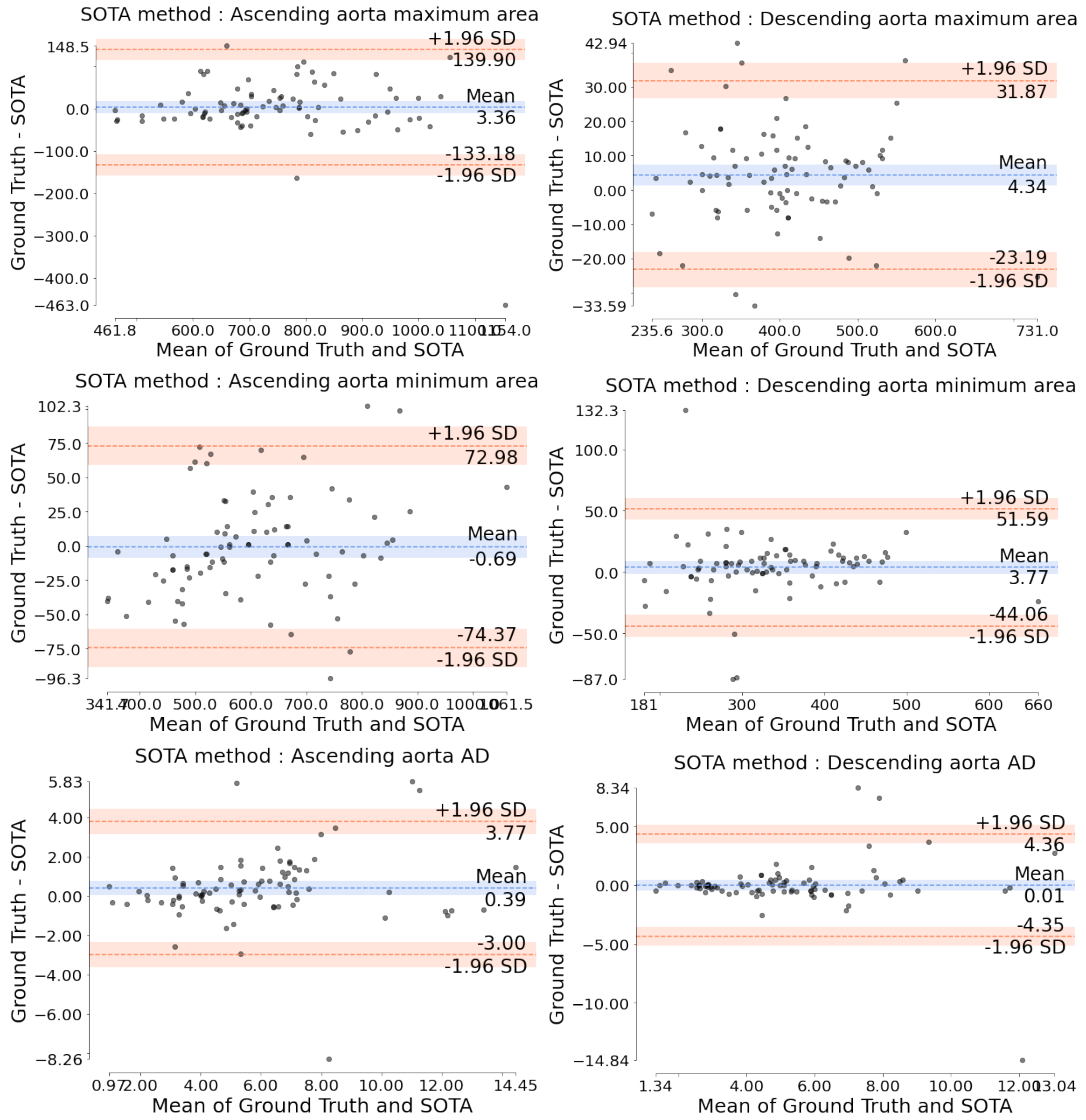


Figure 2.3: Bland-Altman analysis for graphically comparing the SOTA method to the ground truth with respect to aorta maximum areas, aorta minimum areas and aortic distensibility (AD) values. Y-axis gives the difference between the two methods, whereas X-axis represents their mean. The area is measured in mm^2. AD is measured in 10^{-3} mmHg^{-1}. SD is the standard deviation.

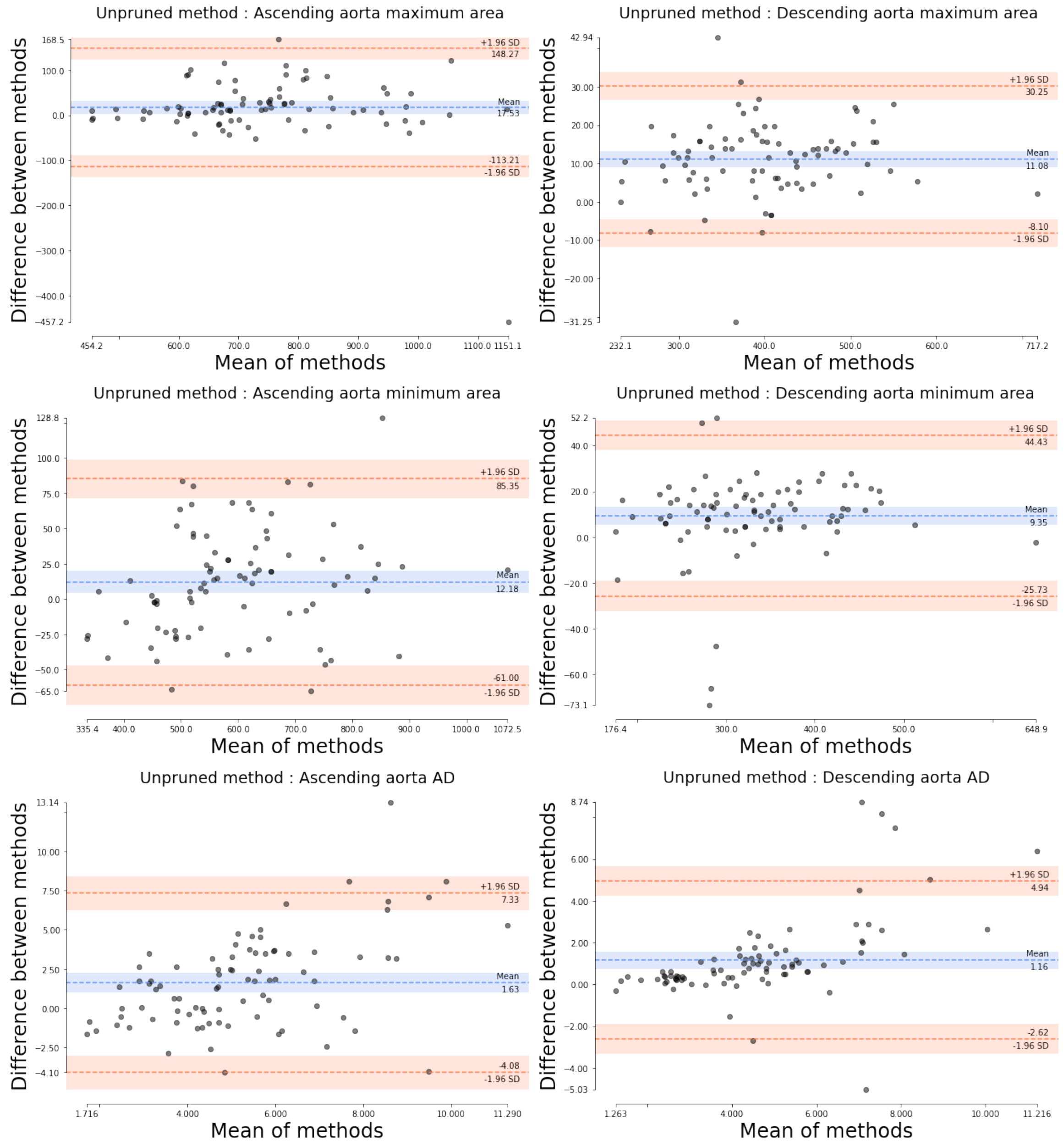


Figure 2.4: Bland-Altman analysis for graphically comparing the unpruned method to the ground truth with respect to aorta maximum areas, aorta minimum areas and aortic distensibility (AD) values. The Y-axis gives the difference between the two methods, whereas the X-axis represents their mean. The area is measured in mm^2. AD is measured in 10^{-3} mmHg^{-1}. SD is the standard deviation.

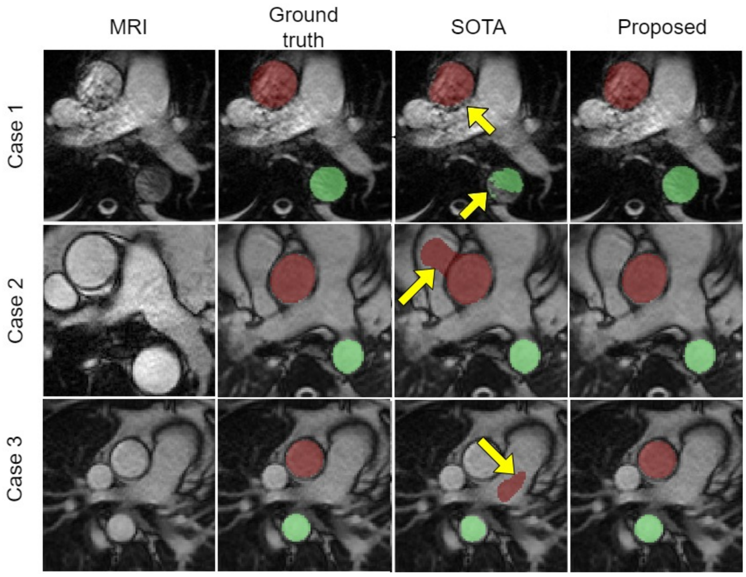
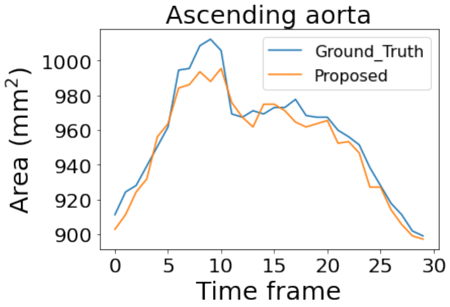
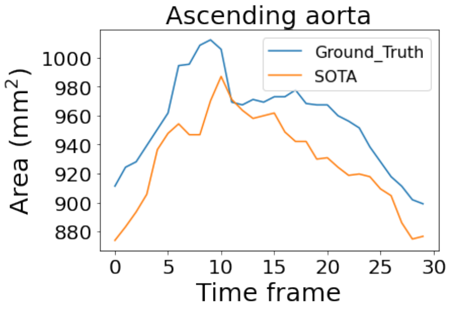
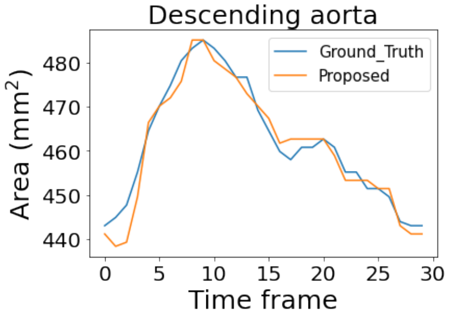
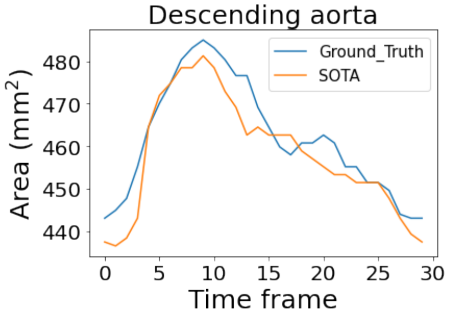


Figure 2.5: Qualitative comparison of the proposed method with the state-of-the-art (SOTA) method in one-time frame of the cardiac cycle for three representative cases. First column: MRI. Second column: Ground truth (semi-automated segmentation). Third column: Segmentation results of the SOTA method. Fourth column: Segmentation results of the proposed method. The yellow arrows indicate the errors in segmentation. The ascending aorta (AAo) is denoted by the red area and the green area denotes the descending aorta (DAo).

Table 2.3: Qualitative and quantitative comparisons of ascending and descending aorta time-area curves for one cardiac cycle of a representative case. Boldface indicates best performance.

Temporal curve	**Fréchet distance**	**Hausdorff distance**	**Dynamic time warping (DTW) distance**
	**16.822**	**16.822**	**165.424**
	37.384	37.384	337.390
	**5.607**	**4.673**	**63.552**
	9.346	6.542	85.048

SOTA: State-Of-The-Art, DTW: Dynamic Time Warping.

Table 2.4: Resource efficiency evaluation of the proposed method using the generated carbon emissions. The consumed energy and the equivalent distance a car could travel during the final training (250 epochs). Also shown are the training and average inference times. The experiments were conducted on a workstation with an NVIDIA RTX A6000 (48GB) GPU. Boldface indicates best performance

Model	CO_2eq (g)	Energy (kWh)	Equivalent distance travelled by car (km)	Training time (h:min:s)	Average inference time (ms)
Proposed	**2093.571**	**5.984**	**17.388**	**06:46:11**	**2.768**
SOTA	5785.498	23.184	48.052	35:09:20	7.544
Unpruned	11031.780	39.574	91.626	63:06:35	12.25

SOTA: State-Of-The-Art.

Table 2.5: Ablation over the number of features using the proposed framework.

Model	Absolute error in area (mm^2)		Absolute error in AD ($10^{-3}mmHg^{-1}$)		Dice coefficient	
	AAo	DAo	AAo	DAo	AAo	DAo
Proposed, mean (±SD)	**7.346 (±2.257)**	**4.749 (±2.567)**	**0.394 (±0.401)**	**0.544 (±0.908)**	**0.989 (±0.003)**	**0.991 (±0.004)**
No full labels, mean (±SD)	29.964 (±25.754)	8.649 (±7.776)	4.017 (±3.230)	3.038 (±2.581)	0.983 (±0.011)	0.970 (±0.016)
No focal Tversky, mean (±SD)	30.034 (±28.886)	10.399 (±6.815)	25.499 (±148.856)	2.161 (±1.807)	0.981 (±0.009)	0.969 (±0.012)
No non-linearities, mean (±SD)	30.589 (±20.173)	10.367 (±7.943)	30.761 (±210.163)	49.617 (±437.580)	0.980 (±0.016)	0.967 (±0.0154)
No dense pruning, mean (±SD)	30.034 (±28.886)	10.399 (±6.815)	25.483 (±148.859)	2.134 (±1.86)	0.981 (±0.008)	0.969 (±0.012)
No filter pruning, mean (±SD)	28.966 (±28.717)	9.322 (±6.152)	10.901 (±58.238)	1.991 (±1.701)	0.982 (±0.009)	0.971 (±0.012)
No BConvLSTM pruning, mean (±SD)	46.811 (±64.332)	17.288 (±8.793)	46.502 (±274.518)	2.116 (±2.227)	0.971 (±0.011)	0.962 (±0.014)

"No full labels" is trained using labels obtained by propagating ES and ED labels using non-rigid image registration." No focal Tversky" is trained using Dice coefficient loss, ignoring the severe class imbalance in the data. "No non-linearities" is trained by merging the encoder and decoder feature maps through linear concatenation. "No dense pruning" is trained using three densely packed convolutional blocks in the final encoding step. "No filter pruning" is trained using four times more filters in the convolutional layers than the proposed model. "No BConvLSTM pruning" is trained using BConvLSTM unit in three steps instead of one. Boldface indicates best performance. AAo : Ascending Aorta, DAo: Descending Aorta, SD: Standard Deviation.

2.4 Discussion

2.4.1 Strengths of this Study

Compared to previous work, this study has several strengths. We did not treat each time frame of the cardiac cycle as a separate entity for semantic segmentation. Instead, we integrated information related to space and time into our task by combining encoder and decoder feature maps in the second up-sampling layer using a non-linear function, specifically a BConvLSTM. Using the hyperbolic tangent function to merge the output of the forward and backward paths aids in the network's ability to learn complex data structures [31]. We implemented the focal Tversky loss during training to address the issue that the ROI class is much smaller than the background class. To leverage the enhanced credibility of the ground truth targets in all cardiac cycle time frames, we conducted end-to-end hierarchical learning and testing of the aortic lumen area from cine CMR images rather than focusing solely on a limited range of time frames. Additionally, we used multi-centre, multi-vendor data from a diverse patient cohort, which greatly enhances the generalization ability of the proposed aortic lumen segmentation algorithm. Site, vendor, and patient heterogeneity are crucial when testing a model for effective clinical implementation and accreditation agency approval. Unlike the model [31] that inspired this paper, our algorithm uses a building block that returns the sequence of feature maps over all time steps since we are dealing with video inputs. Moreover, to achieve a more resource-efficient architecture, our algorithm: (i) employs four times fewer filters in the convolutional layers, (ii) uses BConvLSTM layers in only one-third of the steps, and (iii) contains only one-third of the densely packed convolutional blocks in the final encoding step.

2.4.2 Main Findings

The segmentation masks for AAo and DAo, as predicted by our model, aligned well with the semi-automated assessments of CMR experts, as evidenced by the low area discrepancies and elevated Dice coefficient metrics. This is in agreement with the findings of [65], asserting that a single densely populated convolution block at the terminal layer of the contraction sequence is adequate for feature diversity. Our approach was bench-marked against other DL methodologies [27, 31] synonymous with the current SOTA and their unpruned counterparts. Quantitative evaluations spotlighted our model's superior performance against both SOTA and unpruned models, achieving minimal aortic and AD errors while boasting elevated Dice coefficient scores for both AAo and DAo. The BA analysis, when contrasted with the gold standard, indicated our method's predicted values for maximum and minimum aorta areas, along with AD metrics, oscillated approximately ∼3.9 times less compared to the SOTA and unpruned models. This resonates with the conclusions of recent literature [66], which underlines the efficacy of intertwining temporal dynamics with spatial processing in video datasets. A distinguishing attribute of our technique is its capability to pinpoint aortic sections in an atypical instance, where the DAo's positioning deviates from the norm in relation to the AAo. Such anomalies are typically evident in individuals diagnosed with scoliosis or those exhibiting abnormalities in aortic arch branching and orientation. Contrary to prior works [27, 28], models assessed in our study displayed slightly worse performance on the AAo as opposed to the DAo, possibly attributed to the intricate nature of structures flanking the AAo within our varied dataset. We verified the model's uniformity throughout the cardiac cycle by showcasing time versus area plots for AAo and DAo, alongside estimating deviations from the gold standard for a prototypical instance. Regarding resource usage and environmental impact, our model was assessed against SOTA and the unpruned versions employing the Carbontracker toolkit [63]. It emerged that our model required ∼3.9 times less energy, was approximately ∼2.8 times more environmentally friendly during the training phase than the SOTA. Moreover, the model demanded ∼5.2 times less training duration, with inference speeds surpassing by ∼2.7 times than SOTA. As expected, when juxtaposed against the unpruned models, our technique showcased ∼5.2 times reduced environmental implications, consumed ∼6.6 times less energy, demanded ∼9.3 times shorter training intervals, and the inference was accelerated by nearly ∼4.4 times. We found through ablation studies that (i) The exclusion of non-linearities resulted in the largest prediction errors; (ii) Increasing the model size did not lead to improved accuracy. This interesting outcome suggests that the model may have been overfitting and indicates that a simpler model architecture might be more effective for our particular use case or some cases; (iii) The precision of AAo AD calculations was notably affected by the lack of complete labels and the focal Tversky loss. This effect was mainly due to increased false positives, especially from neighbouring artefacts. Thus, it underscores the importance of these components in enhancing the accuracy and focus of our model.

2.4.3 Clinical Implications

Prior research carried out in our laboratory has demonstrated [37] that the AD aortic stiffness metric exhibits greater reproducibility compared to pulse wave velocity. The exceptional findings detailed in this paper significantly decrease the time required to extract aortic structural and functional phenotypes from CMR data while enhancing the outcomes' dependability. As a result, our pipeline could serve as an invaluable tool for investigating genome-wide associations between AD and aortic areas in relation to cognitive [67, 68] performance within large-scale biomedical databases (such as the UK Biobank) that more accurately represent the broader population. Acquiring quantitative CMR phenotypes on this scale remains a considerable challenge today. This type of analysis would allow us to examine potential causal connections between aortic measurements and aortic aneurysms, brain small vessel disease, and the reciprocal relationship with blood pressure indices. Identifying the responsible mechanisms could ultimately lead to (i) a deeper comprehension of factors contributing to cognitive decline and dementia and (ii) the discovery of novel therapeutic targets. Additionally, the model designed for the AD quantification task can be repurposed as a foundation for various clinically pertinent tasks (through transfer learning), enabling swift advancements and superior performance. The proposed framework can be easily incorporated into CMR analysis software. We have made the code for our image analysis pipeline available online [1]

2.4.4 Resource Efficiency Considerations

Research has shown [69] that between 2012 and 2018, the computational demands for DL research increased 300,000-fold, far outpacing the historical growth of computational requirements. These computations necessitate enormous [30] amounts of energy, contributing to greenhouse gas emissions and an outsized carbon footprint [30], potentially exacerbating global climate change. Although natural language processing applications with large-scale models are primarily responsible for this issue, energy consumption by typical medical image analysis DL models is still significant. Our model's energy consumption during training was estimated at 5.984 KWh, compared to 23.183 KWh for the SOTA model. The resulting carbon emissions were equivalent to a car travelling 17.388 km for our model and 48.052 km for SOTA. However, these figures only pertain to the final training session. A standard development process for determining the optimal model typically involves multiple training runs for hyperparameter tuning and experimenting with various model architectures. To better understand the energy usage and carbon footprint of a complete DL model development pipeline, a recent study [70] discovered that

[1] `https://github.com/tuanaqeelbohoran/Aortic-Distensibility.git`

constructing and testing a final paper-worthy DL model required training 4789 models over six months. The implications of these alarming figures are immense, particularly when considering the global adoption of healthcare DL applications. Although concerns about the energy usage and carbon footprint of DL research have begun to surface [30, 63, 70], the majority of healthcare computer vision DL research focuses on improving accuracy while neglecting resource efficiency. In this paper, we explicitly considered energy usage and CO_2eq emissions. We support previous recommendations [30, 70, 71, 72, 73] to evaluate DL research using key metrics like energy usage and carbon emissions alongside accuracy-related measures, fostering innovation in DL algorithmic efficiency. The merit of DL models should be determined by their intelligence per joule. This approach could help: (i) mitigate the negative environmental impact of DL research during training and development and (ii) promote inclusivity in DL research [73] by enabling broader participation. As we face an energy-intensive future and escalating rates of natural disasters, it is crucial [30] to explore methods to control DL's energy consumption and carbon cost. Additionally, reemphasising resource efficiency may increase the portability and applicability of our method, facilitating broader adoption on devices with lower computational power and memory and ultimately enabling more widespread and systematic DL-driven evaluations of CMR-derived aortic stiffness.

2.4.5 Study Limitations

Several limitations exist within this study. Although the training dataset is extensive and diverse, our network may yield less accurate outcomes when presented with pathologies, age groups, ethnicities, and scanners not represented in the training set. This challenge is the primary obstacle preventing the implementation of any DL model in real-world settings. To address this issue partially, we utilised data augmentation methods that emulate various potential data distributions. As a result, our pipeline demonstrated an enhanced capacity to generalise to unobserved non-representative cases compared to SOTA. Another significant limitation of our model is DL algorithms' inherent "black-box" nature, resulting in a lack of interpretability. Explainable tools are essential [74] for fostering trust in DL models, and creating such a module will be a focus of future research. Additionally, our model has not been assessed against adversarial attacks [75]. Although Carbontracker supports numerous environments and platforms, minor discrepancies may exist between the reported energy consumption, carbon emission values, and figures. Such deviations could arise from the quality of the estimated carbon intensity of electricity production, which can vary by geographic location (depending on local energy sources) and time of day (as energy demand and capacity fluctuate). However, all DL experiments in this study were conducted on the same workstation and at the same time of day. Moreover, Carbontracker employs "real-time" carbon intensity values, which are updated every 15 minutes during training using the tool's supported application programming interfaces. In this study, the predicted maximum and minimum aortic areas utilised for AD quantification did not consistently correspond with the diastolic and systolic cardiac cycle phases observed in the handcrafted analysis. Nonetheless, this discrepancy did not significantly impact the AD measurements. Our study required full annotation of all temporal frames in the cine dataset instead of SOTA, which required only sparse annotation. However, the SOTA method also needs extra resources (people, time) to identify the ED and ES phases. Significant resources (hardware, time) were also needed to perform the computationally intensive non-rigid registration. Lastly, the annotation in our study was not so time-consuming because it was largely supported by the JIM software that performed automated label propagation, which was then verified by the clinicians. Therefore, the consensus on the standards is that our study is concerned only with training resources. Finally, other recent fully automated segmentation approaches exist [76, 77], potentially performing better than ours in the aortic lumen delineation task from cine CMR images. However, this work aimed to propose a DL-based framework that surpasses the SOTA methods for this particular task while being more resource-efficient rather than to carry out an exhaustive survey of semantic segmentation, the literature of which is huge.

Chapter 3

Enhanced Right Ventricular Volume Prediction and Uncertainty Estimation: From Supervised Tree Kernel Ensembles to Feature Tokeniser Transformer-based Regression on 2D Echocardiography Planimetry Data

3.1 Introduction

3.1.1 Clinical Background

Right ventricular (RV) volume assessment plays a pivotal role in evaluating RV size and function, which are crucial for diagnosing and managing a wide spectrum of CVDs. These diseases encompass conditions such as congenital heart defects, pulmonary hypertension, coronary artery disease, and heart failure [78]. Accurate and reliable assessment of RV volumes is essential for risk stratification, guiding therapeutic decision-making, and monitoring treatment response in these patient populations.

Cardiac magnetic resonance imaging (CMR) has emerged as the gold standard for quantifying RV volumes due to its exceptional accuracy and reproducibility [79, 80, 81]. CMR offers unparalleled visualisation of the RV anatomy and function, allowing for precise volume measurements through various techniques. However, the widespread adoption of CMR is hampered by several limitations. Firstly, access to CMR scanners is severely limited, particularly in regions with resource constraints. In the United Kingdom, for instance, there are only 6.1 magnetic resonance imaging scanners per million people [82]. This limited availability creates significant logistical challenges for patients requiring RV volume assessment.

Secondly, the high cost associated with CMR examinations can be a barrier for some healthcare systems and patients. The complex infrastructure and skilled personnel required for CMR operation further contribute to its cost. Thirdly, CMR scan times can be lengthy, impacting patient throughput and potentially increasing wait times for those in need of the procedure. Finally, certain patient populations, such as those with claustrophobia or metallic implants, may not be suitable candidates for CMR due to safety concerns.

Despite its limitations, CMR remains the gold standard for RV volume quantification. However, the aforementioned challenges necessitate the exploration of alternative imaging modalities that are more readily available, cost-effective, and time-efficient.

Two-dimensional echocardiography (2DE) emerges as the primary alternative imaging modality for RV evaluation due to its widespread availability, portability, and cost-effectiveness [83].

3.1.2 The Challenge and Related Work

Unlike CMR, which provides a comprehensive view of the heart from multiple angles, 2DE relies on ultrasound waves to generate cardiac images from specific windows. This inherent limitation (poor quality and inability to penetrate deeply into organs) in image acquisition can lead to challenges in accurately depicting the complex three-dimensional geometry of the RV.

The accuracy of 2DE-based RV volume estimations is further compromised by the absence of precise geometric models that can accurately capture the intricate morphology of the RV [84]. These models are mathematical representations of the RV shape used to calculate volumes from 2DE planimetry measurements. The limitations of current geometric models often lead to under or overestimating accurate RV volumes.

Three-dimensional echocardiography (3DE) has the potential to address this gap by enabling quantitative RV assessment without relying on geometric assumptions. 3DE technology allows for the reconstruction of 3D images of the heart from multiple 2DE views. This reconstruction provides a more comprehensive representation of the RV, facilitating more accurate volume calculations. However, 3DE is hampered by limitations in image quality compared to CMR. The reconstruction process can introduce artefacts that can affect the accuracy of volume measurements. Additionally, 3DE requires specialised equipment and extensive training for operators, limiting its widespread adoption. Furthermore, the cost and computational demands associated with 3DE are significantly higher compared to 2DE.

These limitations, coupled with the pressing need for accurate and accessible RV volume measurements, pose a critical challenge for healthcare systems worldwide. The backlog of cardiac scans, including those for RV volume assessment, is growing, and patients awaiting CMR imaging studies face extended delays. These delays can lead to increased mortality and complications, particularly for patients with critical conditions requiring timely diagnosis and intervention. Therefore, the challenge is twofold: to improve the reliability and accuracy of RV volume measurements using 2DE and mitigate the limitations in the availability and accessibility of CMR scanners.

Recent advancements in machine learning (ML) and deep learning (DL) offer exciting possibilities for overcoming the limitations of traditional 2DE-based RV volume assessment. Gradient boosted regression trees (GBRTs) are known for their effectiveness in handling tabular data, which is the type of data readily available from 2DE examinations [85]. Transformers [86] are a type of deep neural network architecture that has demonstrated remarkable success in various natural language processing tasks.

These techniques have the potential to learn complex relationships between readily obtainable 2DE data, such as area measurements from various standardised views and corresponding RV volumes obtained from CMR (considered the gold standard). By leveraging this knowledge, ML and DL models can potentially generate more accurate and reliable estimates of RV volumes from 2DE data. On top of this, the uncertainty quantification of each ML/DL model RV volume prediction would be crucial for clinicians, as it provides insights into the confidence level they can place in the model's estimates. However, no ML/DL studies exist to date that predict RV volumes and uncertainty levels based on 2DE planimetry data.

3.1.3 Our Contribution

The contributions of this study are:

- We propose to employ gradient boosted regression trees (GBRTs) in conjunction with a k-nearest neighbour method, supported by a supervised tree kernel, to estimate the RV volume and the uncertainty associated with the model's predictions [87, 88, 89]. The proposed framework relies on 2DE standardised planimetry data (area measurements) and age, cardiac phase, and gender information.
- We explore the potential of tabular Feature Tokeniser Transformers for enhanced RV volume prediction.
- Both methods are evaluated on a small dataset of 100 RV volumes.
- We propose to make use of explainability methods (the gain value of GBRTs) to understand what features contribute to the model the most, and only train on those features. We show that this contributes to reducing the burden of annotation and save resources by using fewer 2DE views.

3.2 Materials and Methods

3.2.1 Study Population and Image Dataset

The foundation of any robust medical study lies in the careful selection and comprehension of its study population. This study encompasses a retrospective cohort comprising 50 adult patients, all of whom underwent both 2DE and CMR within a 30-day window as part of their routine cardiac care [90]. This time-frame ensures the relevance and applicability of the data to real-world clinical scenarios. Patients with a history of significant cardiac interventions or conditions that could confound RV measurements were excluded to maintain the data's integrity. Ethical approval for this study was obtained from both the Columbia University Irving Medical Center Institutional Review Board and the Nottingham Trent University Ethics Committee, underscoring our commitment to ethical research standards.

3.2.2 Image Acquisition

Imaging protocols for 2DE and CMR have been previously described [91]. In summary, 2DE was performed according to the comprehensive protocol of the Columbia University Irving Medical Center (CUIMC) Echocardiography Laboratory and the American Society of Echocardiography recommendations [92] with a minimum of 22 anatomic views per study using commercially available equipment (Philips Epic 7C, CVx and i33 ultrasound systems).

CMR acquisition was obtained using commercially available equipment (Signa 1.5 Tesla MRI scanner, General Electric, Milwaukee, WI). Short-axis cine images were acquired using a SSFP pulse sequence with the following typical parameters: Repetition time (TR) 3.0 ms, echo time

(TE) 1.0 ms, flip angle of 60°, 16 views per segment, field of view 35×35 mm, acquisition matrix 256×256, slice thickness 8 mm with no gap, and receiver bandwidth 125 kHz.

3.2.3 Data Pre-processing and Annotation

The RV endocardial-myocardial interface was traced in ES and ED by a single cardiologist with expertise in echocardiography. This process of planimetry was performed for eight standardised RV views using commercially available software (Syngo Dynamics, Siemens): parasternal long axis (PLAX), RV inflow (RVInflow), parasternal short axis at the level of the aortic valve (PSAXAV), base (PSAXbase), mid (PSAXmid) and apex of the left ventricle (PSAXdistal), standard four chamber (FourC) and subcostal four chamber (SubC). A focussed RV view was used if the FourC had poor image quality. The software automatically calculated an area for each ED and ES tracing. The cardiologist was blinded to the CMR results [90]. CMR analysis was performed using commercially available software (cvi42 v5.11, Circle Cardiovascular Imaging, Calgary, Canada) by experienced cardiovascular radiologists who were blinded to the 2DE results. Cine loops were used to select images in ED and ES. Four chamber cine images were used as a reference to help define the atrioventricular valves and apical planes. Endocardial-myocardial segmentation was performed by manual tracing of each ED and ES short-axis view and used to calculate right ventricular end-diastolic volume (RVEDV), right ventricular end-systolic volume (RVESV) and right ventricular ejection fraction (RVEF). We applied gender-specific CMR cutoffs for RV dilatation and dysfunction as proposed by Petersen et al. [92] The eight area measurements described above, along with the patient's age, constituted the numerical input variables for our model, while gender and cardiac phase information served as categorical input variables. For each patient, the CMR-derived RVEDV and RVESV were recorded, resulting in a total of 100 data points.

3.2.4 Intra- and Inter-observer Variability

The reliability of measurements of RV tracings from eight different 2D echocardiography views in ED and ES were analysed for both intra- and interobserver agreement. Two additional sets of measurements were performed for a random subset of 10 patients to allow two clinicians to assess intra- and interobserver variability. We used the intraclass correlation coefficient (ICC) to evaluate the intra- and inter-observer reliability of 2DE tracings using a 2-way fixed and a 2-way random effects model, respectively [91]. An absolute agreement criterion was applied in both cases. ICC values of less than 0.5 indicated poor reliability, whereas values between $0.5-0.75, 0.75-0.9$ and above 0.9 indicated moderate, good and excellent reliability, respectively.

3.2.5 Baseline Characteristics

Table 3.1 summarises the study cohort's baseline clinical, echocardiographic and CMR characteristics. Patients had median age 51, interquartile range $32 - 62$ and 42% were women. Nineteen patients (38%) had a clinical history of heart failure and 6 (12%) had prior left sided cardiac surgery. Seven patients (14%) had hypertrophic cardiomyopathy, and 6 (12%) had simple congenital heart disease (atrial/ventricular septal defect, anomalous coronary artery, bicuspid/unicuspid aortic valve). There were no significant differences in clinical characteristics between the training ($n = 40$) and testing ($n = 10$) subsets (Table 3.1).

By 2DE, 28% of patients had left ventricular systolic dysfunction. RV dilatation and dysfunction were identified in 20% and 16% of patients, respectively, using 2DE. No patient had severe RV dilatation or dysfunction. The mean RV basal diameter was 3.8 ± 0.8 cm. Tricuspid annular plane systolic excursion (TAPSE) and maximum tissue Doppler velocity in systole (RVS) were 18.8 ± 5.6 mm and 11.8 ± 2.5 cm/s, respectively. The testing subset did not include any patient with RV dilatation by 2DE, however this difference was not statistically significant ($p = .317$).

The mean time interval between 2DE and CMR was 9.9 ± 5.6 days (median six days, interquartile range $2 - 20$ days). By CMR, mean RVEDV and RVESV were 163 ± 70 mL and 86 ± 45 mL, respectively. Mean RVEF was $50 \pm 8\%$. Six patients (12%) had RV dilatation, and 9 (18%) had RV dysfunction. According to the same cutoffs, there was no difference in RV dilatation or dysfunction between the training and testing subsets (13% vs. 10% and 18% vs. 20%, respectively).

Table 3.1: Baseline clinical and imaging characteristics (presented as mean±SD or frequency (%)) of the right ventricular study.

Clinical Data	**All (n = 50)**	**Training (n = 40)**	**Testing (n = 10)**	**p-value**
Female gender	21 (42)	16 (40)	5 (50)	.567 [90]
Age	47 ± 18	48 ± 18	46 ± 17	.698 [90]
Coronary artery disease	11 (22)	9 (23)	2 (20)	.864 [90]
Diabetes	7 (14)	6 (15)	1 (10)	.684 [90]
Paroxysmal atrial fibrillation	6 (12)	5 (13)	1 (10)	.828 [90]
Hypertrophic cardiomyopathy	7 (14)	5 (13)	2 (20)	.541 [90]
Simple congenital heart disease	6 (12)	5 (13)	1 (10)	.828 [90]
Heart failure	19 (38)	14 (38)	5 (50)	.449 [90]
Prior cardiac surgery	6 (12)	5 (13)	1 (10)	.828 [90]
Echocardiography				
LV systolic dysfunction				.750 [90]
Mild	3 (6)	2 (5)	1 (10)	[90]
Moderate	8 (16)	6 (15)	2 (20)	[90]
Severe	3 (6)	3 (8)	0 (0)	[90]
RV basal diameter (cm) (FourChamber)	3.8 ± .8	3.9 ± .7	3.8 ± .9	.698 [90]
RV dilatation				.317 [90]
Mild	6 (12)	6 (15)	0 (0)	[90]
Moderate	4 (8)	4 (10)	0 (0)	[90]
RV dysfunction				.843 [90]
Mild	7 (14)	5 (13)	2 (20)	[90]
Moderate	1 (2)	1 (3)	0 (0)	[90]
TAPSE (mm)	18.8 ± 5.6	18.6 ± 5.8	19.4 ± 5.3	.808 [90]
RVS (cm/s)	11.8 ± 2.5	11.8 ± 2.6	11.9 ± 1.8	.634 [90]
FAC (%)	46 ± 10	46 ± 10	44 ± 13	.357 [90]
TRE moderate or more	5 (10)	5 (13)	0 (0)	.239 [90]
CMR				
Time between TTE and CMR [days]	9.9 ± 5.6	10 ± 6	8 ± 8 8	.488 [90]
RVEDV (mL)	163 ± 70	164 ± 76	160 ± 39	.628 [90]
RVESV (mL)	86 ± 45	83 ± 45	79 ± 28	.913 [90]
RVEF (%)	50 ± 8	50 ± 8	52 ± 8	.913 [90]
RV dilatation*	6 (12)	5 (13)	1 (10)	.828 [90]
RV dysfunction*	9 (18)	7 (18)	2 (20)	.854 [90]

LV: Left Ventricular, RV: Right Ventricular, TAPSE: Tricuspid Annular Plane Systolic Excursion, RVS: Peak Systolic Right Ventricular Tissue Doppler, FAC: Fractional Area Change, TRE: Tricuspid Regurgitation, TTE: Transthoracic Echocardiography, CMR: Cardiac Magnetic Resonance Imaging, RVEDV: Right Ventricular End-Diastolic Volume, RVESV: Right Ventricular End-Systolic Volume, RVEF: Right Ventricular Ejection Fraction, * denotes that gender-specific cutoffs were applied.

3.2.6 Machine Learning Methods

As mentioned above, the challenge in this study is to improve the reliability and accuracy of 2D planimetry-based RV volume measurements. For that, we propose two ML regression methods.

The first method is ensemble modelling-based, and the second method is transformer-based. Henceforth, we will refer to them as **Method I** and **Method II**, respectively.

Method I: Ensemble Modelling-based Regression

Method I estimates not only RV volume values, but also the uncertainty level associated with this estimation. In the sophisticated domain of probabilistic ML, a particularly challenging task is estimating the conditional probability distribution $P(y|x)$ for a specific target variable y based on an input vector x. This estimation is not just about predicting an outcome but understanding the uncertainty and confidence associated with the prediction. To address this, we borrow the Instance-Based Uncertainty quantification for Gradient boosted regression trees (IBUG) method [87], which combines instance-based learning [93] and supervised tree kernels [94]. IBUG initiates point (RV volume) predictions provided by a GBRT.

Gradient Boosted Regression Trees

Suppose a dataset $D := (x_i, y_i)$, where each instance comprises an input vector $x_i = (x_i^j)_{j=1}^p \in X \subseteq \mathbb{R}^p$ and a corresponding output $y_i \in Y \subset \mathbb{R}$. In the realm of ML, specifically within the context of ensemble learning, gradient-boosting stands out as a powerful algorithm for both regression and classification tasks. Described extensively in [95], its fundamental principle revolves around constructing a predictive model $F : X \to Y$ by iteratively enhancing the model through stage-wise additive modelling while minimising a pre-defined empirical loss function L.

The mechanism of gradient-boosting commences with a base learner, symbolised as $F_0(x)$, and progresses through a series of iterative refinements. Each subsequent model, denoted as $F_m(x)$ for the m^{th} iteration, is an augmentation of its predecessor $F_{m-1}(x)$, adjusted by a scaled weak learner, $\gamma_m h_m(x)$. This scaling factor, γ_m, known as the learning rate, controls the contribution of each weak learner to the overall model.

In the specific case of GBRTs, the typical choice of loss function L is the mean squared error (MSE), a standard metric in regression problems. The initial model parameter $F_0(x)$ is set as the mean of the outputs of the training outcomes, representing a rudimentary yet effective starting point. As GBRTs evolve, regression trees, highly regarded for their interpretability and flexibility, are employed as weak learners. These trees are particularly adept at approximating the residuals, i.e., the discrepancies between the observed and predicted values, at each iteration.

For a deeper understanding, let us delve into the process of selecting a decision tree at each iteration m. The tree is chosen to approximate the residual, conceptualised as the negative gradient of the loss function relative to the current model's prediction,

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \tag{3.1}$$

where r_{im} is the pseudo-residual for the i-th observation at the m^{th} iteration. The mathematical formulation of this process involves solving

$$\min_{\gamma,h} \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \gamma h(x_i)) \tag{3.2}$$

where N is the size of the training dataset. Each decision tree in the ensemble contributes to partitioning the feature space into distinct regions, termed leaves. These leaves, $\{r_m^j\}_{j=1}^{M_{GBRT}}$, where M_{GBRT} is the total number of trees, are the end products of recursive binary splitting, a hallmark of decision trees. For every leaf in a tree, a parameter θ_m^j is computed, often via a Newton-Raphson method [96], considering the second derivative of the loss function. This computation also incorporates a regularisation term lending stability to the model.

The model's final output for a given input x_{te} is an aggregate of the leaf values from all trees through which x_{te} traverses. This aggregation, symbolised as $F(x_{te}) = \sum_m \mathbb{1}^T h_m(x_{te})$, represents the cumulative wisdom of the ensemble, effectively harnessing the collective strength of individual trees to produce a robust prediction.

Regularisation Techniques: To prevent overfitting, a GBRT includes several regularisation techniques:

- Tree Constraints: Limiting the depth and size of the trees.
- Shrinkage: Multiplying the contribution of each tree by a learning rate typically less than 1.
- Stochastic Gradient Boosting: Using a random subset of the data to train each tree.

GBRT: Explainability: A benefit of using GBRTs is that it is straightforward to retrieve feature importance scores that indicate how valuable each attribute was in constructing the model. In this study, feature importance scores were calculated for all models using the 'Gain' metric, which quantifies the relative contributions.

Instance-based Uncertainty Quantification

The GBRT predictions lack information about prediction uncertainty. IBUG extends these point predictions to a probabilistic landscape by modelling the conditional mean of the forecast. This task is achieved using the scalar output of GBRTs.

To further develop the model for a comprehensive output distribution, IBUG employs a supervised tree kernel [94]. This kernel adeptly identifies training instances significantly similar to a target instance (Figure 3.1). The affinity, or the degree of similarity, of a training example x_i to a target example x_{te} is quantified as:

$$A(x_i, x_{te}) = \sum_{m=1}^{M_{GBRT}} \mathbb{1}[R_m(x_i) = R_m(x_{te})] \tag{3.3}$$

where $R_m(x_i)$ denotes the leaf within the m^{th} tree to which x_i is allocated. This affinity measurement (calculated by **Algorithm 1**) leverages the tree structure within the ensemble, assessing how frequently the training and target instances align in their path through the decision-making process.

Unlike other SOTA methods, IBUG provides a flexible model of conditional output distribution. The simplest model employs a normal distribution, using the GBRT's output to approximate the mean $\mu_{F(x_{te})} = F(x_{te})$, while the variance $\sigma^2_{F(x_{te})}$ is derived from the affinity scores. A vital calibration step optimises the variance estimation:

$$\sigma^2_{F(x_{te})} \leftarrow \gamma_f \sigma^2_{F(x_{te})} + \delta_f. \tag{3.4}$$

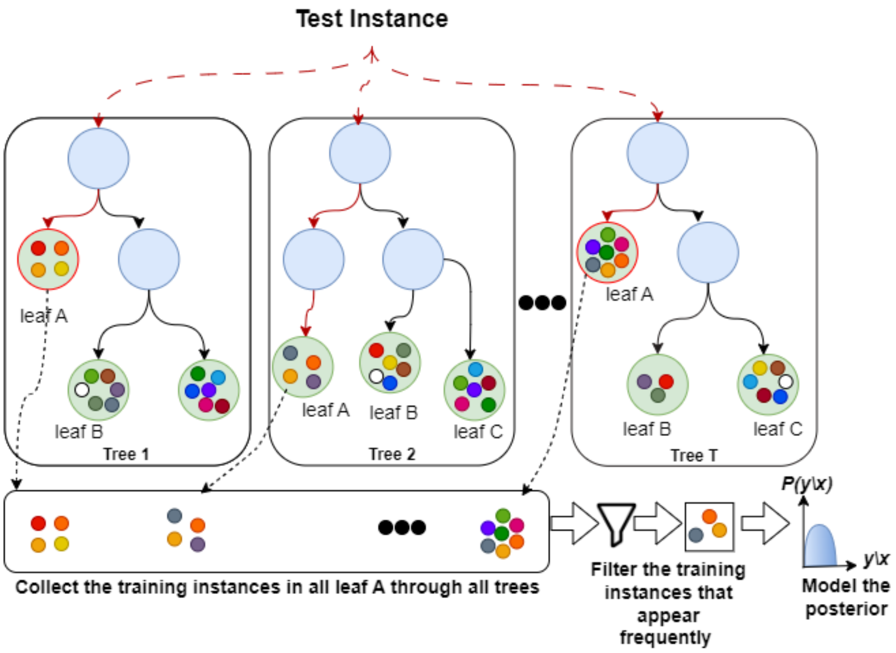

Test Instance
leaf A
leaf B
Tree 1
leaf A
leaf B
leaf C
Tree 2
leaf A
leaf B
leaf C
Tree T
Collect the training instances in all leaf A through all trees
Filter the training instances that appear frequently
P(y\x)
y\x
Model the posterior

Figure 3.1: IBUG flow chart. For a target instance, IBUG collects the training instances at each leaf it traverses, keeps the k most frequent samples, and then uses those instances to model the output distribution.

This step is crucial for aligning the variance with the inherent uncertainty of the data, achieved through tuning parameters γ_f and δ_f on a validation subset.

Beyond normal distributions, IBUG can adapt to various parametric or non-parametric distributions, enhancing its suitability to diverse datasets. This flexibility is demonstrated in the following equation (**Algorithm 2**):

$$\hat{D}_{te} = D_{dist}\left(A^{(k)} \mid \mu_{F(x_{te})}, \sigma^2_{F(x_{te})}\right). \tag{3.5}$$

The selection of k, the number of top affinity scores, is pivotal in shaping the accuracy of the probabilistic predictions. This parameter is tuned using the negative log-likelihood (NLL) on a validation subset $D_{\text{val}} \subset D$. We employ a procedure (**Algorithm 3**) of [87] to expedite this tuning process. This algorithm is designed to avoid the repetition of computationally intensive affinity calculations, significantly enhancing the efficiency of the tuning process. The algorithm also introduces parameter ρ_z to account for instances with abnormally low variance, ensuring the model's performance robustness.

Method II: Feature Tokeniser Transformer-based Regression

Before we delve into the proposed Transformer-based model [97], for **Method II**, we will explore a transformer's basic components and principles.

Introduction to Vanilla Transformers: Transformers, introduced in 2017, represent a paradigm shift in DL architectures, particularly in natural language processing (NLP) and beyond [86]. Unlike their predecessors, which relied on recurrent neural networks or CNNs, Transformers are based solely on attention mechanisms, enabling more efficient parallel processing and better handling of sequence data.

Algorithm 1 IBUG Affinity Computation

```
Require: Input instance $x \in X$, GBRT model $F$.
procedure COMPUTEAFFINITIES($x, F$)
$A \leftarrow \vec{0}$    ▷ Init. train affinities
for $t = 1...T$ do
    Get instance set $I_l^t$ for leaf $l = R_t(x)$    ▷ Visit each tree
    for $i \in I_l^t$ do
        $A_i \leftarrow A_i + 1$    ▷ Increment affinities
    end for
end forreturn $A$
```

Algorithm 2 IBUG Probabilistic Prediction

```
Require: Input $x \in X$, GBRT model $F$, $k$ highest-affinity neighbors $A^{(k)}$, min. variance $\rho_z$, variance calibration parameters $\gamma_f$ and $\delta_f$, target distribution $D_{dist}$.
procedure PROBPREDICT($x, F, A^{(k)}, \rho_z, \gamma_z, \delta_z, D_{dist}$)
$\mu_{F(x)} \leftarrow F(x)$    ▷ GBRT scalar output
$\sigma^2_{F(x)} \leftarrow \max(\sigma^2(A^{(k)}), \rho_z)$    ▷ Ensure $\sigma^2 > 0$
$\sigma^2_{F(x)} \leftarrow \gamma_f \sigma^2_{F(x)} + \delta_f$    ▷Var. calibration. return $D_{dist}(A^{(k)}|\mu_{F(x)}, \sigma^2_{F(x)})$
```

Algorithm 3 IBUG Accelerated Tuning of k

```
Require: Validation dataset $D_{\text{val}} \subseteq D$, GBRT model F, list of candidates $K_{cand}$, target distribution $D$, probabilistic scoring metric $V$, minimum variance $\rho = 1e - 15$.
procedure FastTuneK($D_{\text{val}}, F, K_{cand}, D, V, \rho_z$)
for $(x_j, y_j) \in D_{\text{val}}$ do
    $A \leftarrow$ ComputeAffinities($x_j, F$)    ▷ Algorithm 1
    $A \leftarrow$ Argsort $A$ in descending order
    for $k \in K_{cand}$ do
        $A^{(k)} \leftarrow$ Take first $k$ training instances $(A, k)$    ▷ Use same ordering for each $k$
        $\hat{D}^k_{y_j} \leftarrow$ ProbPredict($x_j, F, A^{(k)}, \rho_z, 1, 0, D_{dist}$)    ▷ Algorithm 2
        $S^k_j \leftarrow V(y_j, \hat{D}^k_{y_j})$    ▷Save validation score
    end for
end for
$k \leftarrow$ Select best $k$ from $S$
$p \leftarrow$ Select minimum $\sigma_z^2$ from $\hat{D}_k$ return $k, p$
```

The Core Components of a Vanilla Transformer

Architecture Overview: Transformer comprises two primary parts: The encoder and the decoder. Each part is constructed from layers that include multi-head attention mechanisms and feed-forward NNs [86].

Multi-Head Attention Mechanism: At the heart of a Transformer lies the multi-head attention mechanism. This mechanism allows the model to focus on different parts of the input sequence simultaneously, facilitating the understanding of complex dependencies [98]. Unlike traditional single-attention mechanisms, multi-head attention performs multiple parallel attention computations, enhancing the model's ability to capture various aspects of the data.

Feed-Forward Neural Networks: Following the attention layers, Transformers employ feed-forward NNs. These networks process the outputs of the attention layers independently for each position, adding another level of transformation to the data [98].

Proposed Neural Network Architecture

The proposed pipeline was inspired by [99]. It constitutes an adaptation of the Transformer architecture to tabular (both categorical and numerical) data. Figure 3.2 portrays the inner workings of the proposed pipeline. In brief, all categorical and numerical inputs are tokenised and then forwarded to cascaded Transformer layers.

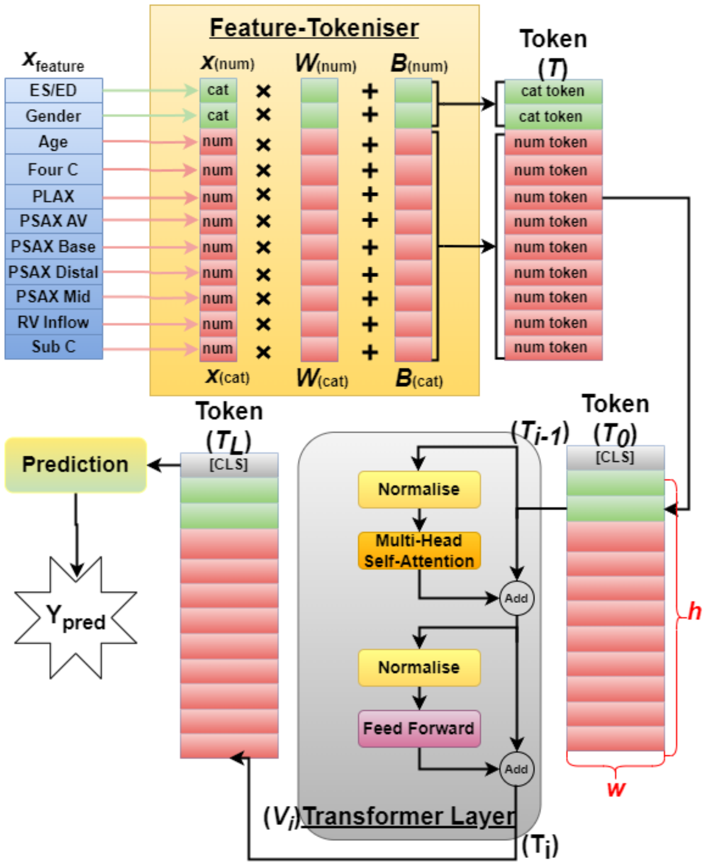


Figure 3.2: The proposed DL model. $X_{feature}$ is the input. Token (T) is the input embedding produced by the Feature Tokeniser. T_0 is the [CLS] token appended embedding matrix. V_i is the i^{th} Transformer layer. T_L is the output product of the cascaded Transformers.

The **Feature-Tokeniser** module transforms inputs $X_{feature}$ to embeddings denoted by Token $(T \in \mathbb{R}^{h \times w})$. The embedding for a feature Xf_i is obtained by:

$$T_i = B_i + f_i(Xf_i) \in \mathbb{R}^w \qquad f_i : \mathbb{X}_i \to \mathbb{R}^w \tag{3.6}$$

where B_i is the i^{th} feature bias, $f_{i(num)}$ is the element-wise multiplication of $Xf_{i(num)}$ (numerical features) with vector $\mathrm{W}_{i(num)} \in \mathbb{R}^w$, whereas $f_{i(cat)}$ is a lookup table with $\mathrm{W}_{i(cat)} \in \mathbb{R}^{S_i \times w}$ for $Xf_{i(cat)}$ (categorical features), where S_i is the number of categories for the i^{th} categorical feature. The whole process is described by:

$$T_{i(num)} = B_{i(num)} + Xf_{i(num)} \cdot W_{i(num)} \ \in \mathbb{R}^w, \tag{3.7}$$

$$T_{i(cat)} = B_{i(cat)} + e_i^T W_{i(cat)} \ \in \mathbb{R}^w, \tag{3.8}$$

$$\text{Token}(T) = \text{stack}[T_{1(cat)}, T_{2(cat)}, T_{1(num)}, T_{2(num)}, ..., T_{9(num)}] \ \in \mathbb{R}^{w \times h}, \tag{3.9}$$

where e_i^T is the one-hot vector for the corresponding categorical variable.

The embeddings T are transformed into a classification token in Transformer layers as described in [100]. Then the [CLS] token is appended to T and LT Transformer layers V_1, ..., V_{LT} are applied giving:

$$T_0 = \text{stack}[[CLS], T] \quad T_i = V_i(T_{i-1}). \tag{3.10}$$

Following [99], we have used the PreNorm variant for easier optimisation [101]. In addition, we discarded the first normalisation from the first Transformer to achieve a good performance, as stated previously. The final representation of the [CLS] token (also used for prediction) is:

$$Y_{pred} = \text{Linear}(\text{ReLU}(\text{LayerNorm}(T_L^{[CLS]}))). \tag{3.11}$$

3.2.7 AdaMax Optimisation

Adamax is a variant of the Adam algorithm, also introduced in [102], which uses the infinity norm (max norm) instead of the second moment. This makes Adamax more robust to large gradients and effective in certain situations where Adam might struggle.

Adamax updates the parameters using the following equations [102]:

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot (\nabla_\theta J(\theta_{t-1})) \tag{3.12}$$

$$v_t = \max(\beta_2 \cdot v_{t-1}, |\nabla_\theta J(\theta_{t-1})|^\infty) \tag{3.13}$$

$$\theta_t = \theta_{t-1} - [\frac{\alpha}{(1 - \beta_1^t)}] \cdot \frac{m_t}{v_t} \tag{3.14}$$

where:

- t is the time step;
- θ_* represents the model parameters at time $*$;
- ∇_θ is the gradient at time step t;
- m_t is the first moment estimate at time step t;
- v_t is the bias with infinity norm at time step t;
- β_1 is the exponential decay rate for the first moment;
- β_2 is the exponential decay rate for the infinity norm moment;
- $\nabla_\theta J(\theta_{t-1})$ is the gradient of the objective function with respect to the parameters, assigned the parameter values at the previous time step;

- $[\frac{\alpha}{(1-\beta_t^2)}]$ is the learning rate with the bias-correction term for the first moment.

AdaMax provides a robust alternative to Adam, particularly in scenarios with large gradients. It is effective in various ML applications where stability and convergence speed are crucial. AdaMax is a generalisation of Adam from the l_2 norm to the l_∞ norm.

3.2.8 Implementation and Training

Training Environment

The development and execution of our models were conducted leveraging a synergy of Cython and Python, with PyTorch providing the robust framework necessary for our DL undertakings. The computational experiments were performed on a computing system equipped with an *Intel Core™ i9-10900K Comet Lake* processor. This advanced processor features 10 cores and 20 threads, capable of reaching a peak clock speed of 5.3 GHz, which is instrumental in expediting the execution of computationally intensive tasks. Complementing this processing power, the system boasts 128 GB of DDR4 RAM, operating at a frequency of 2.6 GHz, to ensure smooth data handling and efficient model training processes. For graphical computations and parallel processing tasks that are crucial in ML operations, an *NVIDIA RTX A6000* graphics card with 48 GB of memory was utilised, underscoring the high-performance capability of our experimental setup.

Dataset Split

We employed 5-fold cross-validation to generate five different 80/20 train/test folds. For each fold, the 80% training set was randomly divided into a 60/20 train/validation set for hyperparameter tuning. After tuning the hyperparameters, the model was retrained using the complete 80% training set.

Method I: Implementation

Method I: Tuning Hyperparameters IBUG was applied to Extreme Gradient Boosting (XGBoost) [103], Light Gradient-Boosting Machine (LightGBM) [104], and Categorical Boosting (CatBoost) [105]. We tuned the hyperparameters using ranges as follows:

- k using values: [3, 5, 7, 9, 11, 15, 31, 61].
- γ_f and δ_f using values ranging from 1×10^{-8} to 1×10^{3} with additional multipliers $[1.0, 2.5, 5.0]$.
- Number of trees, M_{GBRT}, using values $[10, 25, 50, 100, 250, 500, 1000, 2000]$ (early stopping [106] was used for NGBoost).
- Learning rate using values $[0.01, 0.1]$.
- Maximum number of leaves using values $[15, 31, 61, 91]$.
- Minimum number of leaves using values $[1, 20]$.

- Maximum depth using values $[2, 3, 5, 7, -1]$ (indicating no limit).
- Adjusted the ρ_z parameter based on the minimum variance obtained from the validation set predictions.

Method I: Posterior Modelling To test IBUG's flexibility in posterior modelling, we modelled each probabilistic prediction using various distributions: normal, skew-normal, log normal, Laplace, student t, logistic, Gumbel, Weibull, and kernel density estimation (KDE).

Method II: Implementation

Method II: Loss Function and Training Strategy We employed the mean squared error (MSE) loss function to optimise our model. MSE is a common choice for regression tasks and is well-suited for our volume prediction problem. We aimed to minimise the MSE between our predicted volumes and the ground truth.

The training process spanned 500 epochs, with a batch size of 1. This approach allowed our model to learn the features and relationships in the data iteratively. Throughout the training, we continually monitored the validation performance to prevent overfitting.

Method II: Hyperparameters The efficacy of any deep learning model hinges on the selection of appropriate hyperparameters. We meticulously fine-tuned our hyperparameters to achieve optimal results. We tuned the hyperparameters between the boundaries as follows:

- Layers : [1 - 12].
- Feature embedding size : [4 - 1024].
- Residual dropout : [0.1 - 0.9].
- Attention dropout : [0.1 - 0.9].
- Feed forward network dropout : [0.1 - 0.9].
- Learning rate : [0.1 - 1e-6].
- Weight decay : [0 - 0.8].
- Optimiser : [SGD, Adam, Adamax, AdamW].

These hyperparameters were chosen with precision to strike a balance between model complexity and training stability.

We aimed to create a robust and effective model for ED and ES volume prediction by meticulously configuring our training process and optimising hyperparameters.

3.2.9 Model Evaluation and Statistical Analysis

Point Performance Evaluation Metrics

Root Mean Squared Error (RMSE) Root Mean Squared Error (RMSE) is a frequently employed metric in the field of statistics and ML for quantifying the difference between values predicted by a model and the values actually observed. The efficacy of RMSE stems from its capacity to measure the magnitude of error in prediction models on a scale that is commensurate with the original data [107].

Given a series of n predictions, $\{\hat{y}_i\}_{i=1}^n$, and the corresponding observed values, $\{y_i\}_{i=1}^n$, the RMSE is calculated using the formula [108]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}. \tag{3.15}$$

Squaring the errors before averaging them penalises larger errors, which means that RMSE is particularly sensitive to outliers in the data set. The subsequent square root transformation ensures that the units of RMSE are consistent with the units of the original measurements, thereby facilitating an intuitive interpretation of the model's performance.

RMSE serves as a standard gauge for the accuracy of predictive models. It is a crucial component in the repertoire of tools for model evaluation, offering a clear indication of model performance by assigning higher penalties to larger errors. This makes RMSE a robust measure against variance in error size, promoting models that offer consistency in their predictive accuracy [109].

As RMSE is a measure of prediction error, a model that predicts the observed values perfectly would yield an RMSE of 0. Conversely, the greater the discrepancy between the predicted and observed values, the higher the RMSE. Due to its ability to summarise the residuals' dispersion in a single measure, the RMSE is considered an essential metric in the domain of regression analysis, pattern recognition, and forecast evaluation.

Mean Absolute Error (MAE) MAE is a robust measure used in statistics and ML to quantify the accuracy of continuous variables. MAE provides an intuitive measure of average error magnitude unlike other complex metrics. It is defined as the average of the absolute differences between forecasted values and observed values [109, 110]. For a set of observations y_i and predictions $\hat{y}_i$, where $i = 1, 2, ..., n$, the MAE is computed as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|, \tag{3.16}$$

where $|\hat{y}_i - y_i|$ denotes the absolute error between the forecasted and observed value for the i-th pair. The appeal of MAE lies in its clear interpretability; the errors are scaled directly to the measured quantity, enabling straightforward comparisons across different contexts and datasets.

One of the primary advantages of MAE is its resistance to the influence of outliers, making it a reliable measure of central tendency in error distributions. It ensures that all individual differences are weighted equally in the average, providing a measure of predictive accuracy less sensitive to large deviations than RMSE [110].

MAE is particularly valuable in applications where the costs of positive and negative errors are equivalent. It has been utilised in a myriad of fields, including meteorology, finance, and health sciences, where accurate predictions are crucial for decision-making. Despite its simplicity, MAE remains a fundamental yardstick for model assessment, enabling practitioners to benchmark the performance of predictive algorithms effectively.

Mean Absolute Percentage Error (MAPE) MAPE is a statistical metric commonly utilised to gauge the performance of forecasting models. MAPE measures the size of the error in percentage terms, offering a perspective on the accuracy of predictions relative to the actual observed values. It is particularly favoured for its interpretability; the metric encapsulates the average absolute percentage difference between each observed value y_i and its corresponding forecasted value $\hat{y}_i$, across all n data points [107, 111]. The formula for MAPE is given by:

$$MAPE = \left(\frac{100\%}{n}\right) \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|, \tag{3.17}$$

where the absolute value of the percentage error for each data point is aggregated and then averaged. This percentage-based error measure is particularly useful when comparing the accuracy of forecasting models across different data scales or when the focus is on relative rather than absolute errors.

MAPE's interpretability comes with a caveat: it can be disproportionately affected by low values of y_i, leading to undefined or infinite percentages if y_i equals zero. Additionally, the asymmetry in the penalty of overestimates versus underestimates can sometimes yield misleading insights in scenarios where these differences are materially significant [112].

Despite these limitations, MAPE remains a staple in the pantheon of error metrics for forecasting due to its ease of understanding and implementation. It provides a quick, heuristic measure of model accuracy in percentage terms, facilitating communication with stakeholders who might be less familiar with more complex statistical measures.

R-squared (R^2) The R-squared (R^2) measure, often referred to as the coefficient of determination, is a statistical metric representing the proportion of variance in the dependent variable that is predictable from the independent variables in a regression model. It gauges the strength and effectiveness of the relationship between the model's predictions and the actual data. R^2 is expressed as a value between 0 and 1, where 1 indicates that the regression predictions perfectly fit the data, and 0 suggests no linear correlation between the predicted values and actual observations [113].

The calculation of R^2 is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \tag{3.18}$$

where y_i denotes the observed values, $\hat{y}_i$ represents the predicted values by the regression model, and $\overline{y}$ is the mean of the observed data. The numerator captures the sum of squared residuals, and the denominator quantifies the total variance in the observed data, implying that a higher R^2 value signals a model with greater explanatory power.

It is crucial to interpret R^2 with caution. A high R^2 does not inherently imply a causative relationship, nor does it guarantee that the model will perform well on unseen data. Furthermore, R^2 alone cannot determine whether the coefficient estimates and predictions are biased, which is why it should be used in conjunction with other statistics like the adjusted R^2, RMSE, and analysis of residual plots to validate a regression model's accuracy and reliability [114].

In practice, R^2 is often used to compare the fit of different regression models, provided that the models are nested or estimate the same outcome variable. Despite some criticism, it remains a widespread and informative metric in regression analysis, encapsulating the percentage of the response variable variation that a linear model explains.

Pearson's Correlation Coefficient Correlation is a statistical measure describing the extent to which two or more variables fluctuate. A correlation coefficient quantifies the degree to which a change in one variable is associated with a change in another. Pearson's correlation coefficient (denoted as r) is the most widely used among several types of correlation coefficients. It measures the linear relationship between two continuous variables. It ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 signifies no linear relationship [115].

The formula for Pearson's correlation coefficient is given by:

$$r = \frac{\sum_{i=1}^{n}(x1_i - \overline{x1})(x2_i - \overline{x2})}{\sqrt{\sum_{i=1}^{n}(x1_i - \overline{x1})^2}\sqrt{\sum_{i=1}^{n}(x2_i - \overline{x2})^2}}, \tag{3.19}$$

where $x1_i$ and $x2_i$ are the individual sample points indexed with i, $\overline{x1}$ and $\overline{x2}$ are the mean values of the respective sample sets. This coefficient is sensitive only to a linear relationship between two variables, which means it may underestimate the strength of non-linear relationships.

Spearman's rank correlation coefficient and Kendall's tau coefficient are non-parametric measures of statistical dependence. They are more suitable when the relationship between variables is not linear or when the data do not meet the normality assumptions required by Pearson's correlation [116, 117].

It is crucial to interpret correlation coefficients within the context of the research study, considering both the coefficient's size and significance. Correlation does not imply causation; a high correlation between two variables does not mean that one variable causes the change in another. Other statistical techniques are required to establish causal relationships.

Bland-Altman Analysis Bland-Altman (BA) analysis [56], also known as a Tukey mean-difference plot, is a statistical method used to visually assess the agreement between two measurements of the same quantity. The method is beneficial when the two measurements are obtained using different methods or instruments, and their agreement needs to be evaluated.

The BA plot shows the difference between the two measurements on the y-axis and the average of the two measurements on the x-axis. The plot also includes three horizontal lines: one representing the mean difference between the two measurements and the other representing the limits of agreement, defined as the mean difference plus or minus two standard deviations of the differences.

The BA analysis allows for the identification of systematic bias between the two methods and the presence of any outliers or trends in the differences. It also provides a measure of the overall variability of the differences.

The mathematics behind the BA analysis is straight-forward. Let the two measurements be denoted as x_{BA} and y_{BA}, and let their mean difference be denoted as $\bar{d}_{BA}$. $\bar{d}_{BA}$ can be calculated as follows:

$$\bar{d}_{BA} = \frac{1}{n_{BA}} \sum_{i=1}^{n_{BA}} d_{BAi}, \tag{3.20}$$

where n_{BA} is the number of observations.

The limits of the agreement are given by:

$$\bar{d}_{BA} \pm 1.96 s_d, \tag{3.21}$$

where s_d is the standard deviation of the differences, which can be calculated as:

$$s_d = \sqrt{\frac{1}{n_{BA} - 1} \sum_{i=1}^{n_{BA}} (d_{BAi} - \bar{d}_{BA})^2}. \tag{3.22}$$

Probabilistic Performance Evaluation Metrics

Continuous Ranked Probability Score (CRPS) CRPS is a measure [118] used to evaluate the accuracy of probabilistic forecasts. It is particularly effective when the forecasts are expressed as cumulative distribution functions (CDFs). The CRPS generalises the concept of the MAE to probabilistic forecasts, effectively quantifying the difference between the forecasted CDF and the empirical CDF of the observations. Mathematically, the CRPS is defined as:

$$CRPS(D_{CRPS}, x) = \int_{-\infty}^{\infty} (F_D(y) - \mathbb{1}(y \geq x))^2 \, dy, \tag{3.23}$$

where F_D is the CDF of the forecasted distribution D_{CRPS}, x is the observed value, and $\mathbb{1}$ is the indicator function which is equal to 1 if $y \geq x$ and 0 otherwise. The integral quantifies the area between the forecast CDF and the step function representing the observation. A perfect forecast would result in a CRPS of 0, indicating no difference between the forecasted and observed distributions [118].

The CRPS is inherently a proper score, which implies that the expected score is minimised when the forecast distribution corresponds to the true distribution of the outcomes. This characteristic ensures that the CRPS encourages honest reporting of the forecast probabilities. Moreover, the CRPS is a strictly proper score, meaning it is minimised only by the true distribution. Therefore, it is widely employed in the meteorological sciences and increasingly in other fields requiring reliable probabilistic forecasts.

The implementation of CRPS is straightforward when the forecast distribution is expressed parametrically, as integrals of standard distribution functions are typically available. Numerical integration methods may be employed to evaluate the CRPS for non-parametric distributions.

In the context of ensemble forecasts, where a set of simulations represents the forecast distribution, the CRPS can be computed using the empirical CDF of the ensemble members. The ensemble CRPS thus allows for a nuanced assessment of ensemble forecast performance, including evaluating both the calibration and sharpness of the probabilistic forecasts.

The utility of CRPS in practical scenarios is further enhanced by its decomposability into terms that measure different aspects of forecast quality, such as reliability, resolution, and uncertainty. This decomposition allows researchers to diagnose specific areas where forecasts may be improved, facilitating a more targeted approach to forecast enhancement.

Given its favourable properties and flexibility, the CRPS has emerged as a cornerstone for verifying probabilistic forecasts, fostering advancements in forecast systems to cater for the inherent uncertainty in real-world scenarios.

Negative Log-Likelihood (NLL) NLL [119, 120] is a statistical measure that is widely employed in the realms of statistical and ML models to assess the performance of a model in terms of its ability to represent the data. The NLL function quantifies the disparity between the predicted probability distribution by a model and the actual distribution of the data. It is inherently a measure of loss, where lower values correspond to models with better predictive accuracy.

For a given continuous probability model, the likelihood function $\mathcal{L}(\theta|X_0)$ measures the probability of the observed data X_0 under the parameter set θ. Consequently, the log-likelihood is expressed as the logarithm of the likelihood function:

$$\log \mathcal{L}(\theta|X_0) = \sum_{i=1}^{N} \log f_{PD}(x_i|\theta), \tag{3.24}$$

where $f_{PD}(x_i|\theta)$ denotes the probability density or mass function for the i-th observation. Taking the negative of the log-likelihood gives us the NLL:

$$NLL(\theta|X) = -\sum_{i=1}^{N} \log f(x_i|\theta). \tag{3.25}$$

Minimising the NLL corresponds to maximising the likelihood, a fundamental principle in statistical inference known as the Maximum Likelihood Estimation (MLE). The MLE seeks to find the parameter values that make the observed data most probable, which is particularly advantageous for its consistency and efficiency properties under mild regularity conditions [120].

The application of NLL is not confined to purely statistical models. In ML, particularly in the training of neural networks, NLL serves as a loss function for classification and regression tasks when the model outputs can be interpreted as probabilities. It is especially prevalent in models where the outputs are probabilities, such as logistic regression for binary classification or softmax for multi-class classification.

A distinct advantage of NLL is that it naturally penalises incorrect predictions with a high degree of certainty, which aligns with the intuition that confident but incorrect predictions should incur a larger penalty. This property encourages the model to be accurate and calibrated, meaning that predicted probabilities should reflect true probabilities.

Regularisation terms are often added to the NLL to control for model complexity and prevent overfitting. Such an approach ensures that the model fits the training data well and generalises to new, unseen data. The versatility and theoretical soundness of the NLL make it a cornerstone in developing probabilistic models, as it fosters the creation of both interpretable and robust models.

Check Score The Check Score [119], also known in the literature as the tick score or threshold Brier score, is a verification tool for probabilistic forecasts. This score is particularly useful when assessing the accuracy of predictions for a specific event occurrence. The Check Score is formulated as a generalisation of the Brier score. It focusses on a binary event with a threshold value, thus allowing for the evaluation of forecasts against the actual outcome in a probabilistic manner.

Given a set of forecasts and corresponding observations, the Check Score can be expressed mathematically for a forecast-observation pair (g_i, o_i) as follows:

$$\widetilde{S}(g_i, o_i) = \begin{cases} (1 - g_i)^2, & \text{if } o_i = 1, \\ g_i^2, & \text{if } o_i = 0, \end{cases} \tag{3.26}$$

where g_i is the forecast probability for the i-th instance, and o_i is the binary observation indicating the presence (1) or absence (0) of the event [121].

The Check Score is advantageous because it evaluates the calibration of forecast probabilities and the resolution of the forecast system. It penalises both the lack of reliability in the forecast probabilities and the sharpness of the forecasts, which refers to the concentration of forecast probabilities away from the extremes.

Implementing the Check Score across a suite of forecasts provides a robust assessment of the model's performance. Moreover, it affords meaningful insights into the forecast system, which can be instrumental in informing model improvements and guiding decision-making processes where probabilistic forecasts play a pivotal role.

Interval Score NLL serves as a cornerstone for evaluating probabilistic models. However, it solely focuses on point forecasts, neglecting the potential value of information regarding predictive uncertainty. To address this limitation, alternative scoring rules that incorporate point predictions and the associated uncertainty quantification have been proposed. One such prominent scoring rule is the Interval Score (IS) [122].

IS measures the average disagreement between the predicted probability distribution $\widetilde{P}$ and the observed outcome y [123]. It is calculated as the integral of the absolute difference between the predicted CDF $\widetilde{P}(x)$ and the indicator function $\mathbb{1}(x \geq y)$ representing the true outcome:

$$IS(\widetilde{P}, y) = \int_{-\infty}^{\infty} |\widetilde{P}(x) - \mathbb{1}(x \geq y)| dx. \tag{3.27}$$

In simpler terms, the IS penalises the model for both misplaced point predictions and overly narrow or wide confidence intervals. A well-calibrated model with accurate uncertainty estimates will achieve a lower IS than a model with poorly calibrated predictions or overly narrow confidence intervals.

IS offers several advantages over NLL. Firstly, it incorporates uncertainty information, encouraging models to predict point values and quantify the associated confidence. Secondly, IS is a proper scoring rule, implying that a model minimising IS on average recovers the actual underlying distribution [123]. This property makes IS particularly appealing for tasks where accurate probabilistic forecasts are crucial.

Here are some additional points to consider:

IS can be challenging to compute for complex probability distributions. However, efficient numerical integration methods can alleviate this issue. IS is sensitive to outliers, potentially leading to higher scores even for well-calibrated models with occasional extreme predictions. In conclusion, IS is a valuable tool for evaluating probabilistic models, particularly when uncertainty quantification is essential. By incorporating point predictions and uncertainty information, IS offers a more holistic assessment of model performance than NLL.

Statistical Comparison

Wilcoxon Signed-Rank Test The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two related samples or repeated measurements on a single sample to assess whether their population mean ranks differ. It is an alternative to the paired Student's t-test when the population cannot be assumed to be normally distributed [57, 124]. The test is applicable in situations such as evaluating the effect of an intervention where the measurements are taken before and after the intervention on the same subjects.

The test involves ranking the absolute differences between the pairs of samples without considering the signs, then applying signs to the ranks based on the sign of the differences. The sum of the positive ranks (W+) and the sum of the negative ranks (W-) are calculated, and the test statistic is the smaller of these sums. The null hypothesis of the test, which suggests that the median difference between the pairs is zero, is rejected if the calculated statistic is smaller than the critical value from the Wilcoxon signed-rank test distribution for a given significance level [125].

It is essential to note that the Wilcoxon signed-rank test assumes that the differences are symmetrically distributed about the median and that the data are measured at least on an ordinal scale. This test is highly useful in medical, psychological, and other scientific research where the assumptions of parametric tests cannot be met, providing a powerful tool for statistical inference in these domains.

F-test for Interclass Correlation Coefficient F-test for ICCs is a statistical test used to evaluate the reliability of measurements or ratings within a class or group. ICC assesses the degree of agreement or consistency among different raters or measurements and is particularly useful in studies where measurements are made on the same subjects by different raters or under different conditions. The test is based on an analysis of variance (ANOVA) framework and can be used to determine if the variation between groups is significantly larger than the variation within groups [126, 127].

ICC is calculated by comparing the variability of different ratings of the same subject to the total variability across all ratings and subjects. F-test is then used to determine if the observed ICC is significantly greater than zero, indicating that the measurements are not just random but show a degree of consistency or agreement.

In practice, the F-test for ICC is applied in various fields such as psychology, education, and medical research to ensure that the instruments or raters used in a study are reliable. The test assumes that the data are normally distributed and that the subjects are randomly selected. There are different forms of ICC depending on the study design and the assumptions about the raters and subjects [128].

Evaluation Process

This section discusses the evaluation metrics used to assess both probabilistic and point performance.

For point performance evaluation, of **Method I**, the following metrics were utilised: RMSE [107, 108, 109], MAE [109, 110], MAPE [107, 111, 112], R^2 [113, 114] and correlation coefficient [115, 116, 117]. To evaluate the probabilistic performance, of **Method I**, we employed the following metrics, where lower values indicate better performance: CRPS [118], NLL [120], Check Score [121] and IS [123]. We also compared the performance of **Method I** with three recent gradient-boosting algorithms that provide probabilistic predictions: Natural gradient boosting (NGBoost) [106], probabilistic gradient boosting machines (PGBMs) [129] and CatBoost with uncertainty (CBU) [130]. To facilitate a comprehensive understanding of our results, we provide conditional output distributions and the respective confidence intervals (CIs) for two high- and two low-accuracy (test set) predictions.

We assessed the relative importance of each of the eight 2DE RV views (features) for predicting RV volumes using gain metric-powered explainability analysis. In summary, we first used GBRTs as ensemble models to predict RV volumes from the areas of the eight standardised RV views [131]. Following that, we calculated the relative feature importance (RFI) of each view feature to the RV volume. We did not distinguish between RVEDV and RVESV when determining the RFI of RV views.

The evaluation metrics used to assess only point performance for **Method II** were RMSE [107, 108, 109], MAE [109, 110], MAPE [107, 111, 112], and R^2 [113, 114]. These point performance metrics provide insights into the accuracy of our predictions and their alignment with the reference (CMR) values. In addition to evaluating **Method II**, we also compared it with other SOTA methods for tabular data. These methods included both shallow tree-based ensemble models such as XGBoost [103] and CatBoost [105], as well as attention-based deep architectures like the Tab-Transformer [132]. All SOTA methods were implemented using recommended hyperparameter values from [99]. The Wilcoxon signed-rank test [57, 124, 125] was used to evaluate the statistical differences between the proposed and SOTA methods. Apart from volume prediction, we extended our evaluation to include RV ejection fraction (RVEF) values. RVEF is a critical clinical metric used to describe the percentage of blood leaving the right ventricle with each contraction. We visualised and compared the predicted volumes and calculated EF values against the reference (CMR) values for **Method II** and the other SOTA methods to provide a holistic view of the results. These visualisations offer valuable insights into the performance and alignment of our approach with established methods. The training and evaluation phases were conducted across various combinations of the eight available RV views, guided by insights from the ML model explainability analysis. Our objective was to pinpoint a set of up to three RV views that would sustain high diagnostic accuracy. This specific limit was imposed to enhance the clinical feasibility of the approach. Sensitivity analysis was undertaken for combinations encompassing two to four views, aiming to identify the most informative set that supports accurate RV volume and function estimation.

3.3 Results

3.3.1 Variability Analysis of 2DE RV Planimetry

Measurements of areas from tracings of all available RV views were assessed for variability in both ED and ES. Results are shown in Tables 3.2, 3.3 for intraobserver and Tables 3.4, 3.5 for interobserver [91]. We found poor reliability of the planimetered areas for the SubC view in both ED and ES and poor to moderate reliability for the PSAXmid and PSAXdistal views in both ED and ES. However, tracings from all other views showed good (ICC=0.75-0.90) to excellent (ICC>0.90) intra- and interobserver variability.

Table 3.2: Results for intraobserver variability in end-diastole. Intraclass correlation coefficients for intraobserver reliability of measured RV tracings. Based on a 2-way fixed effects model and absolute agreement.

RV tracing	**Type**	**Intraclass Correlation Coefficient**	**95% Confidence Interval**		**F-test with true value 0**	
			Lower Bound	**Upper Bound**	**Value**	***p*-value**
PLAX	Single	0.891	0.452	0.975	27.59	<0.001
	Average	0.798	0.247	0.949		
RVInflow	Single	0.679	0.041	0.915	8.68	0.002
	Average	0.809	0.043	0.955		
PSAXAV	Single	0.851	-0.032	0.972	43.99	<0.001
	Average	0.920	-0.065	0.985		
PSAXbase	Single	0.760	-0.069	0.952	32.40	<0.001
	Average	0.863	-0.148	0.975		
PSAXmid	Single	0.455	-0.139	0.824	2.79	0.071
	Average	0.625	-0.323	0.903		
PSAXdistal	Single	0.570	0.021	0.868	4.27	0.021
	Average	0.726	0.040	0.929		
FourC	Single	0.720	0.076	0.937	16.59	<0.001
	Average	0.837	-0.164	0.967		
SubC	Single	0.362	-0.378	0.798	2.03	0.154
	Average	0.532	-1.218	0.888		

PLAX: Parasternal Long Axis, RVInflow: Right Ventricular Inflow, PSAXAV: Parasternal Short Axis at the Level of the Aortic Valve, PSAXbase: Parasternal Short Axis at the Base of the Left Ventricle, PSAXmid: Parasternal Short Axis at the Mid Left Ventricle, PSAXdistal: Parasternal Short Axis at the Apex of the Left Ventricle, FourC: Standard Four Chamber, SubC: Subcostal Four Chamber, RV: Right Ventricular.

Table 3.3: Results for intraobserver variability for end-systole. The F-test intraclass correlation coefficients, and 95% CIs for intraobserver reliability of measured RV tracings. Based on a 2-way fixed effects model and absolute agreement.

RV tracing	**Type**	**Intraclass Correlation Coefficient**	**95% Confidence Interval**		**F-test with true value 0**	
			Lower Bound	**Upper Bound**	**Value**	***p*-value**
PLAX	Single	0.882	0.230	0.975	32.68	<0.001
	Average	0.937	0.374	0.987		
RVinflow	Single	0.831	0.257	0.960	17.94	<0.001
	Average	0.907	0.409	0.979		
PSAXAV	Single	0.860	-0.015	0.973	44.76	<0.001
	Average	0.925	-0.031	0.986		
PSAXbase	Single	0.658	-0.088	0.922	17.32	<0.001
	Average	0.794	-0.195	0.960		
PSAXmid	Single	0.561	-0.048	0.870	5.45	0.009
	Average	0.719	-0.102	0.937		
PSAXdistal	Single	0.407	-0.162	0.800	2.58	0.087
	Average	0.578	-0.388	0.889		
FourC	Single	0.682	-0.064	0.922	11.52	0.001
	Average	0.811	-0.138	0.959		
SubC	Single	0.454	-0.262	0.833	2.50	0.094
	Average	0.624	-0.712	0.909		

PLAX: Parasternal Long Axis, RVInflow: Right Ventricular Inflow, PSAXAV: Parasternal Short Axis at the Level of the Aortic Valve, PSAXbase: Parasternal Short Axis at the Base of the Left Ventricle, PSAXmid: Parasternal Short Axis at the Mid Left Ventricle, PSAXdistal: Parasternal Short Axis at the Apex of the Left Ventricle, FourC: Standard Four Chamber, SubC: Subcostal Four Chamber, RV: Right Ventricular, CI: Confidence Interval.

Table 3.4: Results for interobserver variability for end-diastole. The F-test intraclass correlation coefficients, and 95% CIs for interobserver reliability of measured RV tracings. Based on a 2-way random effects model and absolute agreement.

RV tracing	**Type**	**Intraclass Correlation Coefficient**	**95% Confidence Interval**		**F-test with true value 0**	
			Lower Bound	**Upper Bound**	**Value**	***p*-value**
PLAX	Single	0.975	0.909	0.993	77.57	<0.001
	Average	0.987	0.952	0.997		
RVinflow	Single	0.937	0.782	0.983	32.45	<0.001
	Average	0.967	0.878	0.991		
PSAXAV	Single	0.936	0.781	0.983	31.11	<0.001
	Average	0.967	0.877	0.991		
PSAXbase	Single	0.862	0.473	0.965	18.32	<0.001
	Average	0.926	0.642	0.982		
PSAXmid	Single	0.648	0.040	0.901	4.33	0.020
	Average	0.786	0.077	0.947		
PSAXdistal	Single	0.944	0.781	0.986	42.15	<0.001
	Average	0.971	0.877	0.993		
FourC	Single	0.930	0.723	0.982	34.66	<0.001
	Average	0.964	0.839	0.991		
SubC	Single	0.282	-0.352	0.752	1.80	0.197
	Average	0.440	-1.08	0.858		

PLAX: Parasternal Long Axis, RVInflow: Right Ventricular Inflow, PSAXAV: Parasternal Short Axis at the Level of the Aortic Valve, PSAXbase: Parasternal Short Axis at the Base of the Left Ventricle, PSAXmid: Parasternal Short Axis at the Mid Left Ventricle, PSAXdistal: Parasternal Short Axis at the Apex of the Left Ventricle, FourC: Standard Four Chamber, SubC: Subcostal Four Chamber, RV: Right Ventricular, CI: Confidence Interval.

Table 3.5: Results for interobserver variability for end-systole. Intraclass correlation coefficients for interobserver reliability of measured RV tracings. Based on a 2-way random effects model and absolute agreement.

RV tracing	**Type**	**Intraclass Correlation Coefficient**	**95% Confidence Interval**		**F-test with true value 0**	
			Lower Bound	**Upper Bound**	**Value**	***p*-value**
PLAX	Single	0.956	0.836	0.988	40.85	<0.001
	Average	0.977	0.910	0.994		
RVinflow	Single	0.931	0.750	0.982	25.78	<0.001
	Average	0.964	0.857	0.991		
PSAXAV	Single	0.926	0.732	0.981	30.37	<0.001
	Average	0.961	0.845	0.990		
PSAXbase	Single	0.916	0.285	0.983	51.47	<0.001
	Average	0.956	0.443	0.991		
PSAXmid	Single	0.836	0.470	0.956	10.30	0.001
	Average	0.911	0.639	0.977		
PSAXdistal	Single	0.974	0.906	0.993	82.21	<0.001
	Average	0.987	0.951	0.996		
FourC	Single	0.913	0.709	0.977	22.46	<0.001
	Average	0.954	0.829	0.988		
SubC	Single	0.368	-0.334	0.797	2.09	0.144
	Average	0.538	-1.004	0.887		

PLAX: Parasternal Long Axis, RVInflow: Right Ventricular Inflow, PSAXAV: Parasternal Short Axis at the Level of the Aortic Valve, PSAXbase: Parasternal Short Axis at the Base of the Left Ventricle, PSAXmid: Parasternal Short Axis at the Mid Left Ventricle, PSAXdistal: Parasternal Short Axis at the Apex of the Left Ventricle, FourC: Standard Four Chamber, SubC: Subcostal Four Chamber, RV: Right Ventricular.

3.3.2 Method I: Results

Accuracy

The final set of hyperparameters for each method and the corresponding tuning and training times are listed in Table 3.6. In Table 3.7, the point performances of **Method I** (with CatBoost as the base learner) and three SOTA probabilistic prediction methods are provided. Overall, **Method I** displayed the best performance. Table 3.8 compares the probabilistic performance of all methods. **Method I** (with CatBoost as the base learner) provided the lowest average scores in all CRPS, NLL, Check Score and Interval Score indices. Table 3.9 shows the importance of variance calibration in the probabilistic performance of **Method I**. Table 3.10 demonstrates that the logistic (parametric) distribution better fits the underlying data than assuming normality.

We also illustrate the conditional output distributions for four representative test cases (two that were predicted with high accuracy and two that were predicted with low accuracy when normal (Figures 3.3 and 3.4) and logistic (Figures 3.5 and 3.6) probabilistic density functions were used for modelling, respectively. Table 3.11 lists the above cases' 95% and 99% CIs. These results showcase the appropriateness of the proposed framework for providing uncertainty scores for RV volume predictions.

Importance of 2DE views for RV volume prediction

The outcomes of the "Gain" explainability analysis are illustrated in Figure 3.7. The analysis identified that the three paramount views, as determined by their Relative Feature Importance (RFI), were the orthogonal views: PLAX, with an RFI of 0.215; FourC, with an RFI of 0.132; and PSAXbase, with an RFI of 0.127. The fourth significant view was identified as RVInflow, exhibiting an RFI of 0.105. Notably, age and gender were observed to be the least influential factors, with RFIs of 0.042 and 0.030, respectively, in predicting RV volumes.

Table 3.6: The final set of hyperparameters used for each method. The tuning and training times are also shown. **Method I** was applied to CatBoost, XGBoost, and LightGBM.

Parameter	CatBoost	XGBoost	LightGBM	NGBoost	PGBM	CBU
k	5	15	3	-	-	-
δ_f	1	0.5	0.5	5	10	1
Operation	add	mult	mult	add	add	add
Minimum scale	6.164	13.826	2.055	-	-	-
Estimators (trees)	100	25	25	244	250	250
Maximum depth	5	2	-1	-	-	-
Learning rate	0.1	0.1	0.1	0.01	0.01	0.1
Minimum data in leaf	1	-	-	-	20	1
Minimum child weight	-	20	20	-	-	-
Number of leaves	-	-	15	-	15	15
Maximum bin	255	255	255	255	255	255
Tune+train time (s)	67.369	19.932	14.260	5.370	872.663	81.929

CatBoost: Categorical Boosting, XGBoost: Extreme Gradient Boosting, LightGBM: Light Gradient Boosting, NGBoost: Natural Gradient Boosting, PGBM: Probabilistic Gradient Boosting Machine, CBU: Categorical Boosting with Uncertainty.

Table 3.7: Point performance comparison on the test set (five folds). **Method I** results are for the case when CatBoost was the base learner. Boldface indicates the best performance.

Method	MAE	RMSE	MAPE	R^2	Correlation
Method I	**22.75**	**26.292**	**20.22**	**0.666**	**0.824**
NGBoost	28.114	32.269	24.406	0.496	0.736
PGBM	27.479	31.27	25.147	0.527	0.768
CBU	26.378	30.127	22.974	0.561	0.772

NGBoost: Natural Gradient Boosting, PGBM: Probabilistic Gradient Boosting Machine, CBU: Categorical Boosting with Uncertainty, MAE: Mean Absolute Error, RMSE: Root Mean Squared Error, MAPE: Mean Absolute Percentage Error.

Table 3.8: Probabilistic performance comparison on the test set (five folds). **Method I** results are for the case when CatBoost was the base learner. **Method I** results have been averaged over all nine posterior output distributions. Boldface indicates the best performance.

Method	NLL	CRPS	Check Score	IS
Method I	**4.747**	**15.398**	**7.775**	**73.380**
NGBoost	7.571	22.174	11.177	141.618
PGBM	6.136	20.796	10.492	122.401
CBU	5.780	19.524	9.853	110.140

NGBoost: Natural Gradient Boosting, PGBM: Probabilistic Gradient Boosting Machine, CBU: Categorical Boosting with Uncertainty, NLL: Negative Log-Likelihood, CRPS: Continuous Ranked Probability Score, IS: Interval Score.

Table 3.9: Probabilistic performance comparison of **Method I** with and without variance calibration. **Method I** results are for the case when CatBoost was the base learner. Boldface indicates the best performance.

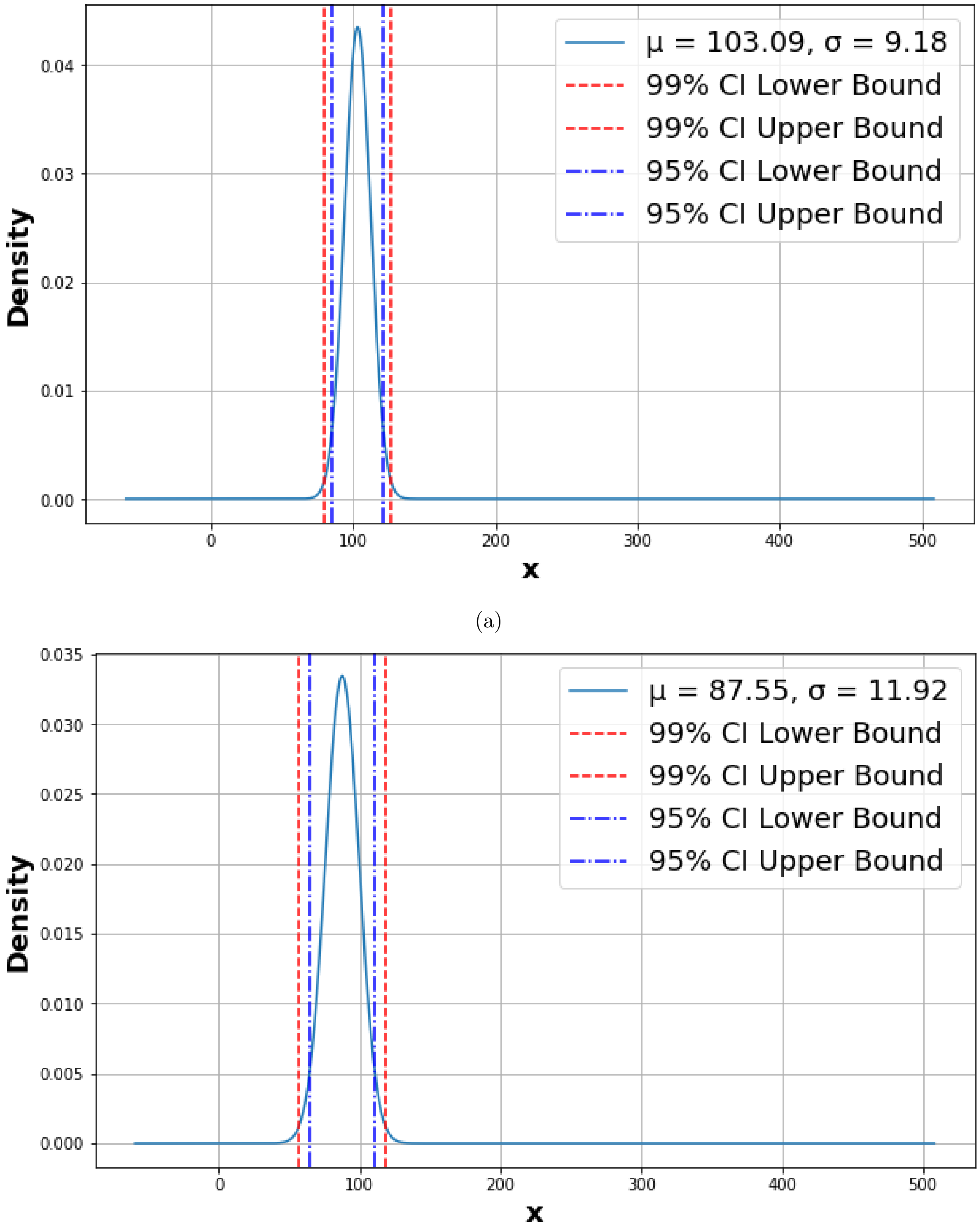
Operation	NLL	CRPS	Check Score	IS
With Calibration	**4.747**	**15.398**	**7.775**	**73.38**
Without Calibration	4.781	15.457	7.805	74.044

NLL: Negative Log-Likelihood, CRPS: Continuous Ranked Probability Score, IS: Interval Score.

Table 3.10: Probabilistic performance comparison when assuming normal and logistic distributions for modelling the underlying data. Boldface indicates the best performance.

Distribution	NLL
Normal	5.10466
Logistic	**5.00837**

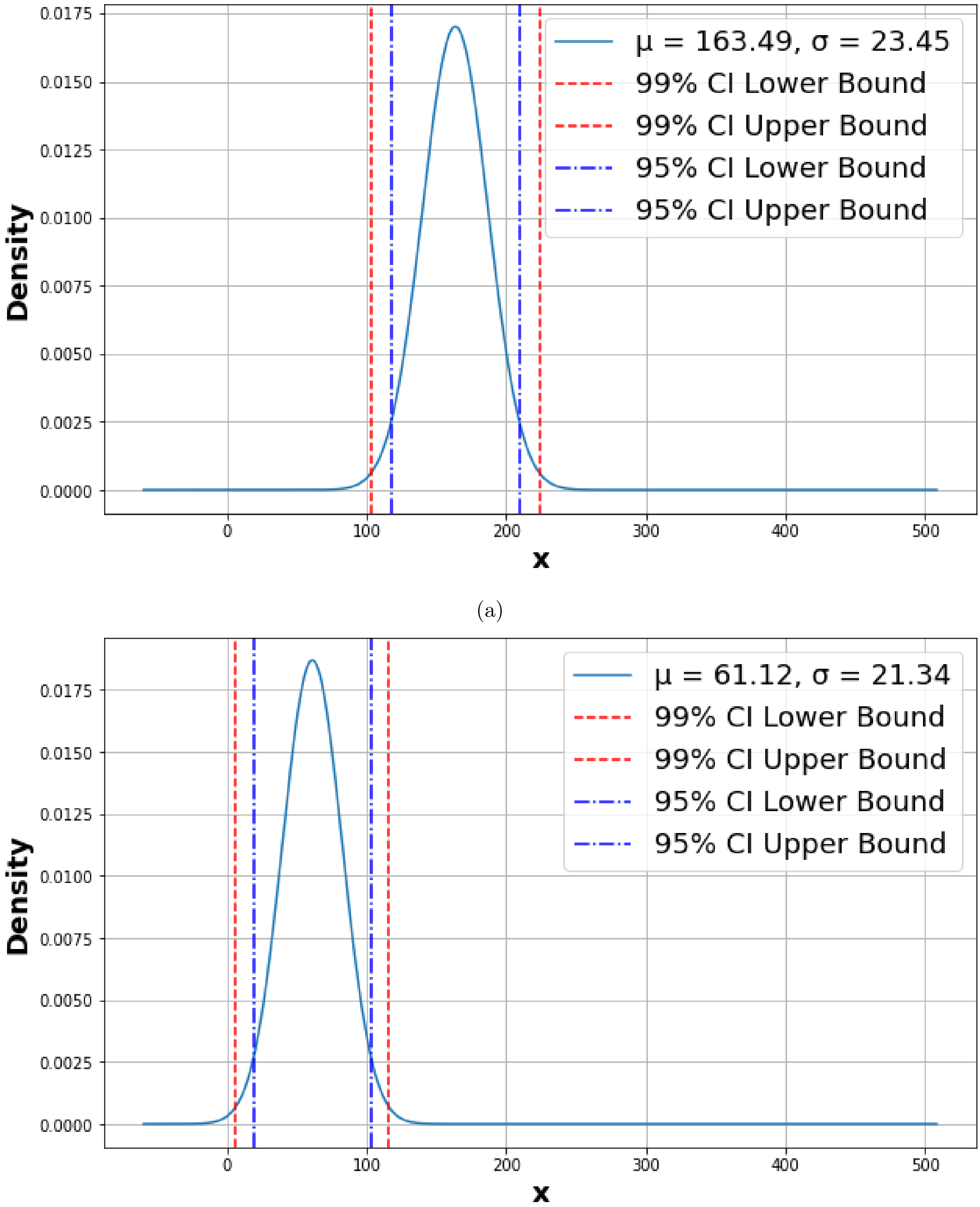
NLL: Negative Log-Likelihood.



(a)

(b)

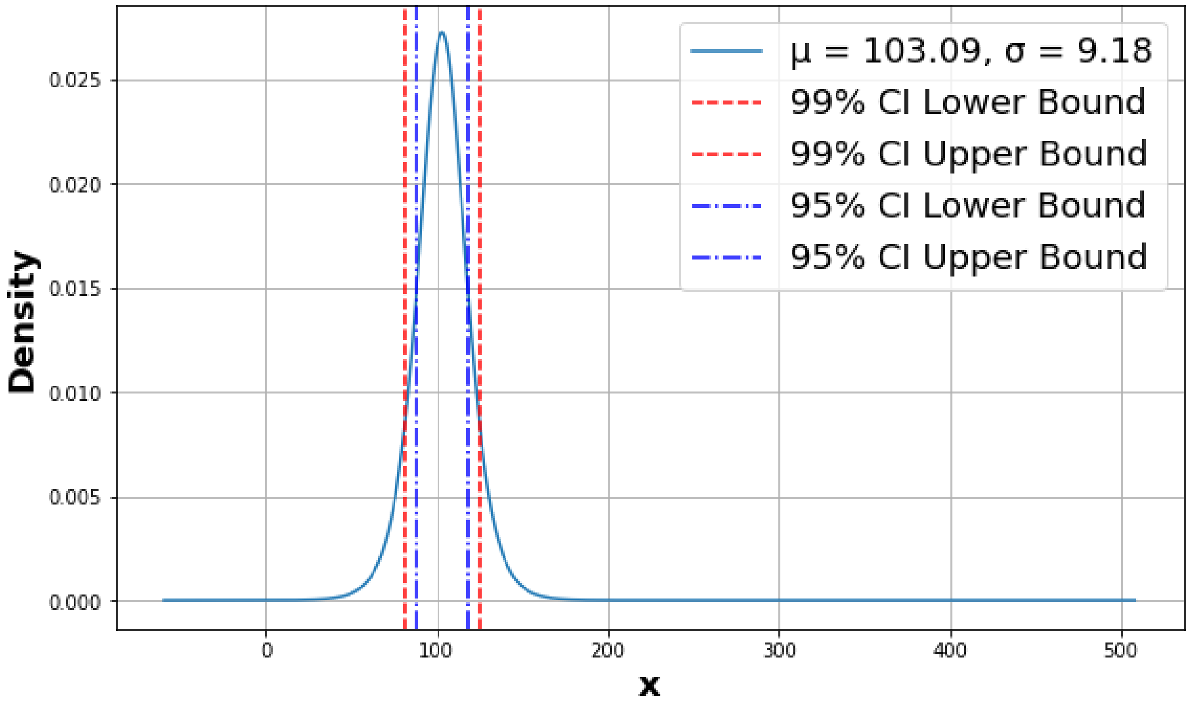
Figure 3.3: The conditional output normal distributions for two test instances that were predicted with high accuracy.



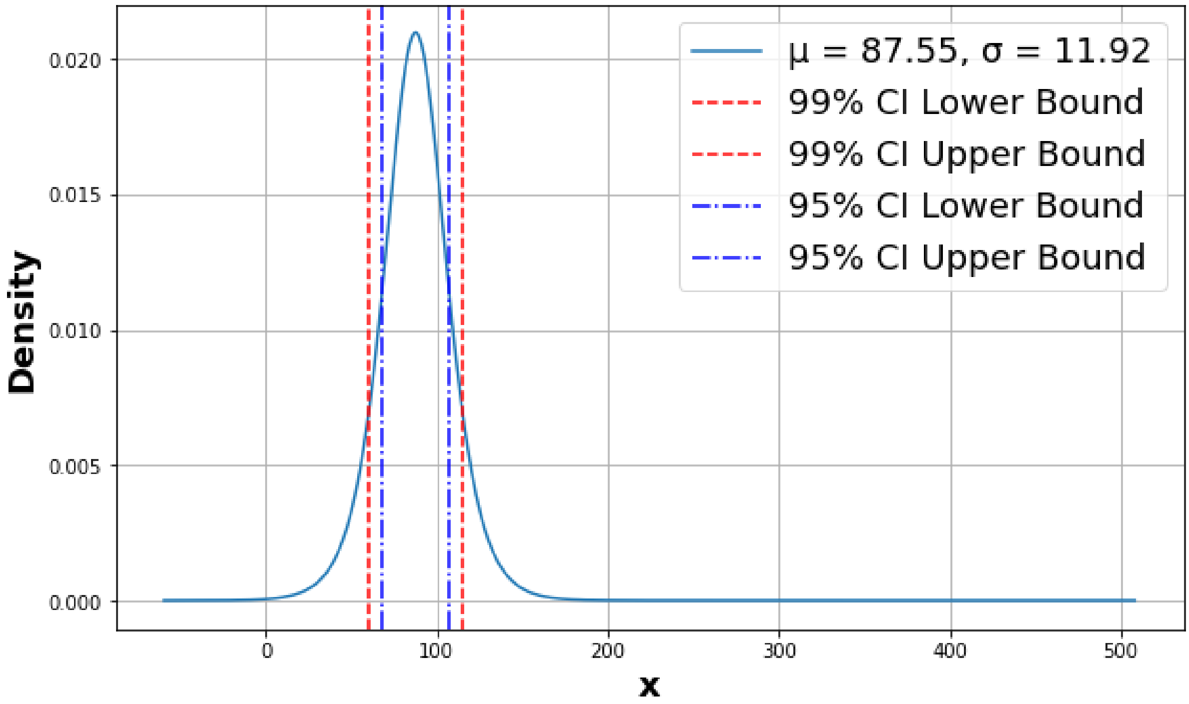
(a)

(b)

Figure 3.4: The conditional output normal distributions for two test instances that were predicted with low accuracy.

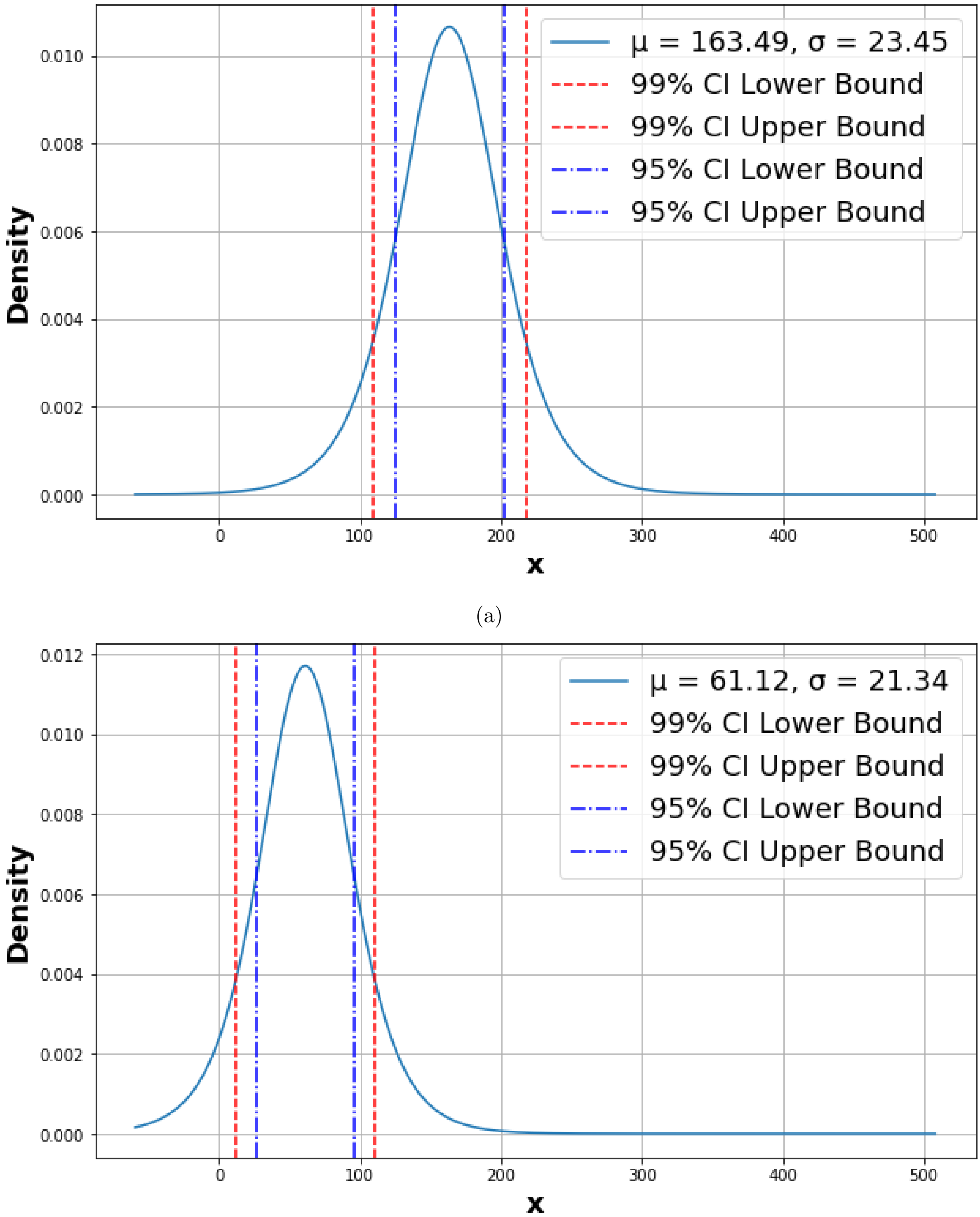


(a)



(b)

Figure 3.5: The conditional output logistic distributions for two test instances that were predicted with high accuracy.



(a)

(b)

Figure 3.6: The conditional output logistic distributions for two test instances that were predicted with low accuracy.

Table 3.11: The 95% and 99% Confidence Intervals (CIs) for both normal and logistic distributions for four representative test set cases, two that were predicted with high accuracy (low APE) and two that were predicted with low accuracy (high APE).

Prediction Accuracy	APE (%)	Point Prediction	Normal Distribution		Logistic Distribution	
			95% CI	99% CI	95% CI	99% CI
High	3.090	103.090	35.979	47.286	30.196	42.697
	0.508	87.553	46.739	61.428	39.227	55.466
Low	10.169	163.492	91.932	120.825	77.157	109.099
	10.119	61.119	83.650	109.939	70.206	99.270

APE: Absolute Percentage Error, CI: Confidence Interval.

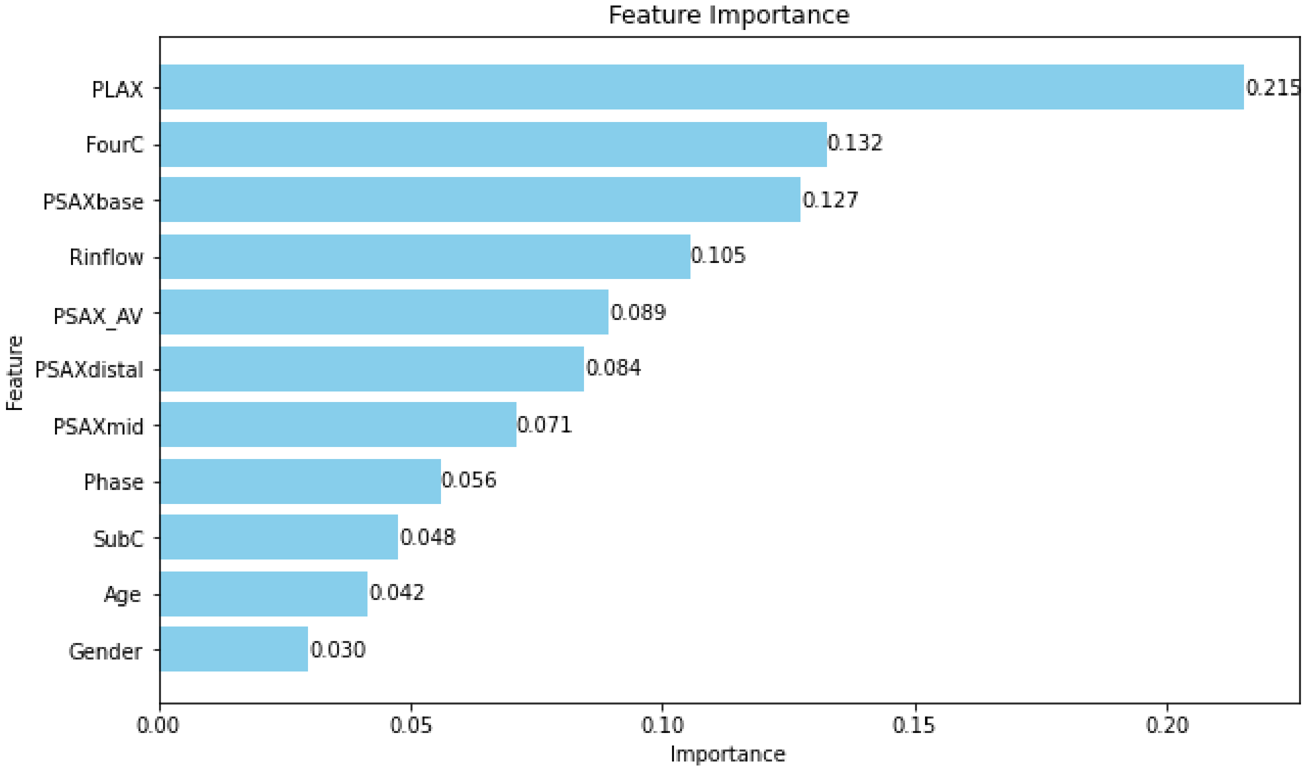


Figure 3.7: Feature importance plot for the **Method I** with CatBoost as the base learner. PLAX: Parasternal Long Axis, RVInflow: Right Ventricular Inflow, PSAXAV: Parasternal Short Axis at the Level of the Aortic Valve, PSAXbase: Parasternal Short Axis at the Base of the Left Ventricle, PSAXmid: Parasternal Short Axis at the Mid Left Ventricle, PSAXdistal: Parasternal Short Axis at the Apex of the Left Ventricle, FourC: Standard Four Chamber, SubC: Subcostal Four Chamber.

3.3.3 Method II: Results

Final hyperparameters

The final hyperparameters of the proposed DL model are shown in Table 3.12 and are compared to the respective values from the initial study that inspired our design.

Table 3.12: The final hyperparameter values employed in **Method II**.

Parameter	**Method II**	Recommended by [99]
# of layers	3	6
Feature embedding size	16	512
Residual dropout	0.3	0.2
Attention dropout	0.3	0.5
Feed forward network dropout	0.3	0.5
Feed forward network factor	4/3	4/3
Learning rate	0.01	LogUniform [3e-5,3e-4]
Weight decay	0	LogUniform [3e-6,3e-3]
Optimiser	Adamax	AdamW

Accuracy

The proposed **Method II** for RV volume prediction achieved good accuracy (Table 3.13, Figure 3.8) with $R^2 = 0.975$ and APE = 5.46% ± 4.87%. It also outperformed SOTA ML algorithms for tabular data, namely XGBoost $\left(R^2 = 0.600, \text{APE} = 15.90\% \pm 18.62\%\right)$, CatBoost $\left(R^2 = 0.797\right.$ APE= 15.13% ± 10.81%) and Tab-Transformer ($R^2 = 0.784$, APE = 21.05% ± 16.74%). Similar accuracy was achieved (Table 3.14, Figure 3.9) for RVEF with APE = 5.80% ± 3.91%. Bland-Altman analysis, presented as mean bias ±95% limits of agreement, also revealed good agreement between CMR and the proposed method for RVEDV (1.27 ± 23.35 mL) and RVESV (−2.61 ± 19.63 mL), and RVEF(−1.97% ± 7.04%) (Figure 3.10).

Using CMR-derived RV volumes and RVEF, one patient had RV dilatation, and three had RV dysfunction in the testing dataset ($n = 10$). Of note, qualitative 2DE analysis did not show any patient with RV dilatation in the testing dataset. Two patients with RV dysfunction were detected of whom none had RV dysfunction by CMR. Therefore, there was no correlation at all between qualitative 2DE analysis and CMR (0% accuracy for both dilatation and dysfunction). **Method II** correctly classified one patient with RV dilatation and did not detect any other patients with RV dilatation (100% diagnostic accuracy). **Method II** correctly identified the three patients with RV dysfunction by CMR (100% sensitivity) and additionally detected a 4th patient that did not have RV dysfunction by CMR (75% specificity), yielding a diagnostic accuracy of 90%.

Verifying Importance of 2DE views for RV volume prediction

In Table 3.15, we display the performance metrics of **Method II** under conditions of varying the number of input views. These specific views were selected following guidance from the feature importance depicted in Figure 3.7. A discernible trend emerges from this analysis: a positive correlation exists between the model's accuracy and the number of views incorporated. Increasing the number of views entered results in increased accuracy.

Moreover, the rigorous statistical examination conducted via the Wilcoxon signed-rank test, as detailed in Table 3.16, further substantiates the impact of view selection on prediction accuracy. Remarkably, the reduction of views from eight to four (PLAX, FourC, PSAXbase and RVInflow) does not significantly impair the model's performance, suggesting that the proposed model retains its predictive robustness over a moderately reduced view spectrum.

However, the scenario alters markedly when the number of views diminishes to less than four. This transition is characterised by a substantial decrement in model accuracy, highlighting a critical inflexion point in the relationship between view availability and model performance. Such a decline in accuracy with reduced views underscores the indispensability of a comprehensive feature set for maintaining the integrity of the model's predictive capability.

These findings collectively advocate for a strategic approach in selecting 2DE views for the proposed model training, emphasising the balance between proposed model simplicity and the retention of essential predictive features. The evidence points to the potential of reducing the of clinicians by 50% by using four 2DE views, thus enhancing the clinical feasibility.

Table 3.13: Quantitative comparison of the predicted RV volumes between **Method II** and other SOTA ML methods. Boldface indicates best performance. *p*-values were obtained from the Wilcoxon signed-rank test ($\alpha = .05$).

Method	R^2 Score	APE (%) mean (±SD)	***p*-value**
Method II	**0.975**	**5.46 (±4.87)**	**-**
TabTransformer	0.784	21.05 (±16.74)	2.1×10^{-5}
CatBoost	0.797	15.13 (±10.81)	6.3×10^{-6}
XGBoost	0.600	15.90 (±18.62)	7.3×10^{-6}

APE: Absolute Percentage Error, SD: Standard Deviation, TabTransformer: Tabular Transformer, CatBoost: Categorical Boosting, XGBoost: Extreme Gradient Boosting, RV: Right Ventricular, ML: Machine Learning, SOTA: State-Of-The-Art.

Table 3.14: Quantitative comparison of the calculated RV ejection fraction between **Method II** and other SOTA ML methods. Boldface indicates best performance. *p*-values were obtained from the Wilcoxon signed-rank test ($\alpha = .05$).

Method	APE in RVEF (%) mean (± SD)	*p*-value
Method II	**5.80 (± 3.91)**	**-**
TabTransformer	11.29 (± 6.11)	0.004
CatBoost	13.48 (± 5.31)	0.048
XGBoost	21.93 (± 20.63)	0.013

APE: Absolute Percentage Error, RVEF: Right Ventricular Ejection Fraction, SD: Standard Deviation, TabTransformer: Tabular Transformer, CatBoost: Categorical Boosting, XGBoost: Extreme Gradient Boosting; RV, Right Ventricular; ML, Machine Learning; SOTA, State-Of-The-Art.

Table 3.15: Summary of **Method II** performance using eight, four or fewer 2DE views as model inputs.

View 1	View 2	View 3	View 4	R^2 Score	APE	SD	MSE	SD
FourC	PLAX			0.946	10.293	16.072	266.061	16.311
FourC	PSAXbase			0.565	18.395	42.144	2147.621	46.342
FourC	RVInflow			0.578	20.345	39.612	2080.598	45.614
FourC	PSAXbase	RVInflow		0.478	22.669	40.099	2573.024	50.725
FourC	PLAX	RVInflow		0.836	15.701	23.267	806.538	28.400
FourC	PLAX	PSAXbase		0.964	8.091	12.068	177.252	13.314
FourC	PLAX	PSAXbase	PSAXAV	0.938	10.137	17.544	307.793	17.544
FourC	PLAX	PSAXbase	RVInflow	0.973	4.907	11.330	131.528	11.469
8 Views				0.975	5.460	4.870	71.860	8.477

FourC: Four Chamber, PLAX: Parasternal Long Axis, PSAXbase: Parasternal Short Axis at the Base of the Left Ventricle, RVInflow: Right Ventricular Inflow, 2DE: Two-Dimensional Echocardiography, APE: Absolute Percentage Error, SD: Standard Deviation, MSE: Mean Squared Error.

Table 3.16: Wilcoxon signed-rank test (p-values) for subsample comparisons.

Subsamples (Volumes)	**Wilcoxon signed-rank test (p-values)**
3 Views Vs 2 Views	0.068010725
3 Views Vs 4 Views	5.97×10^{-9}
3 Views Vs 8 Views	7.62×10^{-9}
4 Views Vs 4 Views(RV_Inflow Vs PSAX_AV)	3.83×10^{-8}
8 Views Vs 4 Views	0.481479371

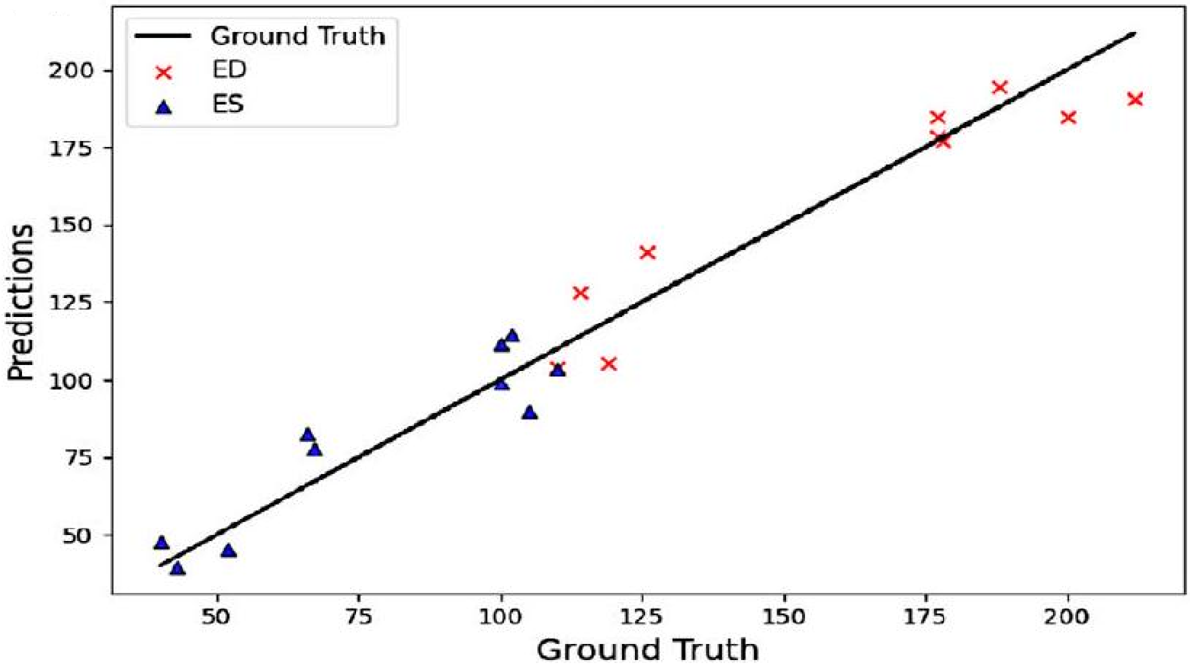


Figure 3.8: Predicted RVEDV and RVESV vs. ground truth (CMR) using the proposed **Method II**. CMR: Cardiac Magnetic Resonance Imaging, RVEDV: Right Ventricular End-Diastolic Volume, RVESV: Right Ventricular End-Systolic Volume.

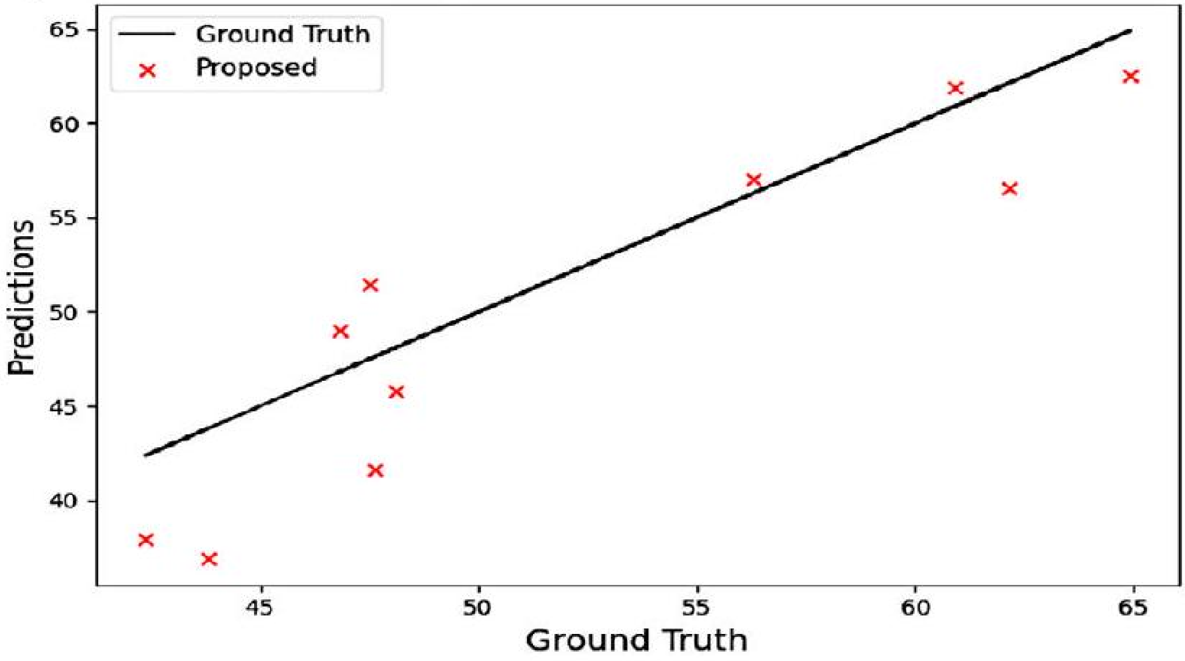
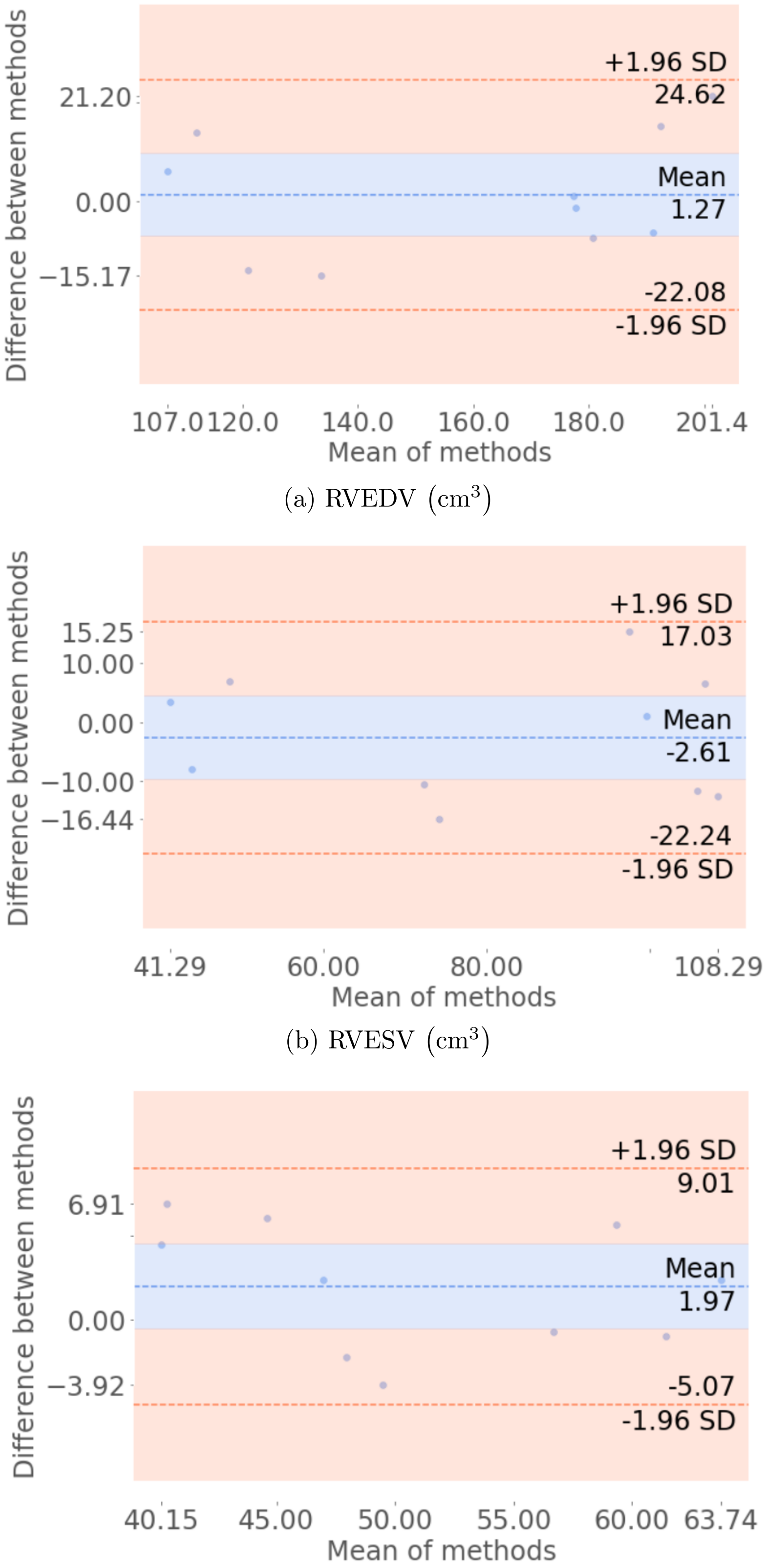


Figure 3.9: Predicted RVEF vs. ground truth (CMR) using the proposed **Method II**. CMR: Cardiac Magnetic Resonance Imaging, RVEF: Right Ventricular Ejection Fraction.



(a) RVEDV (cm^3)

(b) RVESV (cm^3)

(c) RVEF (%)

Figure 3.10: Bland-Altman analysis for predicted RVEDV (a) and RVESV (b), as well as RVEF (c) versus CMR. CMR: Cardiac Magnetic Resonance Imaging, RV, Right Ventricular, RVEDV: Right Ventricular End-Diastolic Volume, RVESV, Right Ventricular End-Systolic Volume, RVEF: Right Ventricular Ejection Fraction.

3.4 Discussion

This discussion summarises the findings from our investigation into the efficacy of ML and DL in enhancing the precision of RV evaluation through non-invasive imaging techniques. This chapter leveraged a novel application of multi-headed attention-based DL algorithms for the volumetric and functional quantification of the RV from 2DE planimetry data, challenging conventional approaches and offering new insights into the future of cardiovascular imaging.

3.4.1 Strengths of this Study

The primary strength of this study lies in the pioneering use of a non-geometric-based DL method for the prediction of RV volumes and RVEF from planimetered 2DE views. This work also questions the superiority of tree-based ensembles over DL for tabular data in the context of cardiovascular imaging. The appropriateness of the proposed tree-based ensemble method provides uncertainty scores for RV volume predictions to alleviate trustworthiness in artificial intelligence and reduce the risks [133]. The methodological rigour, incorporating ML explainability analysis and multi-head attention-based models, sets a new benchmark in the domain. The adoption of the powerful Transformer architectures, traditionally reserved for NLP and computer vision, into the analysis of tabular 2DE data underscores our innovative approach. Moreover, the validation of our models against CMR, the gold standard for RV evaluation, highlights the reliability and accuracy of our findings [133].

3.4.2 Main Findings

Our research demonstrates that accurate (R^2=0.975) RV volume and function predictions can be achieved with a reduced number of 2DE views, challenging the necessity for extensive and time-consuming imaging protocols. The successful application of DL algorithms, devoid of geometric assumptions, enables a closer approximation to CMR accuracy levels, thereby addressing a significant limitation of current 2DE evaluation methods [134]. Importantly, our study reveals that a four-view combination, while practical for clinical application, closely mirrors the predictive accuracy of a eight-view approach, suggesting an optimal balance between efficiency and diagnostic precision.

3.4.3 Clinical Implications

The implications of our findings are manifold. Firstly, the enhanced accuracy in RV evaluation provided by our DL framework can significantly improve the diagnosis and treatment of RV dysfunction, a condition with established morbidity and mortality implications. By offering a non-invasive, accessible, and efficient alternative to CMR, our approach stands to democratise high-quality RV assessment, extending its reach beyond specialised centres to routine clinical practice. This democratisation is crucial, given the prevalence of RV dysfunction and the critical role of timely and accurate diagnosis in patient management.

3.4.4 Resource Efficiency Considerations

The adoption of our DL-based method for RV evaluation promises not only clinical but also operational benefits. By reducing the number of required 2DE views without compromising diagnostic accuracy, our approach enhances the resource efficiency of cardiac imaging. This efficiency is not limited to time savings for healthcare professionals but extends to the potential for reducing the financial burden associated with extensive imaging studies. Furthermore, by serving as a gatekeeper for CMR, our method could optimise the utilisation of high-cost imaging resources, ensuring they are reserved for cases where they add the most value.

3.4.5 Study Limitations

Despite its strengths, this study is not without limitations. The use of a relatively small and select patient cohort may restrict the generalisability of our findings. Moreover, the exclusion of patients with severe RV dysfunction or complex cardiac conditions from our analysis could limit the applicability of our results to the broader population of patients with RV abnormalities. As a result of a small-scale dataset, overfitting was observed in training **Method I**. The "gain" explainability was taken from **Method I** and applied to **Method II**; they are not from the same model. The ground truth (CMR) and the inputs (echocardiography) were not taken on the same day. Future research should aim to validate our DL framework across diverse patient populations, including those with significant RV dysfunction, to enhance its clinical relevance and applicability.

Chapter 4

Fast-tracking the Deep Residual Network Training for Arrhythmia Classification by Leveraging the Power of Dynamical Systems

4.1 Introduction

4.1.1 Background

Many people have irregular heartbeats, which can be fatal in some cases [135]. According to the World Heart Report 2023, CVDs (CVDs) continue to affect more than half a billion people worldwide, accounting for 20.5 million deaths in 2021 [136]. This is close to a third of all deaths globally and an overall increase in the estimated 121 million CVD deaths [136]. With proper precautions, Up to 80% of premature heart attacks and strokes can be prevented [136]. As a result, arrhythmia classification by analysing the widely used electrocardiogram (ECG) signals has received significant attention in recent years [137]. Manual heartbeat analysis is labour-intensive and subject to human errors [138]. Furthermore, clinicians' manual analysis of ECGs can only include a finite number of predictors and cannot execute complex analyses [139], necessitating the advent of automated approaches. Within this context, deep learning (DL) architectures, particularly residual networks [140], have been successfully applied lately towards detecting ECG signal anomalies [141].

4.1.2 The Challenge

Evidence in existing literature underscores the pivotal role of network depth in enhancing the accuracy and capabilities of residual models [142]. However, the computational demands during training accompanying an increased depth pose significant hurdles, limiting such models' practical applicability. In this paper, in search of resource-economical development of systems for machine learning-enabled arrhythmia detection, we propose to adjust the computation workload dynamically during residual network training.

4.1.3 Our Contribution

The contributions of this study are:

- We propose to initiate the training process with a shallow network, gradually increasing its depth as training progresses. To this end, we leverage the dynamical systems perspective of residual networks that conceptualises a network composed of L_{RN} residual blocks as an ordinary differential equation with L_{RN} temporal intervals [143].
- Unlike other residual network training acceleration strategies, which are primarily heuristic [144], the proposed approach is theoretically grounded.
- Our pipeline is evaluated on the large-scale freely-available PhysioNet's Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH) arrhythmia dataset [145, 146].
- Drawing inspiration from some recent deep learning sound classification studies that provided more accurate results in noisy conditions [147], we use heartbeat spectrograms to train deep residual networks. The spectrogram being an image itself means that it aligns seamlessly with the input requisites of deep residual networks.
- Apart from training time reductions, we also gauge savings on energy consumption and environmental costs by using the proposed pipeline.

4.2 Materials and Methods

4.2.1 Study Population and Dataset

This study utilises the PhysioNet MIT-BIH Arrhythmia ECG dataset [145, 146]. In our experiments, we have used only the ECG lead II. The MIT-BIH dataset consists of 47 patients (109446 data points per example at sampling frequency 125 Hz).

4.2.2 Annotation and Pre-processing

The ECG signals are annotated into five classes by at least two cardiologists according to the Association for the Advancement of Medical Instrumentation (AAMI) EC57 standard [148]. Table 4.1 lists the AAMI EC57 annotation standards for the five categories. The predefined five class dataset consists of 87554 ECG signals in the training set and 21892 ECG signals in test set. We randomly chose 10% of the ECG signals from the training set as the validation set. Drawing inspiration from recent DL sound classification studies that provided more accurate results in noisy conditions [147], each heartbeat was converted to a spectogram which was then used to train the deep residual network. A representative heartbeat and the generated spectrogram are illustrated in Figure 4.1.

Table 4.1: Summary of mappings between beat annotations and AAMI EC57 categories.

Category	**Annotations**
Nonectopic beat (N)	• Normal • Left/Right bundle branch block • Atrial escape • Nodal escape
Supraventricular ectopic beat (S)	• Atrial premature • Aberrant atrial premature • Nodal premature • Supra-ventricular premature
Ventricular ectopic beat (V)	• Premature ventricular contraction • Ventricular escape
Fusion beat (F)	• Fusion of ventricular and normal
Unknown beat (Q)	• Paced • Fusion of paced and normal • Unclassifiable

AAMI:Association for the Advancement of Medical Instrumentation.

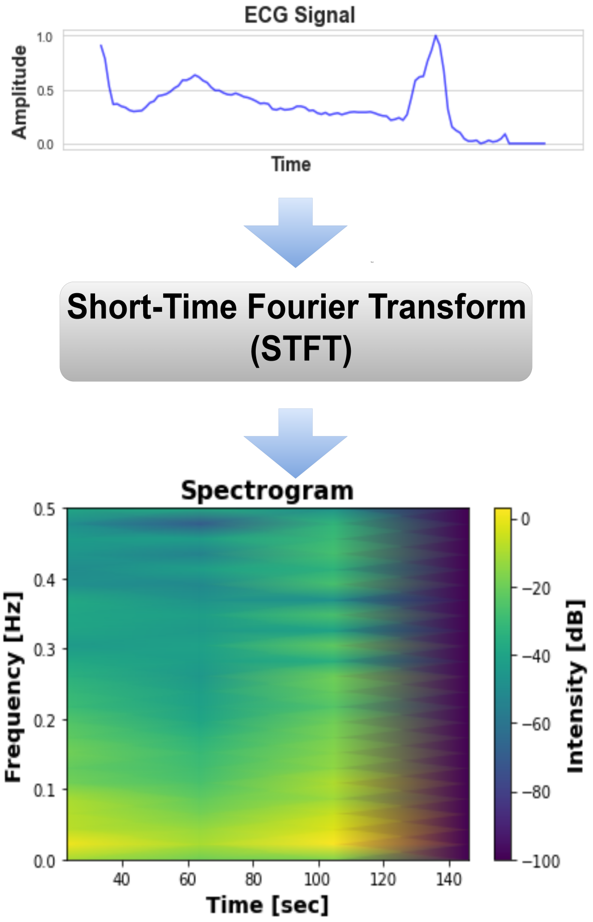


Figure 4.1: Instance of an ECG signal and the generated spectrogram. The brighter the colour, the higher the energy of the signal. ECG: Electrocardiogram, STFT:Short-Time Fourier Transform.

4.2.3 Dynamical Systems Viewpoint

The forward propagation in a residual network block can be expressed as:

$$z_{j+1} = z_j + q\mathcal{E}(z_j, w_j), \qquad j = 0, 1, \ldots, L_{RN}, \tag{4.1}$$

where $\mathcal{E}$ is the residual module, and L_{RN} is the number of layers. Here, $q > 0$ is a sufficiently small parameter that has been included without loss of generality. $\mathcal{E}$ encompasses BN, ReLU activation, and convolutional layers. Eq. (4.1) can be rewritten as

$$\frac{z_{j+1} - z_j}{q} = \mathcal{E}(z_j, w_j), \tag{4.2}$$

Eq. (4.2) can be seen as the forward Euler discretisation for the following initial value ordinary differential equation (ODE).

$$\dot{z}(t) = \mathcal{E}(z(t), w(t)), \quad z(0) = z_0, \quad \text{for} \quad 0 \leq t \leq T_{evol}, \tag{4.3}$$

where features $z(t)$ and parameters $w(t)$ are viewed in their continuous limit as functions of time $t \in [0, T_{evol}]$, the evolution time T_{evol} corresponds to the network depth L_{RN}, $z(0)$ is the input feature map after the initial convolution, and $z(T_{evol})$ is the output feature map before the softmax classifier. Consequently, learning the model parameters $w(t)$, is equivalent to solving an optimal control problem involving the ODE in Eq. (4.3).

The theoretical analyses for the network growing dynamics and the feasibility of effective residual network growing during training can be found in [149].

4.2.4 Automated Adaptive Training Algorithm

We briefly describe the adaptive training algorithm [149] that we borrowed. In this study, we go further and test the algorithm's validity in a different target research area, particularly in a challenging cardiovascular healthcare task.

In each training epoch, the model parameters are updated as usual. Then, a growing scheduler determines if it is needed to increase the network depth. Given that the upper bound of the temporal error is monotonously correlated with the maximum Lipschitz constant of $\mathcal{E}(L_{RN})$, then the growing scheduler is designed to make sure that the Lipschitz constant will not become too large. Specifically, if it exceeds a predetermined risk tolerance r_{tol}, then the network growth is triggered.

Given that a residual block comprises convolutional layers, ReLU activation layers, and BN, then the Lipschitz constant of the aggregate function is simply the product of the individual Lipschitz constants of each component. The latter is suited for efficient calculations [150]. In fact, it has been shown that this Lipschitz constant calculation imposes a negligible overhead on the total training time [151].

To incorporate adaptive growing, the learning rate scheduler is designed such that, after each growth, the cycle in a standard cosine learning rate scheduler is reset as

$$\eta = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos\left(\frac{T_{\text{cur}} - T_{\text{grow}}}{T_{\text{tot}} - T_{\text{grow}}}\pi\right)\right), \tag{4.4}$$

where $\eta_{\min}$ and $\eta_{\max}$ represent the minimum and maximum learning rates, respectively. T_{cur} denotes the current epoch, T_{tot} is the total number of epochs, and T_{grow} refers to the epoch at the last growth occurrence. Initially, $T_{\text{grow}} = 0$.

To ensure effective training, cloning initialisation is applied as a growth method. This method simply clones the residual blocks from the nearest time points of the previous network to populate the new network. This approach also ensures efficient continual optimisation after growth. An implicit step size scaling is also implemented after growth to maintain a roughly constant sum of residuals. The number of layers is doubled in each growth, whereas a certain number of epochs is reserved exclusively for the training of the final model.

4.2.5 Experiments

We used residual networks with 50 and 74 blocks (ResNet-50 and ResNet-74) to train using the proposed adaptive method. For a fair comparison, vanilla ResNet-50 and vanilla ResNet-74 (fixed models) were also trained. The training and testing mini-batch sizes were 128 and 100, respectively. As with the proposed method, we utilised the SGD optimiser and Cross Entropy Loss function. The learning rate was fixed at 0.1. For the optimiser, we chose the weight decay and momentum constants to be equal to 0.0002 and 0.9, respectively. The growth risk tolerance for the Lipschitz constant was (r_{tol} =) 1.4. Each network was grown two times, doubling the depth each time. All networks were trained for 64 epochs (three runs), making sure to reserve for the proposed method at least 20 epochs to train the final model (after the second growth). All experiments were carried out on a workstation with NVIDIA RTX A6000 48GB GPU and Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz CPU.

4.2.6 Evaluation Metrics

To represent the actual training time, we depict the learning curves with respect to the wall clock time (rather than the epoch number). For quantitative evaluation purposes, on top of validation and test accuracy, we also employ the parameters per epoch (PPE) metric [149]. This measure portrays the computational load (memory and processor that were utilised for training), yet it is detached from hardware settings and its usage. Lastly, we present the total carbon dioxide equivalent emissions (CO_2eq) (in g) and energy consumption (in kWh) for each model using the *Carbontracker* method [63], which relies on the hardware type used. The mean and standard deviation for the three training runs are provided for each quantitative measure.

4.3 Results

Figures 4.2 and 4.3 illustrate representative training and validation error rates (error against wall clock time in minutes) for ResNet-74. By inspection, it is apparent that the proposed method exhibits substantial training acceleration while achieving similar accuracy at the end.

Table 4.2 lists the validation and test accuracies achieved by the proposed (adaptive) and the vanilla (fixed) training methods. The proposed method has slightly higher values for both ResNet-50 and ResNet-74.

Table 4.3 outlines the CO_2eq (in g), energy spent (in kWh) and PPE for the two training methods. In regard to ResNet-74, the proposed method produced 44.45% less environmental pollution, consumed 38.78% less energy, and required 45.21% less parameters to train when compared to the vanilla method. Similarly, regarding ResNet-50, the proposed method emits 32.73% less carbon emissions, is 24.57% more energy efficient, and requires 39.47% fewer parameters to train compared with the vanilla method.

Table 4.2: Comparison of proposed and Vanilla training methods for ResNet-50 and ResNet-74. Validation and test accuracies are presented as mean ± standard deviation over the three training runs. Boldface indicates best performance.

Method	**ResNet**	**Validation Accuracy (%)**	**Test Accuracy (%)**
Proposed	**50**	$\mathbf{96.33 \pm 0.46}$	$\mathbf{97.14 \pm 0.08}$
	74	$\mathbf{96.41 \pm 0.42}$	$\mathbf{97.10 \pm 0.05}$
Vanilla	50	96.19 ± 0.44	97.09 ± 0.03
	74	96.38 ± 0.51	96.93 ± 0.16

ResNet: Residual Network.

Table 4.3: Comparison of proposed and Vanilla training methods for ResNet-50 and ResNet-74 in terms of CO_2eq, energy consumption, and Parameters Per Epoch (PPE). Results are presented as mean ± standard deviation over the three training runs. Boldface indicates best performance.

Method	**ResNet**	**CO_2eq (g)**	**Energy (kWh)**	**PPE** ($\times 10^6$)
Proposed	**50**	**411.34**	**1.32**	$\mathbf{0.46 \pm 0.01}$
	74	**410.09**	**1.31**	$\mathbf{0.63 \pm 0.06}$
Vanilla	50	611.49	1.75	0.76
	74	738.23	2.14	1.15

ResNet: Residual Network, CO_2eq: Carbon Dioxide Equivalent, PPE: Parameters Per Epoch.

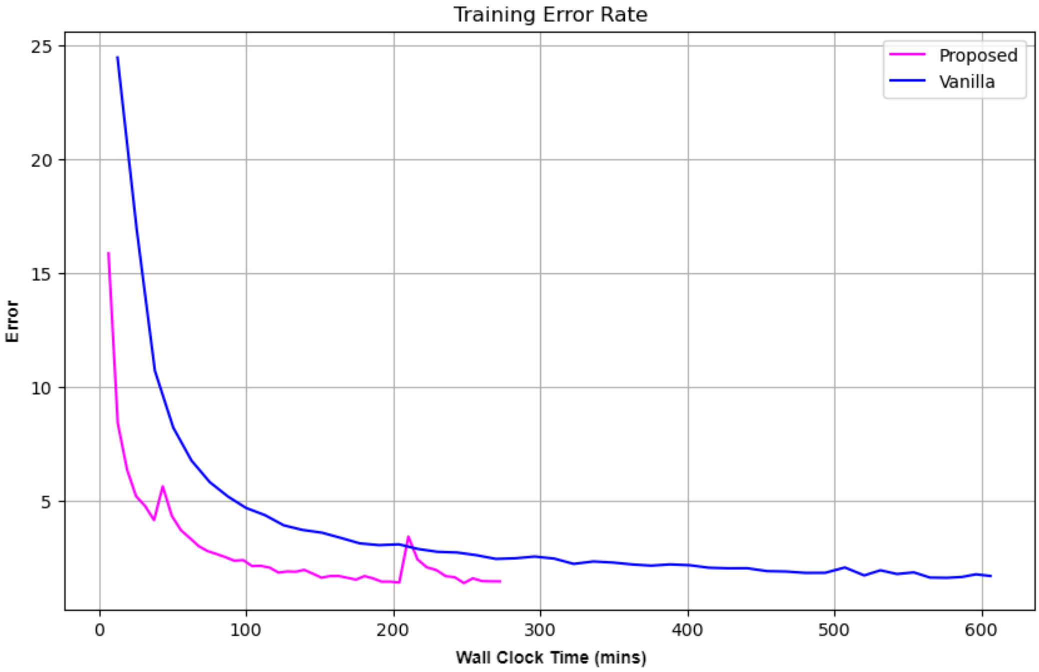


Figure 4.2: The training error over wall clock time comparison of the proposed and vanilla training methods. Time is measured in minutes. Note that the curve of the proposed model stops much earlier due to the shorter epoch durations at the beginning of the training process following the significantly lesser number of parameters.

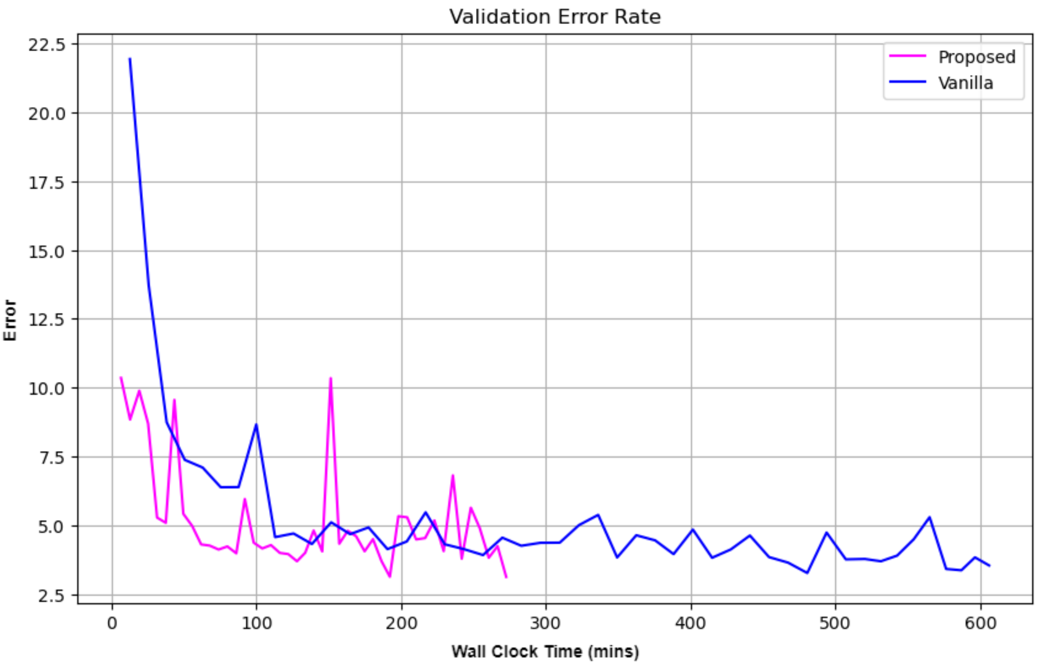


Figure 4.3: The validation error over wall clock time comparison of the proposed and vanilla training methods. Time is measured in minutes. Note that the curve of the proposed model stops much earlier due to the shorter epoch durations at the beginning of the training process following the significantly lesser number of parameters.

4.4 Discussion

4.4.1 Strengths of this Study

The study introduced an adaptive training methodology for ResNet architectures, marking a significant leap forward in the realm of ML methodologies. This method's unique ability to expedite the training process while maintaining a high level of accuracy is a testament to its robustness and effectiveness. This approach addresses some of the most pressing challenges in the field, particularly those related to the computational demands associated with training deep NNs. The proposed method's innovative design and implementation strategies offer a promising solution to these challenges.

4.4.2 Main Findings

The primary findings of this study underscore the success of the proposed method in achieving accuracy levels that are comparable, if not superior, to those achieved by traditional, fixed-depth training approaches. Furthermore, the method demonstrates significant improvements in training speed, highlighting its potential to substantially enhance the efficiency of ML projects. These findings suggest that the proposed method could be a valuable tool for researchers and practitioners in the field of ML.

4.4.3 Clinical Implications

While the study's focus is primarily on technical and computational aspects, its implications extend well beyond these areas into the clinical domain. By enabling faster and more accurate model training, the proposed method has the potential to expedite the development of ML models for clinical applications. This includes applications related to arrhythmia detection, among others. The resulting models could provide more accurate and timely diagnostic tools, ultimately leading to improvements in patient care and outcomes.

4.4.4 Resource Efficiency Considerations

One of the most notable aspects of the proposed method is its contribution to resource efficiency. The method leads to significant reductions in CO_2eq, energy consumption, and the number of parameters, aligning with sustainable ML practices and addressing critical environmental concerns. These benefits underscore the importance of developing more eco-friendly approaches to ML. This is particularly relevant in light of the growing awareness of the environmental impact of computing technologies and the urgent need for solutions that mitigate this impact.

4.4.5 Study Limitations

Despite promising results, it is essential to acknowledge this study's limitations. While the proposed method shows potential, further research is needed to explore its applicability across different network architectures and datasets. Additionally, the study's focus on ResNet models means that the findings may not directly translate to other types of NNs or ML tasks. Future work should aim to validate and extend these results, ensuring that the benefits of adaptive training can be realised across a broader spectrum of applications.

Chapter 5

Segmetron: Sample-efficient Model-agnostic Cardiac Semantic Segmentation with a Trustworthy Reject Option via PQ learning

5.1 Introduction

5.1.1 Background

Semantic segmentation is a central task in computer vision as it serves as the cornerstone of downstream analysis in autonomous driving medical decision-making (diagnosis and treatment) vision-enabled robots etc. Even though state-of-the-art (SOTA) deep learning (DL) semantic segmentation models shine out within the data training distribution they completely flop outside of it [152, 153]. The disparity between the distribution of the input samples used to train the model and the input distribution encountered during testing/deployment also known as covariate shift [154] is the rule (rather than the exception) in real-world scenarios and a cause of significant performance degradation. The covariate shift is exacerbated in the medical imaging field due to diverse imaging protocols patient population heterogeneity medical conditions prevalence of noise and artifacts among other factors [155]. Automated failure detection techniques for semantic segmentation whose results are rigorously guaranteed in the absence of labelled target data are of paramount importance in delivering DL transformative technologies. Nevertheless there is currently a lack of trust in pertinent techniques.

5.1.2 Related Work

A natural approach for detecting covariate shift in image classification has been to cast the problem as a two-sample statistical hypothesis test. The first sample comprises the training data, whereas the second sample is the latest deployment data. Then, the null hypothesis, $\mathcal{H}_0$, is that the samples were drawn from the same probability distribution, as opposed to the alternative hypothesis $\mathcal{H}_1$ that the two distributions are different.

A non-parametric deep kernel-based two-sample hypothesis test was proposed in [156]. The test statistic was based on the maximum mean discrepancy (MMD), which measures the differences between the two kernel mean embeddings. The kernels were parameterised by deep neural networks trained to optimise the test power, rendering this test particularly suited for high-dimensional (such as image) data. The authors of [157] designated H-divergence as a test statistic for two-sample tests. H-divergence is based on the generalised entropy defined by the maximum log-likelihood of readily available deep generative models. It allows to take advantage of inductive biases for each type of data, leading to improved test power. A baseline alternative for detecting covariate shift is to perform a non-parametric Kolmogorov-Smirnoff (KS) test directly on the distribution of relative Mahalanobis distance (RMD) confidence scores obtained for the two samples [158]. The RMD metric was initially introduced to improve out-of-distribution (OOD) detection [158]. Lastly, another way to conduct a two-sample test for recognising covariate shift is to use a classifier-based method. According to this approach, a binary classifier is trained to differentiate between source and target samples, and the test statistic could be based on the classifier's accuracy in a held-out test sample [159].

All the above studies looked into domain shift complications suffered by DL classification models. Covariate shift in the richer semantic segmentation is far less investigated [160]. Even though semantic segmentation and classification are related tasks, task-specific studies are valuable given that learning algorithms may behave inconsistently across different tasks. The findings obtained in classification studies might not be valid for semantic segmentation. As an illustration, the calibration methods, that have been proposed in the literature to deal with the overconfidence issue in DL classification models, behave differently in semantic segmentation [160]. In addition, semantic segmentation is a task of increased complexity when compared to classification, as the dense individual pixel predictions must ensure that they are spatially consistent and that they pick up adjacent pixel relationships (local context) [161]. Moreover, semantic segmentation models have to effectively deal with occlusions [161].

5.1.3 Our Contribution

In this study we make the following pivotal contributions:

- We develop a reliable, sample-efficient, distribution-free and model-agnostic hypothesis test, named the Segmetron (Figure 5.1), to detect image-level covariate shift in semantic segmentation. To assess an unlabelled target domain, Segmetron relies on an existing (but random) pre-trained semantic segmentation model and the labelled samples (pixels) used to train it. The test statistic of the one-sided hypothesis test is based on the rate of sample disagreement of two ensemble models trained to disagree with the baseline segmenter on unseen samples from the training and deployment sets, respectively.

- To obtain strong performance theoretical guarantees on unknown arbitrary test distributions, we build on recent work on the PQ learning setting of selective classification (SC) and extend it to a different discriminative model (i.e. segmenters).

- To train the enforced disagreement segmenters (EDSs) of each ensemble model to learn the same generalisation region as the pre-trained semantic segmentation model, we innovatively propose loss functions (to agree) which are more apropos to the semantic segmentation task and comply with the training of the baseline segmenter.

- We examine real-world covariate shifts that arise naturally (i.e. without human intervention), as opposed to previous studies on semantic segmentation robustness which relied on synthetic domain shifts, obtained by injecting noise/blur or crafting adversaries [162, 163, 164]. In particular, we analyse two covariate shifts from the cardiovascular magnetic resonance imaging (CMR) field, concerned with both binary (aorta, background) and multi-class (left ventricle, right ventricle, myocardium, background) semantic segmentation tasks, ensuring diversity of set-ups.

- We demonstrate that Segmetron outperforms other SOTA techniques in terms of statistical power on the two semantic segmentation tasks, given access to only one image.

5.2 Materials and Methods

5.2.1 Covariate Shift

Let X and Y be the input and output (label) spaces respectively defining the semantic segmentation task. Then, the input and output data are simply random variables. Labels are discrete (C classes). Assume that the marginal distributions of X and Y in the training (source) and testing (target) domains are denoted by $P_s(X), P_s(Y), P_t(X)$, and $P_t(Y)$ respectively. Similarly, the conditional distributions of the output variables given the input variables in the two domains are denoted by $P_s(Y|X)$ and $P_t(Y|X)$. Covariate shift refers to the situation where the marginal distribution of the input variables varies across the source and target domains ($P_s(X) \neq P_t(X)$) whereas the conditional distribution of the output variables given the inputs remains unaltered ($P_s(Y|X) = P_t(Y|X)$) [165]. To put it simply this type of shift means that the input data distribution is different between the source and target domains but the relationship between the input and output variables is the same.

5.2.2 The PQ Learning Setting of Selective Classification

The goal of SC (a.k.a. classification with a reject option) is to learn a classifier model which is allowed to abstain from making predictions when it's not adequately confident [166]. Unlike standard classifiers, which are forced to provide a prediction for every input, a selective classifier can choose to not classify specific examples if it deems the predictions unreliable. This allows SC models to achieve higher accuracy by reducing the number of misclassifications at the expense of coverage (i.e. the fraction of inputs on which predictions are made).

The PQ learning setting of SC has permitted to obtain strong theoretical guarantees on learning with arbitrary and potentially adversarial future test examples [166]. Their work represented a great leap forward, since the specific problem was considered intractable up until then. The authors showed that both the finite-sample error on the random and unknown test distribution Q and the rejection rate on the training distribution P can jointly remain bounded within an acceptable limit $\widetilde{\epsilon}$ with high probability 1-δ.

Formally, in the PQ learning setting of SC, we are given: (i) a training set of n samples $(x_1, x_2, \ldots, x_n)$, drawn i.i.d. from P over the input space X, (ii) the labels $(f(x_1), f(x_2), \ldots, f(x_n))$ for some unknown target function $f \in F$ of Vapnik–Chervonenkis (VC) dimension d, (iii) an unlabelled test set of n samples $(\dot{x}_1, \dot{x}_2, \ldots, \dot{x}_n)$ (i.i.d. from Q), and (iv) the bound parameter ϵ, which also controls the trade-off between errors and rejections. The output is a selective classifer $h|_S$, which is allowed to only predict on certain examples in a subset $S \subset X$, and otherwise abstain from predicting (S is also an output). The theoretical bound is given by the below definition [166].

Learner $\widetilde{L}$ $(\widetilde{\epsilon}, \delta, n)$-PQ-learns a function class $\mathcal{C}$ if, for any distributions P and Q over the input space X, and any target function $f \in \mathcal{C}$, the output $h|_S = \widetilde{L}(P, f(P), Q)$ satisfies:

$$\Pr_{x \sim P^n, x' \sim Q^n} [\mathtt{Reject}_P + \mathtt{Err}_Q \leq \widetilde{\epsilon}] \geq 1 - \delta,$$

where $\texttt{Reject}_P$ is the rejection rate of the classifier on the training distribution P, and $\texttt{Err}_Q$ is the error rate on the test distribution Q. Learner $\widetilde{L}$ PQ-learns C if it runs in non-exponential time and there is a polynomial p such that $\widetilde{L}(\widetilde{\epsilon}, \delta, n)$ PQ-learns C for every $\widetilde{\epsilon}, \delta > 0, n \geq p(1/\widetilde{\epsilon}, 1/\delta)$. A recent implementation of the PQ learning setting of SC is the Rejectron algorithm [166]. It is shown that Rejectron achieves an error bound of $\widetilde{\mathcal{O}}(\sqrt{d/n})$, for any class C of functions with a bounded VC dimension d, where the $\widetilde{\mathcal{O}}$ notation hides logarithmic factors including the dependence on the failure probability δ.

5.2.3 Problem Set-up: Semantic Segmentation with a Trustworthy Reject Option

Let $f_B : X \to Y$ be a semantic segmentation model from a function class F that maps from space X to a discrete set of classes $Y = \{1, \ldots, C\}$. Suppose f_B was trained on a dataset of labelled pixel samples (x_i, y_i) for $i = 1, \ldots, n$ where each x_i is drawn identically from a distribution P over X. During testing/deployment, $f_B$is asked to predict on new unlabelled samples from an arbitrary distribution Q over X.

We define as “semantic segmentation with a trustworthy reject option” the problem of building an automated segmenter that abstains from predicting on those samples of Q for which there is covariate shift and predicts otherwise. The goal is that the algorithm does so with strong guarantees. Even though samples are individual pixels, the decision whether to predict or not (or, equivalently, whether there is covariate shift or not) is taken at image-level.

5.2.4 Enforced Disagreement Segmenter

The EDS is a modification of the standard segmenter, designed to enhance the model’s sensitivity to shifts in data distribution. Unlike typical segmenters that aim for accurate predictions across all examples, an EDS is tailored to maximise disagreement on specific out-of-distribution data while maintaining consistent predictions on in-distribution image-level data. This approach leverages the inherent variability in the model’s response to different data distributions to detect shifts.

An EDS is characterised by the following properties:

- **Model Consistency:** It belongs to the same model class as the base segmenter and is trained using the same algorithm, ensuring that it does not deviate in fundamental learning capabilities.
- **In-Distribution Performance:** It achieves similar performance on unseen samples that follow the in-distribution, verifying that the model’s utility is retained for familiar data.
- **Maximal Disagreement:** On elements of a dataset $\mathbb{Q}$, representing a potential out-of-distribution set, the EDS is trained to disagree maximally with the predictions of the base segmenter, without compromising its performance on in-distribution data.

The operational mechanism of an EDS involves training the model to identify and emphasise discrepancies between the predicted and actual semantic segmentations on new unseen datasets. This is achieved by:

1. Training the base segmenter on a labelled dataset from distribution P.

2. Developing the EDS by further training on a subset from distribution Q, tweaking it to maximise prediction disagreement specifically on Q while ensuring it agrees with the base segmenter's predictions on P.
3. Applying early stopping in the EDS training, if validation performance drops by a certain amount to avoid catastrophic overfitting in small sample regimes.

The primary utility of an EDS lies in its ability to act as a diagnostic tool for flagging up shifts in data distribution that might affect the model's performance, providing an early warning system for potential degradation in model accuracy due to distribution changes. The development of the EDS represents a strategic shift in handling data distribution changes in semantic segmentation. By focusing on the disagreement in predictions between known and new data distributions, EDS offers a robust mechanism for enhancing the reliability of deployed semantic segmentation systems.

5.2.5 Learning to Agree and Disagree

Our goal is to agree on $\mathbb{P}$. To this end, and unlike [167], we propose two loss functions that are better suited for semantic segmentation tasks, namely the Focal Tversky loss and Dice loss.

The Focal Tversky loss [33] is an enhancement of the Tversky loss [55], aimed at addressing class imbalances in image segmentation tasks. It is defined as:

$$\text{Focal Tversky Loss} = (1 - \text{Tversky Loss})^{\gamma_{FT}} \tag{5.1}$$

where γ_{FT} is a focusing parameter that controls the contribution of hard-to-segment pixels, and the Tversky loss is given by:

$$\text{Tversky Loss} = \frac{\sum_i^N p_{0i} g_{0i}}{\sum_i^N p_{0i} g_{0i} + \alpha_{FT} \sum_i^N p_{0i} g_{1i} + \beta_{FT} \sum_i^N p_{1i} g_{0i}} \tag{5.2}$$

In this formula, p_{0i} is the predicted probability that pixel i belongs to the target class (e.g., aorta), and p_{1i} is the probability that pixel i is part of the background. Similarly, g_{0i} is 1 if pixel i belongs to the target class, and 0 otherwise, while g_{1i} represents the opposite. The parameters α_{FT} and β_{FT} balance the penalties for false positives and false negatives, respectively.

The Focal Tversky loss helps control class imbalances and focuses on difficult cases. In our experiments, we set $\alpha_{FT} = 0.8$, $\beta_{FT} = 0.8$, and $\gamma_{FT} = 1$ to improve model convergence and recall.

The Dice loss is a popular loss function for image segmentation tasks [168], particularly useful when dealing with imbalanced classes. It is based on the Dice coefficient, which measures the overlap between the predicted segmentation and the ground truth. The Dice loss is defined as:

$$\text{Dice Loss} = 1 - \frac{2\sum_i^N p_{0i} g_{0i}}{\sum_i^N p_{0i} + \sum_i^N g_{0i}} \tag{5.3}$$

Here, p_{0i} is the predicted probability that pixel i belongs to the target class, and g_{0i} is 1 if the pixel is correctly segmented as part of the target class, and 0 otherwise.

The Dice loss is particularly effective when there is a significant class imbalance, as it directly optimises for the overlap between predicted and actual segments. Maximising this overlap helps reduce false positives and false negatives, leading to improved segmentation performance.

Inspired by the work in [167], we relied on the disagreement cross entropy (DCE) to train segmenters that maximise disagreement on out-of-distribution data. This loss function extends the classical cross entropy loss by encouraging pixel predictions to diverge from the true class label on new, unseen data distributions, thereby effectively identifying distribution shifts.

The DCE loss for a segmenter predicting a distribution over C classes is defined as:

$$L_{DCE}(\widetilde{y}, f_B(x_i)) = \frac{1}{1-C} \sum_{c=1}^{C} \mathbb{1}_{f_B(x_i) \neq c} \log p(c|x_i), \tag{5.4}$$

where $\widetilde{y}$ is the prediction by the ensemble segmenter, $f_B(x_i)$ is the predicted pixel label by the baseline segmenter, and $p(c|x_i)$ is the probability that the ensemble segmenter predicts class c on pixel i. The indicator function $\mathbb{1}_{f_B(x_i) \neq c}$ equals 1 when $f_B(x_i)$ is not equal to c, pushing the ensemble segmenter to assign higher probabilities to incorrect classes.

The DCE has many attractive properties such as being very stable to optimise using gradiant descend methods, having a bounded global minimum and satisfying

$$\forall \mathtt{p} \in P \quad \min_{\mathtt{q} \in Q} L_{DCE}(\mathtt{q}; y) \leq L_{DCE}(\mathtt{p}; y) \tag{5.5}$$

meaning that for each probability vector in P, there is a corresponding probability vector in Q that attains a score that is at least as low.

Then, the overall training objective L_{EDS} can be obtained by combining one of the two losses proposed above that enforce agreement on P and the L_{DCE} for samples from the new distribution Q:

$$L_{EDS}(P, Q) = \sum_{(x_i, y_i) \in P} L_{trad}(\widetilde{y}, f_B(x_i)) + \lambda \sum_{x_i \in Q} L_{DCE}(\widetilde{y}, f_B(x_i)), \tag{5.6}$$

where L_{trad} is either the Focal Tversky or Dice loss and λ is a tuning parameter that balances fitting to P and learning to disagree on Q.

5.2.6 The Segmetron Hypothesis Test

To detect covariate shift in semantic segmentation, Segmetron expands on [167], which in turn had built on earlier work [166]. Segmetron conducts a statistical hypothesis test between the distributions of a potentially shifted new dataset $\mathbb{Q}$ and a dataset $\mathbb{P}^{\star}$ which was not seen during training but it is known to be from the same distribution as the training data. It adopts a transductive approach in the sense that it's constructed by: (i) training segmenters using L_{EDS} on observed training (and test) cases, (ii) performing reasoning to the specific test data.

Let f_B be a baseline segmenter trained on dataset $\mathbb{P}_{train}$ comprising input data (pixels) sampled from P and the corresponding masks. Assume f_Q is a segmenter which agrees (i.e. segments alike) with f_B on $\mathbb{P}_{train}$ and disagrees on a dataset $\mathbb{Q}$ sampled from an unknown arbitrary Q. We use ϕ_Q to denote the rate at which f_Q disagrees with f_B on n unseen pixels from Q, and ϕ_P to denote the rate at which f_Q disagrees with f_B on n unseen pixels from P. Then, by viewing semantic segmentation as a pixel-wise classification problem outputting a dense mask with a predicted class for every pixel, we argue that ϕ_Q being greater than ϕ_P, implies a covariate shift. The proof for the above is based on the fact that under the null hypothesis ($P = Q$), the upper bound of the probability that f_Q is more likely to disagree on Q than P is 0.5 [167]:

$$P = Q \quad \Rightarrow \quad \mathbb{P}(\phi_Q > \phi_P) \leq \frac{1}{2}\left(1 - 4^{-n}\binom{2n}{n}\right) < \frac{1}{2}. \tag{5.7}$$

Segmetron trains two EDS ensembles f_P and f_Q. f_P is trained to disagree on unseen $\mathbb{P}^\star$ from P, and f_Q is trained to disagree on $\mathbb{Q}$. After the training is completed, ϕ_P and ϕ_Q are calculated. If $\phi_Q > \phi_P$, then there is a covariate shift in the data (alternative hypothesis). If $\phi_P \geq \phi_Q$, then $\mathbb{Q}$ dataset is in P distribution (null hypothesis). Segmetron is distribution-free which means no assumptions about P and Q are made.

Two-sample hypothesis tests are associated with Type I errors (i.e. rejecting the true $\mathcal{H}_0$) and Type II errors (i.e. failing to reject a false $\mathcal{H}_0$). The upper bound of the probability of Type I error is controlled by choosing an appropriate significance level. The probability of not making a Type II error is called test power, and is regarded the main efficacy measure in null hypothesis testing. In this study, in order to test for shift on a set $\mathbb{Q}$, we follow the typical setup [169] by: (i) performing a permutation test to guarantee a significance level (or, else, a bounded Type I error), and (ii) empirically measuring the test power. To obtain the Segmetron result at a significance level α (=0.05), we train Segmetron for C_R (=100) calibration rounds with random $\mathbb{P}^\star$. Then, the Segmetron result is significant at the 5% level if ϕ_Q is greater than the (1-α) percentile of ϕ_P. The pseudocode for Segmetron algorithm is given below (**Algorithm 4**). The flowchart is illustrated in Figure 5.1.

5.2.7 Experiments

Datasets

We validated Segmetron in binary and multi-class semantic segmentation tasks, both from the CMR medical imaging field.

The former task involved segmenting aorta (both ascending and descending) from SSFP cine CMR images. The initial (unshifted) dataset is described in [170]. It consists of 340 (2D + time) training datasets and 84 testing datasets from the same distribution. All patients had aortic stiffness-related diseases. To assess covariate shift detection, we received from the same clinical institution additional 280 (unlabelled) patient datasets, for which the base segmenter failed due to covariate shift caused by irregular heart beats and/or breathing pattern, acquisition hardware variability, elevated flow velocity in the aortic valve, and different patient demographics. Each patient dataset comprised 30 time points (2D images).

The multi-class segmentation task involved four classes, namely left ventricle, right ventricle, myocardium, and background. The dataset from the Automated Cardiac Diagnosis Challenge (ACDC) [171] served as the initial dataset sampled from P. This dataset consists of 100 training 3D datasets (20 healthy patients, 20 patients with previous myocardial infarction, 20 patients with dilated cardiomyopathy, 20 patients with hypertrophic cardiomyopathy, 20 patients with abnormal right ventricle) and 50 testing 3D datasets. These were all 3D SSFP CMR data. The shifted dataset, acquired using different CMR sequences, comes from the Multi-sequence Cardiac MR Segmentation Challenge in 2019 (MS-CMRSeg 2019) [172]. We analysed 45 LGE or T2 CMR 3D datasets from patients who underwent cardiomyopathy.

Algorithm 4 The Segmetron algorithm

```
Input: $\mathbb{P}$: labelled dataset ,$(x_i, y_i)$, $\mathbb{Q}$: unlabelled dataset, $(x_i)$, $L_A$: learning algorithm, $C_R$: calibration rounds = 100, E: ensemble size = 5, $\alpha$: significance level = 0.05, $M_e$: evaluation metric (Dice accuracy), $\epsilon$: tolerance (=0.05), $e_m$: max epochs (=4).
Output: test result for covariate shift at significance level $\alpha$.
Partition $\mathbb{P}$ into $\mathbb{P}_{\text{train}}, \mathbb{P}_{\text{val}}, \mathbb{P}^\star$
$N_S \leftarrow |Q|$, $\phi_P \leftarrow [\quad]$
$f_B \leftarrow L_A(P_{\text{train}}, P_{\text{val}})$
for $C_R \leq 100$ do
    $\mathbb{P}^\star \leftarrow$ RandomSampling$(\mathbb{P}^\star, N_S)$
    while $n > 0$ and epochs $\leq N_S$ do // Train ensembles of EDSs on $\mathbb{P}^\star$
        $\hat{\mathbb{P}}^* \leftarrow \{(x, f_B(x)) \mid x \in \mathbb{P}^*\}$ // Infer pseudo labels on $\mathbb{P}^*$ using $f_B$
      //Dataloader using $\mathbb{P}_{train}$ and $\mathbb{P}^*$
        $\mathbb{P}, \mathbb{P}^* \leftarrow$ Batched$(\{(x, y) \mid (x, y) \in \mathbb{P}_{train} \wedge (x, f_B(x) \in \mathbb{P}^*\})$
        Initialise $f_P \leftarrow f_B$
        $m_0 \leftarrow M_e(f_B, \mathbb{P}_{\text{val}})$ // Compute the validation performance of $f_B$
        while $M_e(f_B, \mathbb{P}_{\text{val}}) > m_0 - \epsilon$ and iterations $< e_m$ do
            for batch in $\mathbb{P}, \mathbb{P}^*$ do
                $x_P, y_P \leftarrow \{(x, y) \mid (x, y) \in$ batch and $(x, y) \in \mathbb{P}_{\text{train}}\}$
                $x_{P^*}, y_{P^*} \leftarrow \{(x, f_B(x)) \mid x \in$ batch $\wedge\, x \in \hat{\mathbb{P}}^*\}$
                Update $f_P$ using $L_A$ for $(x_P, y_P)$ and disagreement update for $(x_{P^*}, y_{P^*})$
            end for
        end while
        return $f_P$
        Filter out agreed pixels $\mathbb{P}^\star \leftarrow \{x \mid x \in \mathbb{P}^\star$ and $f_B(x) = f_P(x)\}$
        Update disagreement rate: $\phi_P \leftarrow 1 - \frac{|\mathbb{P}^\star|}{N_S}$
    end while
    Append $\phi_P$ to $[\phi_P]$ list
end for
while $n > 0$ and epochs $\leq N_S$ do
    $\hat{\mathbb{Q}} \leftarrow \{(x, f_B(x)) \mid x \in \mathbb{Q}\}$ // Infer pseudo labels on $\mathbb{Q}$ using $f_B$
    $\mathbb{P}, \mathbb{Q} \leftarrow$ Batched$(\{(x, y) \mid (x, y) \in \mathbb{P}_{train} \wedge (x, y) \in \mathbb{Q}\})$ //Dataloader using $\mathbb{P}_{train}$ and $\mathbb{Q}$
    Initialise $f_Q \leftarrow f_B$
    $m_0 \leftarrow M_e(f_B, \mathbb{P}_{\text{val}})$ // Compute the validation performance of $f_B$
    while $M_e(f_B, \mathbb{P}_{\text{val}}) > m_0 - \epsilon$ and iterations $< e_m$ do
        for batch in $\mathbb{P}, \mathbb{Q}$ do
            $x_P, y_P \leftarrow \{(x, y) \mid (x, y) \in$ batch and $(x, y) \in \mathbb{P}_{\text{train}}\}$
            $x_Q, y_Q \leftarrow \{(x, f_B(x)) \mid x \in$ batch $\wedge\, x \in Q\}$
            Update $f_Q$ with $L_A$ for $(x_P, y_P)$ and disagreement update for $(x_Q, y_Q)$
        end for
    end while
    return $f_Q$
    Filter out agreed pixels $\mathbb{Q} \leftarrow \{x \mid x \in \mathbb{Q}$ and $f_B(x) = f_Q(x)\}$
    Update disagreement rate: $\phi_Q \leftarrow 1 - \frac{|Q|}{N_S}$
end while
return $\phi_Q > (1 - \alpha)$ quantile of $\phi_P$
```

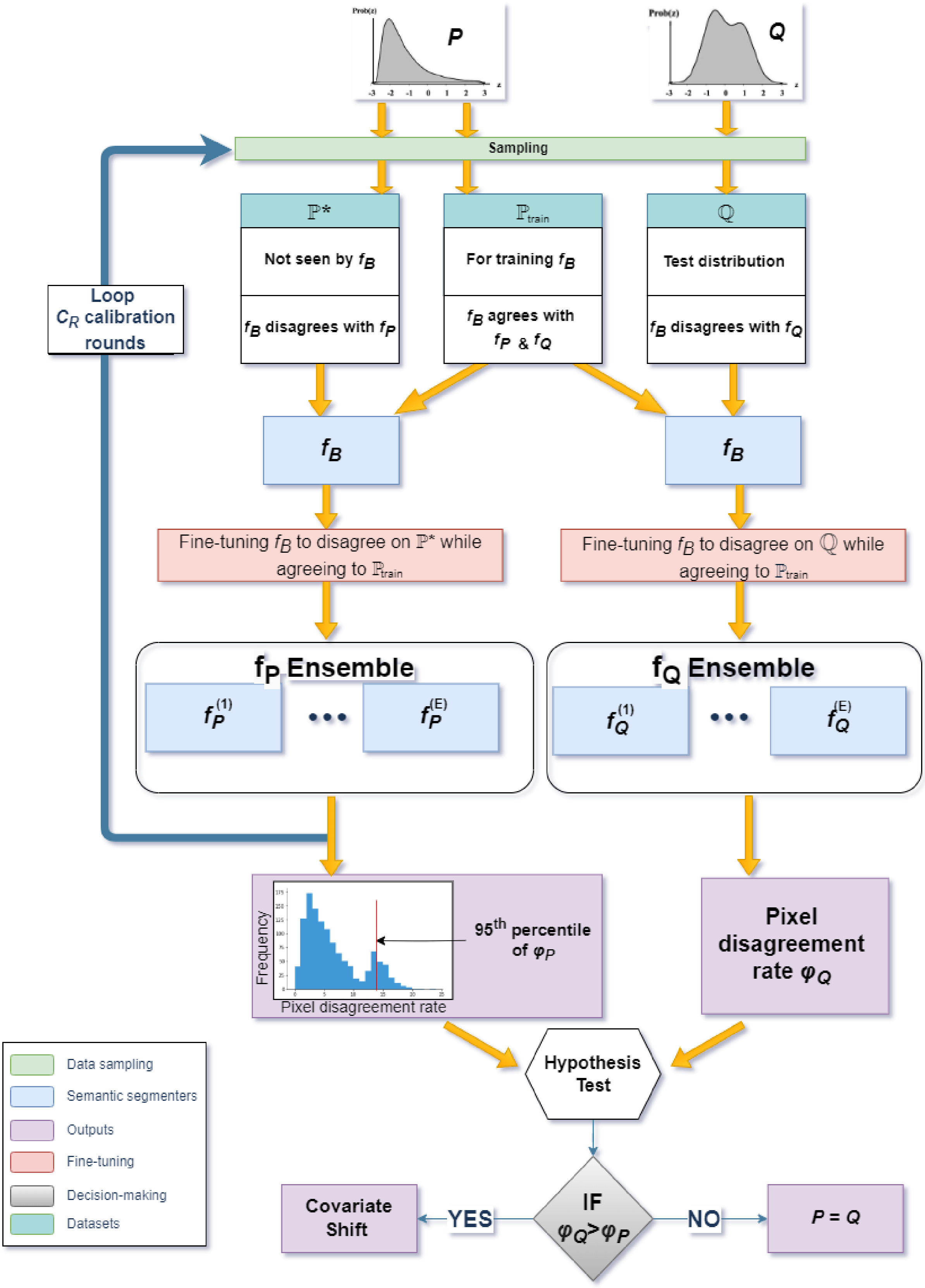


Figure 5.1: The Segmetron hypothesis test.

Base Segmenters

The binary base cardiovascular semantic segmentation model is described in [170]. It is a UNet-based model equipped with 2D ConVLSTM layers that utilise hidden state weights, dropout layers and batch normalisation layers juxtaposed with traditional UNet.

The multi-class base segmentation model we utilised was the overall 3^{rd} place winner in the ACDC challenge in 2019 [173]. This is also a UNet-based model, which is equipped with batch normalisation layers and padding layers that preserve the region of interest (ROI).

State-Of-The-Art Approaches

We juxtaposed Segmetron with tree SOTA techniques for covariate shift detection.

- The Relative Mahalanobis Distance (RMD) is a metric that quantifies the similarity between a sample and the distribution of a known dataset, adjusted by the covariance of the distribution [174]. It has been particularly helpful in detecting OOD samples.
- Deep Kernel Maximum Mean Discrepancy (MMD-D) is a non-parametric distance metric that measures the difference between two probability distributions. MMD-D is widely used in domain adaptation tasks to minimise the distribution shift between the source and target domains [175, 176].
- H-Divergence is a metric that captures the disagreement extent when a discriminative model is applied to different domains [177]. It is typically used to measure the distributional difference between source and target domains.

5.2.8 Implementation

For detecting covariate shift in binary semantic segmentation, $\mathbb{P}_{train}$, $\mathbb{P}_{val}$, $\mathbb{P}^{*}$, and $\mathbb{Q}$ datasets were randomly selected from the respective aorta datasets. All the above datasets involved a single (2D+time) patient dataset, comprising 1966080 ($= 30 \times 256 \times 256$) samples. We trained five EDSs for each ensemble. We employed the Focal Tversky loss to agree on P. Both f_P and f_Q were initialised using the weights of f_B. We trained each EDS for 5 epochs, with early stopping if validation performance dropped by 5%. The datasets from P were permuted across 100 random calibration runs to generate 100 f_P ensemble models, from which we obtained 100 pixel disagreement rate values. The permutation test enabled us to deliver strong statistical guarantees. We then calculated ϕ_P as the 95^{th} percentile of these rates. One test run was used to produce f_Q which gave the pixel disagreement rate value ϕ_Q. To enhance the presentation of the results, we purposefully selected the number of disagreement pixels rather than the rate values. We utilised the exact learning rates and batch sizes recommended by authors in the original study [170]. We chose for the loss function tuning parameter λ to be $\frac{1}{|\mathbb{Q}|+1}$, as recommended by [166]. For obtaining TPR at a 5% Significance Level (TPR@5), we run the above experiment 100 times, each with randomly initiated data. To validate the usefulness of the chosen loss function, we also plotted the pixel disagreement and in-distribution accuracy, both as functions of the ensemble size.

For the multi-class cardiac semantic segmentation task, all the details are the same as above, except that each dataset involved 3D spatial image data, comprising 719,104 ($= 16 \times 212 \times 212$) pixels. In addition, the Dice loss function was chosen to agree on P. In the experiments, we utilised the exact learning rates and batch sizes recommended by the authors of the baseline study [173].

We used the penultimate layer of the pre-trained base models to test for covariate shift using the RMD approach [158]. Next, we performed the KS test directly on the distribution of

RMD confidence scores derived from $\mathbb{P}^*$ and $\mathbb{Q}$ for both binary and multi-class datasets. The `scipy.stats.ks_2samp` implementation of the KS test was employed, giving us directly the p-values.

To conduct the MMD-D test for both binary and multi-class datasets, we relied on the original source code provided by the authors at https://github.com/fengliu90/DK-for-TST.

The H-Divergence implementation involved training variational autoencoder models on both $\mathbb{P}^*$ and $\mathbb{Q}$ datasets individually, and on their uniform mixture $\frac{(\mathbb{P}^*+\mathbb{Q})}{2}$. The variational autoencoder loss, comprising reconstruction loss and a Kullback–Leibler divergence term, was used to compute a test statistic that quantifies the difference between the distributions by comparing the entropy of the mixture to the individual datasets. To realize the H-Divergence, the following general class of continuous functions was chosen

$$\widetilde{\phi}(\theta, \lambda) = \frac{(\theta_s + \lambda_s)^{\frac{1}{s}}}{2} \tag{5.8}$$

for $s > 1$, which generalizes the H-Jensen Shannon divergence for s=1 and the H-Min divergence for $s = \infty$. The loss function $l(x, a)$ was chosen as the negative log-likelihood of x under a distribution a, where a belongs to a model family A. The implementation was validated through a permutation testing scheme to compute the test power, which conducted 100 permutations for each experiment while maintaining an overall significance level at α=0.05. The source code at `https://github.com/a7b23/H-Divergence` [157] was employed.

All the experiments in this study were conducted on an Intel(R) Core(TM) i9-10900K CPU and RTX A600 48 GB GPU. Both Segmetron and the SOTA methods were compiled in a Python 3.8.5 environment. The proposed method was trained using a TensorFlow 2.4.0 DL framework.

5.2.9 Statistical Test and Evaluation

Kolmogorov–Smirnov (KS) Test

The Kolmogorov–Smirnov (KS) test, originally introduced by [178] and further developed by [179], is a non-parametric test that compares two probability distributions by measuring the largest difference between their cumulative distribution functions (CDFs). It is widely used to assess whether a sample comes from a reference distribution (one-sample KS test) or whether two samples are drawn from the same distribution (two-sample KS test).

Formally, for a given empirical distribution function of a sample of size n and a reference distribution, the KS statistic D_n is defined as:

$$D_n = \sup_x |F_n(x) - \acute{F}(x)|$$

where $F_n(x)$ is the empirical CDF of the sample and $\acute{F}(x)$ is the CDF of the reference distribution. In the two-sample KS test, the KS statistic is given by:

$$D_{n,m} = \sup_x |F_n(x) - G_m(x)|$$

where $F_n(x)$ and $G_m(x)$ are the empirical CDFs of the two samples of sizes n and m, respectively. The null hypothesis H_0 in both cases is that the sample(s) come from the same distribution.

The KS test has several important properties. It is sensitive to differences in both the location and shape of the empirical distribution functions. In addition, it does not require the assumption of normality or any particular distribution shape. Moreover, it is applicable to continuous distributions. The KS test is frequently used in various fields such as for comparing empirical distributions in hypothesis testing, for evaluating goodness-of-fit for models, and for detecting covariate shifts in ML models by comparing training and test distributions. While the KS test is versatile, it also has limitations. It is less powerful for detecting deviations at the tails of distributions. Also, the test is sensitive to sample size, as small differences may become statistically significant with large sample sizes.

Permutation Test

A permutation test is a non-parametric statistical method used to determine whether two datasets are significantly different from each other. It involves repeatedly shuffling the data and calculating a test statistic for each permutation, thereby generating a distribution of the test statistic under the null hypothesis. The p-value is then calculated as the proportion of the permuted statistics that are at least as extreme as the observed test statistic. This approach does not make assumptions about the underlying distribution, making it a powerful tool for hypothesis testing.

True Positive Rate

TPR, also known as *Sensitivity* or *Recall*, measures the proportion of actual positives that are correctly identified by the model. It is calculated as follows:

$$TPR = \frac{TP}{TP + FN}$$

where TP is the number of true positives and FN is the number of false negatives. The TPR indicates how well the model is able to identify positive instances.

We report the TPR at a 5% Significance Level (TPR@5) averaged over 100 randomly chosen sets Q. This indicates the frequency with which our method correctly detects covariate shift ($P \neq Q$), while maintaining a false positive rate of only 5%. This is equivalent to the statistical power of a test with a significance level (α) of 5%.

5.3 Results

Figures 5.2 and 5.3 show histograms of 100 calibration and test rounds, respectively, for the binary cardiovascular semantic segmentation task. It can be seen that in all 100 test cases, the number of disagreemnt pixels on $\mathbb{Q}$ is greater that the 95th percentile of the calibration rounds. Therefore, the covariate shift is always detected.

Table 5.1 presents the TPR@5 comparison between Segmetron and the SOTA methods for the binary and multi-class cardiac semantic segmentation tasks. The proposed method achieved the highest statistical power in both tasks among all evaluated approaches.

Figure 5.4 illustrates the Dice (validation) accuracy as the ensemble size increases for the binary semantic segmentation task. Plotted are graphs for both shifted and unshifted datasets from Q. It can be seen that enforcing disagreement does not compromise in-distribution performance.

Figures 5.5 and 5.6 depict the relationship between the number of disagreed pixels and the ensemble size for shifted and unshifted data, respectively. It can be observed that the number of disagreement pixels increases substantially faster (from 32486 to 93942) in the f_Q curve as the number of EDSs rises than the respective increase (from 6870 to 8292) in the f_P curve. This finding corroborates our loss function choices for training the ensemble models.

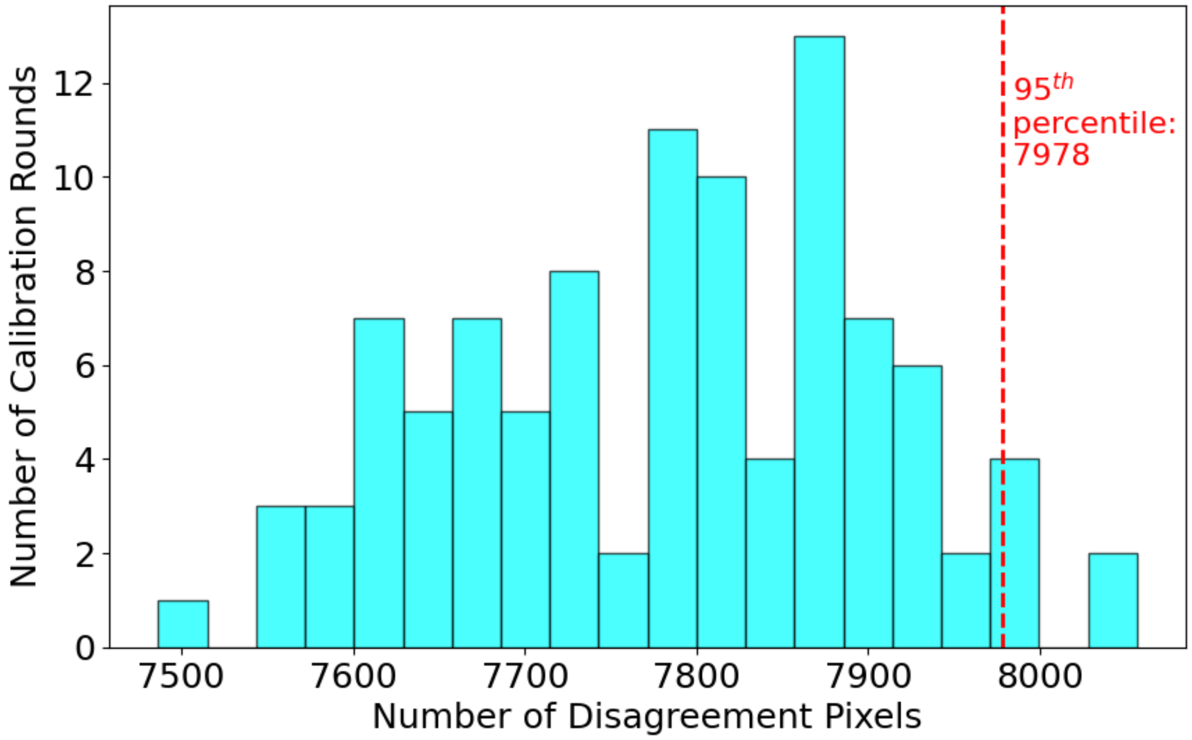


Figure 5.2: The Segmetron hypothesis test: Shown is the relationship between the number of calibration rounds and the number of disagreement pixels. A different random seed for $\mathbb{P}^*$ was used for each calibration round. Also shown in red is the 95^{th} percentile. The plot is taken from the binary cardiovascular semantic segmentation task.

5.4 Discussion

Semantic segmentation enables a higher level of understanding of a depicted scene. It also forms an essential capability towards delivering a plethora of life-changing technologies. However, the covariate shift itself, as well as the lack of trustworthy methods for detecting it, limit the applicability of semantic segmentation. This study introduced Segmetron, which is a segmenter with a reliable reject option that abstains from predicting on test images when semantic segmentation should not be made due to covariate shift. Importantly, Segmetron is able to deal with unforeseeable covariate shifts of any unknown arbitrary distribution that may occur during deployment. Therefore, this work is valuable because it aligns with "Responsible AI" principles and it happens at a time when the machine learning community is striving to increase society's willingness to accept AI. To obtain strong theoretical guarantees, we leveraged recent theoretical work on the PQ learning setting of selective classification [166, 167]. To detect covariate shift, we built two ensembles of segmenters that were enforced to agree on training data and disagree on test data. For training these ensemble models, we proposed novel ways (i.e. loss function components) of agreeing that are better suited to semantic segmentation. Inspired by recent work, we chose the pixel disagreement rate as the discriminative statistic upon which the Segmetron hypothesis test was built.

The ability of Segmetron to detect covariate shift was showcased in two real-world semantic segmentation tasks from the CMR field, involving two (aorta and background) or more (left ventricle, right ventricle, myocardium, background) semantic classes. Our approach was found to have superior statistical power to three SOTA approaches (RMD, H-Divergence, Deep Kernel

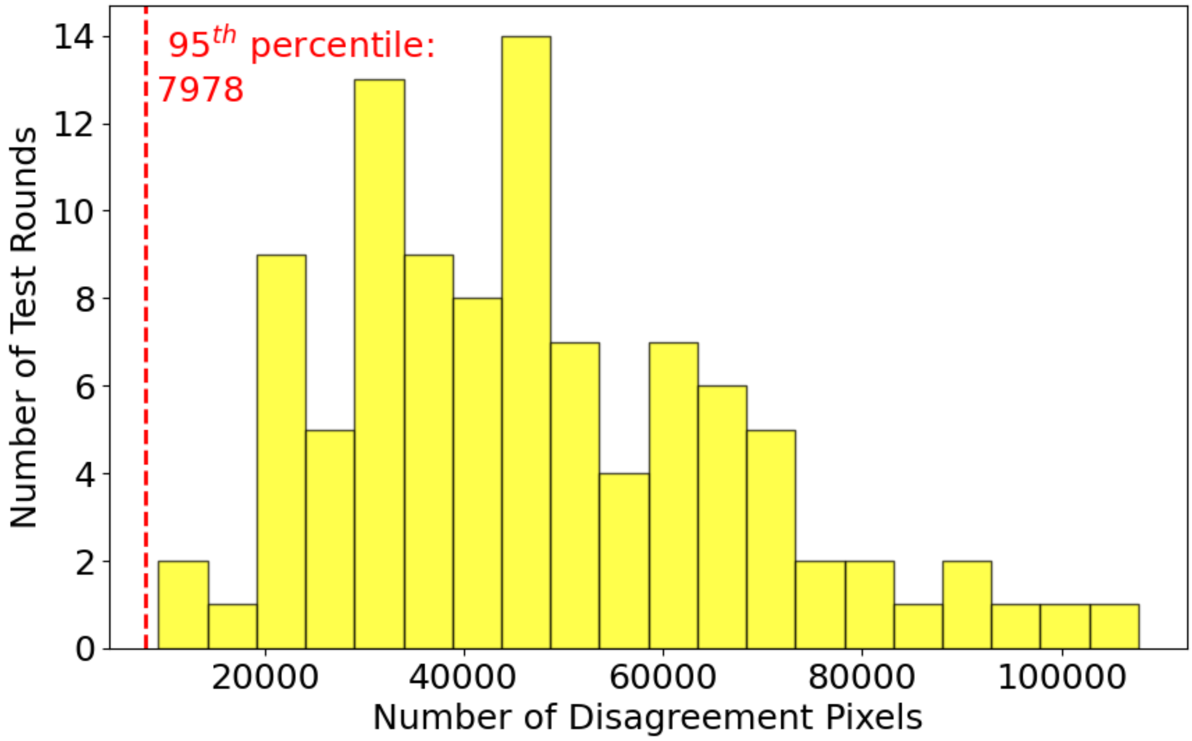


Figure 5.3: The Segmetron hypothesis test: Shown is the relationship between the number of test rounds and the number of disagreement pixels. A different random seed for $\mathbb{Q}$ was used for each test round. Also shown in red is the 95^{th} percentile of the 100 calibration rounds. The plot is taken from the binary cardiovascular semantic segmentation task.

MMD) on both datasets in terms of the TPR at 5

The validity of our design choices was demonstrated by plotting graphs of the relationship between the ensemble size versus disagreement rate and in-distribution accuracy. In particular, it was found that on unseen test sets, the disagreement level ϕQ manifested a much more rapidly increasing trend when compared with the baseline disagreement rate ϕQ. In addition, the EDSs of both ensembles were found to preserve high accuracy on data from the training distribution. Markedly, Segmetron is sample-efficient being able to identify covariate shifts from a single 3D (3D spatial or 2D spatial + time) dataset. Therefore, it could be useful for isolating shifted cases and removing them from automated semantic segmentation pipelines. Last, another favourable characteristic of Segmentron is that it is model-agnostic, since it recognises covariate shifts regardless of the baseline segmentation model that had been used in the initial training.

Segmetron also offers a straight-forward way to assess the intensity of a covariate shift. The severity ordering can be based on the relative disagreement gap between the in-distribution and out-of-distribution samples. The more harsh the domain drift, the larger the gap between the ϕQ and the 95th percentile of ϕP will be. Following this line of argument in our experiments, it was found that the covariate shift in the multi-class cardiac semantic segmentation problem was more severe than in the binary one.

Segmetron has higher computational complexity than the SOTA approaches [Ref9], and is similar in complexity to other ensemble approaches. However, Segmetron is more time-efficient compared to the SOTA methods because it is mainly reliant on GPU. Also, in deployment, deep kernel MMD and H-Divergence may require to train multiple deep ensemble models, whereas Segmetron promotes the utilization of pre-trained models. To run (inference) Segmetron on a

Table 5.1: True positive rates at a 5% significance level for detecting natural covariate shifts for binary and multi-class semantic segmentation tasks from the CMR medical imaging field. Boldface indicates best performance.

Method	TPR@5	
	Binary	Multi-class
Segmetron	**1.0**	**1.0**
RMD	0.95	0.95
MMD-D	0.24	0.88
H-Divergence	0.46	0.49

CMR: Cardiovascular Magnetic Resonance Imaging, TPR@5: True Positive Rate at 5 % significance.

single 3D test dataset, it took approximately 20 minutes per patient. In contrast, the RMD and the Deep Kernel MMD approaches required about one hour. H-Divergence was the most time-consuming, requiring approximately three hours. Notably, all SOTA methods (except Segmetron) run on the CPU and consume a significant amount of RAM. Segmetron, however, leverages GPU resources, which significantly reduces the processing time while also requiring minimal CPU and RAM usage.

5.4.1 Study Limitations

The PQ learning framework assumes that the samples are independent identically distributed, but such an assumption does not hold for semantic segmentation. However, this is not necessarily a problem in the strict sense. Instead, it's a characteristic of the task that is handled by designing modern DL semantic segmentation architectures that capture spatial correlations and multi-scale representations, and by incorporating appropriate training strategies and loss functions. In our framework, we aim to detect covariate shift from as few test samples as possible. Therefore, we estimated disagreement on the same patient dataset that had been used to train the ensemble f_Q. While this could result in high variance and low statistical power, we dealt with this issue by estimating the relative increase in disagreement between the EDSs on Q and P. Last, Segmetron detects covariate shift but it does not correct it. Test-time unsupervised domain adaptation methods could come to our rescue in this aspect, and this will be a topic of future work.

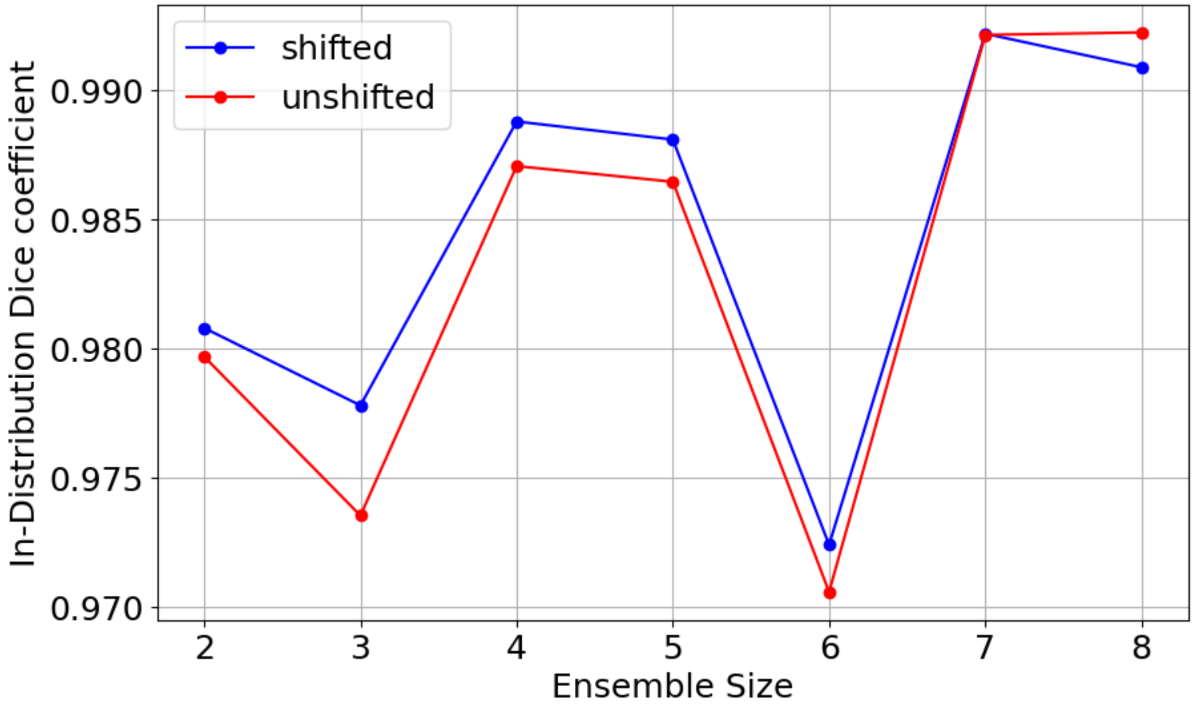


Figure 5.4: Plot of the in-distribution Dice coefficient as a function of the ensemble size. Plotted are graphs for both shifted and unshifted datasets from Q for the binary cardiovascular semantic segmentation task.

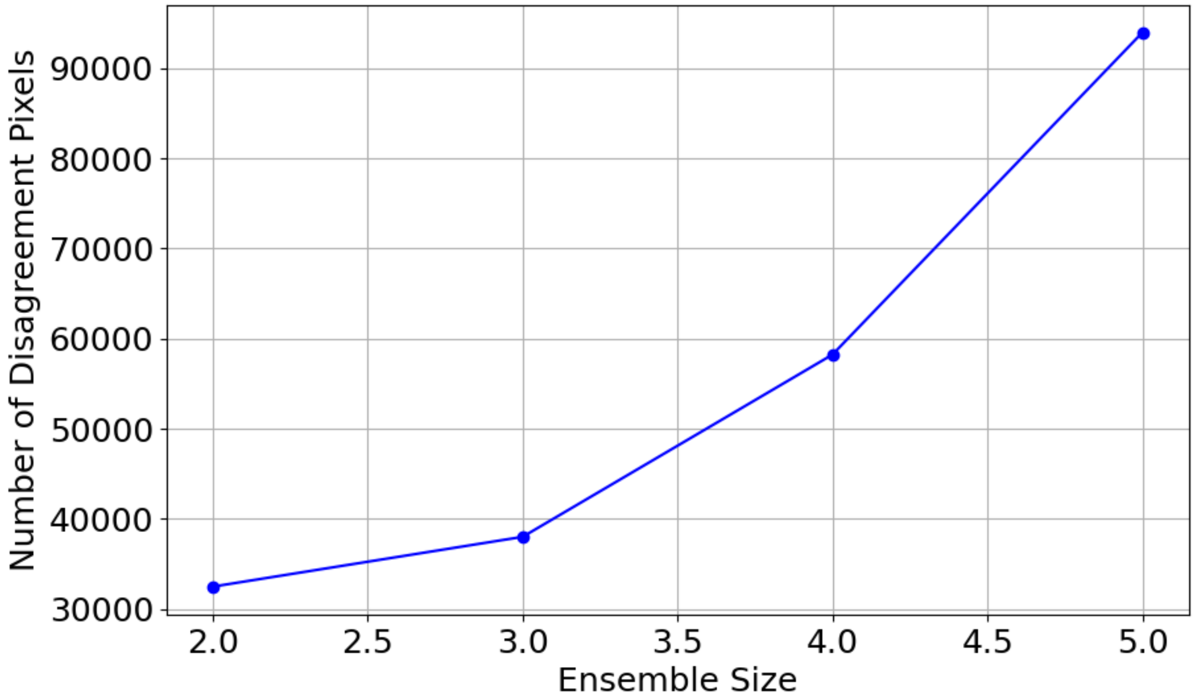


Figure 5.5: Plot of the number of disagreement pixels as a function of the ensemble size. Plotted is the graph for shifted data from Q for the binary cardiovascular semantic segmentation task.

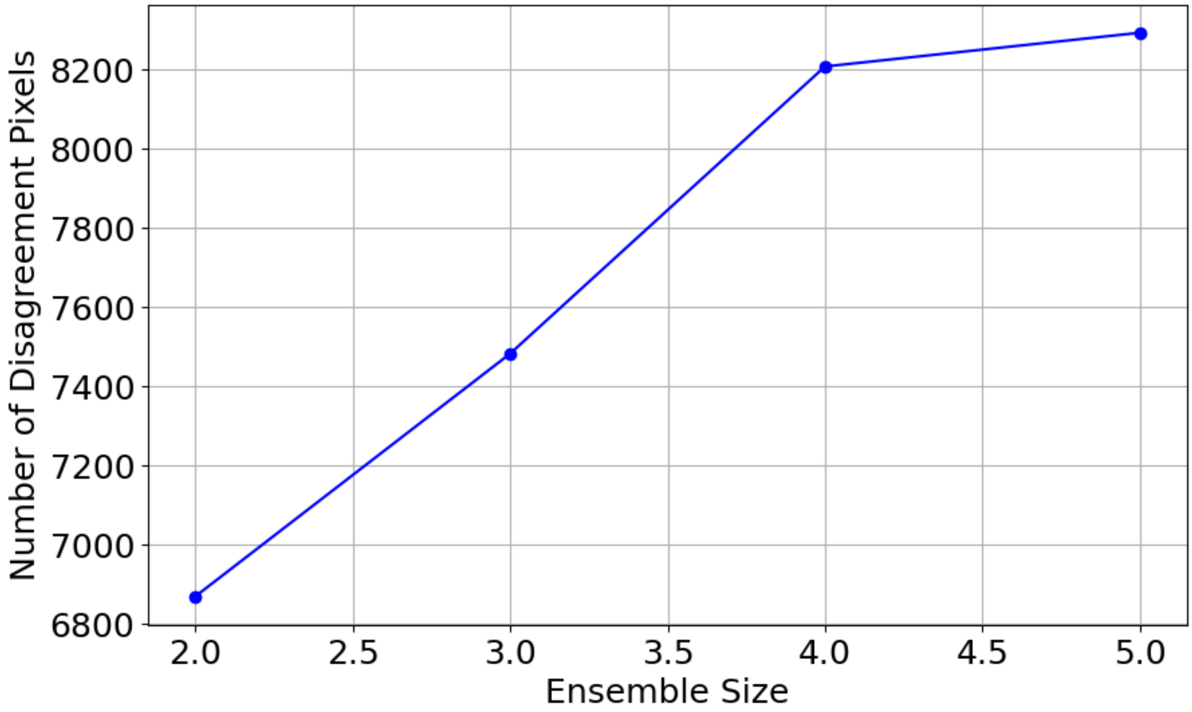


Figure 5.6: Plot of the number of disagreement pixels as a function of the ensemble size. Plotted is the graph for unshifted data from Q for the binary cardiovascular semantic segmentation task.

Chapter 6

Conclusions

6.1 Resource-efficient aortic distensibility calculation by end-to-end spatio-temporal learning of aortic lumen from multi-centre multi-vendor multi-disease CMR images

This study has embarked upon a novel journey in the field of computational cardiology, introducing a ground-breaking, resource-efficient deep learning (DL) model meticulously crafted for the swift, reliable, and fully autonomous segmentation of the ascending aorta (AAo) and descending aorta (DAo) from cine cardiovascular magnetic resonance (CMR) images. This pioneering approach facilitates the quantification of aortic distensibility (AD), a critical biomarker in cardiovascular research, with unprecedented precision and speed.

Our method, tested on an extensive, multi-centre, multi-vendor dataset encompassing a remarkably diverse patient cohort, demonstrated superior accuracy when juxtaposed with the current state-of-the-art (SOTA) methods. A striking feature of our proposed model is its environmental and energy consciousness. It utilised approximately 3.9 times less energy and generated around 2.8 times fewer carbon emissions than its SOTA counterparts. Furthermore, the accuracy of our model eclipsed even the unpruned methodologies, underscoring its exceptional efficacy.

The implications of this study extend far beyond mere technological triumph. By presenting a method that conscientiously addresses the dual aspects of computational efficiency and environmental impact, this work aligns with the burgeoning global narrative emphasising sustainable technological advancement. The energy-conscious nature of our model does not merely represent a technical feature; it embodies a commitment to a future where computational research and environmental stewardship coexist in harmony.

Moreover, the application potential of this model in large-scale biomedical databases, such as the UK Biobank, is particularly noteworthy. By enabling the rapid extraction and analysis of CMR-derived aortic phenotypes, our model opens new vistas in genome-wide association studies. It allows for a deeper exploration of the intricate relationships between AD, aortic dimensions, and cognitive functions, potentially unravelling novel insights into cardiovascular and neurocognitive health.

In an era where the carbon footprint of DL research is increasingly scrutinised, our study sets a new benchmark. It demonstrates that high computational efficiency and environmental responsibility can be achieved without compromising the accuracy and reliability of medical image analysis. This approach, we posit, should become a cornerstone in future DL research, especially in applications with substantial societal and clinical impacts.

In conclusion, our study does not merely present a novel DL model; it proposes a paradigm shift. It advocates for a future where computational efficiency, environmental responsibility, and clinical utility are not competing interests but are harmoniously integrated. The implications of this research are profound, paving the way for more sustainable, efficient, and inclusive advancements in medical imaging and beyond. As we navigate the challenges of an increasingly data-driven world, it is incumbent upon us to ensure that our technological pursuits are aligned with the principles of sustainability and ethical responsibility.

6.2 Enhanced Right Ventricular Volume Prediction and Uncertainty Estimation: From Supervised Tree Kernel Ensembles to Feature Tokeniser Transformer-based Regression on 2D Echocardiography Planimetry Data

In conclusion, Chapter 3 of this thesis has provided a comprehensive discussion of the various approaches and methodologies employed in the field of cardiovascular imaging, specifically in the context of right ventricular (RV) assessment using 2DE echocardiography. Two distinct approaches, namely Gradient Boosting Regression Trees (GBRTs) with instance-based uncertainty (**Method I**), and the Feature Tokeniser Transformer (**Method II**), have been critically evaluated, highlighting their strengths, limitations, and potential contributions to the field.

Method II showcased promising results in predicting RV volumes and ejection fractions, potentially revolutionising clinical practice by enhancing the accuracy and efficiency of RV assessment. However, it is important to acknowledge the common challenges, such as the limited dataset size, the need for further validation on larger and more diverse patient cohorts, and the time-consuming nature of manual tracings.

This chapter has emphasised the significance of these innovative methods in addressing the clinical need for trustworthy and automated RV evaluation, ultimately contributing to improved patient care and timely diagnosis of cardiac diseases. By providing uncertainty estimates, reducing the reliance on complex and time-consuming imaging modalities, and offering more accessible and efficient tools for clinicians, these approaches align with the principles of trustworthy artificial intelligence in healthcare.

In light of the discussed advancements and limitations, future research in this domain should focus on expanding datasets, deriving model interpretability also for attention-tabular models, combining the methods with automated tracing (end-to-end segmentation) pipelines and exploring opportunities for automation to reduce manual efforts. These efforts will pave the way for the widespread adoption of these methods in clinical settings, ultimately aiding both patients and healthcare professionals in the field of cardiovascular medicine.

6.3 Fast-tracking the Deep Residual Network Training for Arrhythmia Classification by Leveraging the Power of Dynamical Systems

This study proposed a method to reduce the training time of deep residual networks for the challenging arrhythmia classification task without compromising the model performance. We exploited the dynamical systems perspective of deep residual networks to achieve our goal.

The Lipschitz constant is calculated at every epoch to decide the growth of the network. The Lipschitz constant has recently received considerable attention in the DL community, mainly to improve stability and robustness against adversarial attacks [180, 181].

Extensive experiments on the MIT-BIH arrhythmia dataset demonstrated that the proposed method required at least 40% fewer parameters per epoch than conventional vanilla training while retaining or improving performance. It also reduced carbon emissions by at least 1/3 and improved energy efficiency by at least 1/4.

The proposed research not only underscores the reductions in training time afforded by our methodology but also foregrounds the associated savings on energy consumption and environmental impact. This investigation contributes a significant step forward in ongoing efforts to improve automated ECG signal analysis by fostering more resource-efficient ML development frameworks for arrhythmia detection. It heralds a new chapter in pursuing sophisticated, efficient, and sustainable solutions in the cardiovascular healthcare technology landscape.

For future work, we will explore further algorithmic refinements and try to extend this framework to encapsulate a broader spectrum of DL architectures. We will also apply the technique to diverse cardiovascular tasks, thus broadening its impact.

6.4 Segmetron: Sample-efficient Model-agnostic Cardiac Semantic Segmentation with a Trustworthy Reject Option via PQ learning

In conclusion, this study introduced Segmetron, which is a sophisticated segmenter with a trustworthy reject option that intelligently abstains from making predictions when it detects conditions unsuitable for semantic segmentation due to covariate shift. Segmetron is designed to handle unforeseeable covariate shifts of any unknown arbitrary distribution that may occur during deployment. To establish strong theoretical guarantees, we leveraged recent theoretical work on the PQ learning setting of selective classification. To detect covariate shift, we constructed two ensembles of segmenters that were enforced to agree on training data and disagree on test data. For training these ensemble models, we proposed novel ways (i.e. loss function components) of agreeing that are specifically tailored for the requirements of semantic segmentation. Inspired by recent work, we chose the pixel disagreement rate as the core discriminative statistic upon which the Segmetron hypothesis test was built. The ability and versatility of Segmetron to detect covariate shift was showcased in two challenging real-world semantic segmentation tasks from the CMR field, involving two (aorta and background) or more (left ventricle, right ventricle, myocardium, background) semantic classes. Our approach was found to surpass three SOTA approaches (RMD, H-Divergence, Deep Kernel MMD) on both tasks in terms of the TPR at 5% significance level, aggregated over 100 randomly selected test sets. Segmetron's strengths lie in its sample efficiency—effectively detecting covariate shifts using a single 3D dataset—and its model-agnostic nature, as it can detect shifts regardless of the underlying segmentation model used during initial training. This work is aligns with "Responsible AI" principles, supporting reliable deployment of AI by enhancing robustness and transparency. As the machine learning community continues to focus on building trust and acceptance of AI in society, Segmetron stands as a critical step forward. It can potentially enable the widespread adoption of semantic segmentation-based DL technologies across various fields.

Bibliography

[1] G. M. London and A. P. Guerin, "Influence of arterial pulse and reflected waves on blood pressure and cardiac function," *American Heart Journal*, vol. 138, pp. 220–224, Sept. 1999.

[2] P. V. Vaitkevicius, J. L. Fleg, J. H. Engel, F. C. O'Connor, J. G. Wright, L. E. Lakatta, F. C. Yin, and E. G. Lakatta, "Effects of age and aerobic capacity on arterial stiffness in healthy adults," *Circulation*, vol. 88, pp. 1456–1462, Oct. 1993.

[3] J. L. Cavalcante, J. A. C. Lima, A. Redheuil, and M. H. Al-Mallah, "Aortic stiffness: current understanding and future directions," *Journal of the American College of Cardiology*, vol. 57, pp. 1511–1522, Apr. 2011.

[4] H. Oxlund, L. M. Rasmussen, T. T. Andreassen, and L. Heickendorff, "Increased aortic stiffness in patients with type 1 (insulin-dependent) diabetes mellitus," *Diabetologia*, vol. 32, pp. 748–752, Oct. 1989.

[5] J. J. M. Westenberg, A. J. H. A. Scholte, Z. Vaskova, R. J. van der Geest, M. Groenink, G. Labadie, P. J. van den Boogaard, T. Radonic, Y. Hilhorst-Hofstee, B. J. M. Mulder, L. J. M. Kroft, J. H. C. Reiber, and A. de Roos, "Age-related and regional changes of aortic stiffness in the marfan syndrome: assessment with velocity-encoded MRI," *Journal of Magnetic Resonance Imaging*, vol. 34, pp. 526–531, July 2011.

[6] Yasmin, C. M. McEniery, K. M. O'Shaughnessy, P. Harnett, A. Arshad, S. Wallace, K. Maki-Petaja, B. McDonnell, M. J. Ashby, J. Brown, J. R. Cockcroft, and I. B. Wilkinson, "Variation in the human matrix metalloproteinase-9 gene is associated with arterial stiffness in healthy individuals," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 26, pp. 1799–1805, May 2006.

[7] S. Nistri, J. Grande-Allen, M. Noale, C. Basso, P. Siviero, S. Maggi, G. Crepaldi, and G. Thiene, "Aortic elasticity and size in bicuspid aortic valve syndrome," *European Heart Journal*, vol. 29, pp. 472–479, Dec. 2007.

[8] W.-Y. Chong, W. H. S. Wong, C. S. W. Chiu, and Y.-F. Cheung, "Aortic root dilation and aortic elastic properties in children after repair of tetralogy of Fallot," *The American Journal of Cardiology*, vol. 97, pp. 905–909, Feb. 2006.

[9] A. Redheuil, W.-C. Yu, C. O. Wu, E. Mousseaux, A. de Cesare, R. Yan, N. Kachenoura, D. Bluemke, and J. A. C. Lima, "Reduced ascending aortic strain and distensibility: earliest manifestations of vascular aging in humans," *Hypertension*, vol. 55, pp. 319–326, Jan. 2010.

[10] K. Cruickshank, L. Riste, S. G. Anderson, J. S. Wright, G. Dunn, and R. G. Gosling, "Aortic pulse-wave velocity and its relationship to mortality in diabetes and glucose intolerance: an integrated index of vascular function?," *Circulation*, vol. 106, pp. 2085–2090, Oct. 2002.

[11] T. Willum-Hansen, J. A. Staessen, C. Torp-Pedersen, S. Rasmussen, L. Thijs, H. Ibsen, and J. Jeppesen, "Prognostic value of aortic pulse wave velocity as index of arterial stiffness in the general population," *Circulation*, vol. 113, pp. 664–670, Feb. 2006.

[12] F. U. S. Mattace-Raso, T. J. M. van der Cammen, A. Hofman, N. M. van Popele, M. L. Bos, M. A. D. H. Schalekamp, R. Asmar, R. S. Reneman, A. P. G. Hoeks, M. M. B. Breteler, and J. C. M. Witteman, "Arterial stiffness and risk of coronary heart disease and stroke: the rotterdam study," *Circulation*, vol. 113, pp. 657–663, Feb. 2006.

[13] A. Redheuil, C. O. Wu, N. Kachenoura, Y. Ohyama, R. T. Yan, A. G. Bertoni, G. W. Hundley, D. A. Duprez, D. R. Jacobs, Jr, L. B. Daniels, C. Darwin, C. Sibley, D. A. Bluemke, and J. A. C. Lima, "Proximal aortic distensibility is an independent predictor of all-cause mortality and incident CV events: the MESA study," *Journal of the American College of Cardiology*, vol. 64, pp. 2619–2629, Dec. 2014.

[14] M. O'Rourke, "Mechanical principles in arterial disease," *Hypertension*, vol. 26, pp. 2–9, July 1995.

[15] I. Voges, M. Jerosch-Herold, J. Hedderich, E. Pardun, C. Hart, D. D. Gabbert, J. H. Hansen, C. Petko, H.-H. Kramer, and C. Rickers, "Normal values of aortic dimensions, distensibility, and pulse wave velocity in children and young adults: a cross-sectional study," *Journal of Cardiovascular Magnetic Resonance*, vol. 14, p. 77, Nov. 2012.

[16] C. Stefanadis, C. Stratos, H. Boudoulas, C. Kourouklis, and P. Toutouzas, "Distensibility of the ascending aorta: comparison of invasive and non-invasive techniques in healthy men and in men with coronary artery disease," *European Heart Journal*, vol. 11, pp. 990–996, Nov. 1990.

[17] A. M. Dart, F. Lacombe, J. K. Yeoh, J. D. Cameron, G. L. Jennings, E. Laufer, and D. S. Esmore, "Aortic distensibility in patients with isolated hypercholesterolaemia, coronary artery disease, or cardiac transplant," *Lancet*, vol. 338, pp. 270–273, Aug. 1991.

[18] L. M. Resnick, D. Militianu, A. J. Cunnings, J. G. Pipe, J. L. Evelhoch, and R. L. Soulen, "Direct magnetic resonance determination of aortic distensibility in essential hypertension: relation to age, abdominal visceral fat, and in situ intracellular free magnesium," *Hypertension*, vol. 30, pp. 654–659, Sept. 1997.

[19] H. B. Grotenhuis, J. J. M. Westenberg, P. Steendijk, R. J. van der Geest, J. Ottenkamp, J. J. Bax, J. W. Jukema, and A. de Roos, "Validation and reproducibility of aortic pulse wave velocity as assessed with velocity-encoded MRI," *Journal of Magnetic Resonance Imaging*, vol. 30, pp. 521–526, Sept. 2009.

[20] G. S. Gulsin, D. J. Swarbrick, W. H. Hunt, E. Levelt, M. P. M. Graham-Brown, K. S. Parke, J. V. Wormleighton, F. Y. Lai, T. Yates, E. G. Wilmot, D. R. Webb, M. J. Davies, and G. P. McCann, "Relation of aortic stiffness to left ventricular remodeling in younger adults with type 2 diabetes," *Diabetes*, vol. 67, pp. 1395–1400, Apr. 2018.

[21] A. Singh, M. A. Horsfield, S. Bekele, J. P. Greenwood, D. K. Dawson, C. Berry, K. Hogrefe, D. J. Kelly, J. G. Houston, P. Guntur Ramkumar, A. Uddin, T. Suzuki, and G. P. McCann, "Aortic stiffness in aortic stenosis assessed by cardiovascular MRI: a comparison between bicuspid and tricuspid valves," *European Radiology*, vol. 29, pp. 2340–2349, Nov. 2018.

[22] A. Herment, N. Kachenoura, M. Lefort, M. Bensalah, A. Dogui, F. Frouin, E. Mousseaux, and A. De Cesare, "Automated segmentation of the aorta from phase contrast MR images: validation against expert tracing in healthy volunteers and in patients with a dilated aorta," *Journal of Magnetic Resonance Imaging*, vol. 31, pp. 881–888, Apr. 2010.

[23] R. J. van der Geest, R. A. Niezen, E. E. van der Wall, A. de Roos, and J. H. Reiber, "Automated measurement of volume flow in the ascending aorta using MR velocity maps: evaluation of inter- and intraobserver variability in healthy volunteers," *Journal of Computer Assisted Tomography*, vol. 22, pp. 904–911, Nov. 1998.

[24] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. `http://www.deeplearningbook.org`.

[25] F. Chollet, *Deep Learning with Python*. USA: Manning Publications Co., 1st ed., 2017.

[26] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: A review," *Frontiers in Cardiovascular Medicine*, vol. 7, p. 25, Mar. 2020.

[27] W. Bai, H. Suzuki, C. Qin, G. Tarroni, O. Oktay, P. Matthews, and D. Rueckert, "Recurrent neural networks for aortic image sequence segmentation with sparse annotations," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018 - 21st International Conference, 2018, Proceedings* (A. Frangi, G. Fichtinger, J. Schnabel, C. Alberola-López, and C. Davatzikos, eds.), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), (Cham), pp. 586–594, Springer-Verlag, 2018.

[28] E. Hann, L. Biasiolli, Q. Zhang, I. A. Popescu, K. Werys, E. Lukaschuk, V. Carapella, J. M. Paiva, N. Aung, J. J. Rayner, K. Fung, H. Puchta, M. M. Sanghvi, N. O. Moon, K. E. Thomas, V. M. Ferreira, S. E. Petersen, S. Neubauer, and S. K. Piechnik, "Quality control-driven image segmentation towards reliable automatic image analysis in large-scale cardiovascular magnetic resonance aortic cine imaging," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019* (D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, eds.), (Cham), pp. 750–758, Springer International Publishing, 2019.

[29] G. Ras, L. Ambrogioni, U. Güçlü, and M. A. van Gerven, "Temporal factorization of 3D convolutional kernels," *arXiv preprint arXiv:1912.04075*, 2019.

[30] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Communications of the ACM*, vol. 63, p. 54–63, nov 2020.

[31] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Bi-directional ConvLSTM u-net with densley connected convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[32] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper ConvLSTM for video salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[33] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention u-net for lesion segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 683–687, 2019.

[34] A. Al-Hussaini, A. M. S. E. K. Abdelaty, G. S. Gulsin, J. R. Arnold, M. Garcia-Guimaraes, D. Premawardhana, C. Budgeon, A. Wood, N. Natarajan, K. Mangion, R. Rakhit, S. P. Hoole, T. W. Johnson, C. Berry, I. Hudson, A. H. Gershlick, A. Ladwiniec, J. Kovac, I. Squire, N. J. Samani, S. Plein, G. P. McCann, and D. Adlam, "Chronic infarct size after spontaneous coronary artery dissection: implications for pathophysiology and clinical management," *European Heart Journal*, vol. 41, pp. 2197–2205, June 2020.

[35] D. R. Webb, Z. Z. Htike, D. J. Swarbrick, E. M. Brady, L. J. Gray, J. Biglands, G. S. Gulsin, J. Henson, K. Khunti, G. P. McCann, H. L. Waller, M. A. Webb, J. A. Sargeant, T. Yates, F. Zaccardi, and M. J. Davies, "A randomized, open-label, active comparator trial assessing the effects of 26 weeks of liraglutide or sitagliptin on cardiovascular function in young obese adults with type 2 diabetes," *Diabetes, Obesity and Metabolism*, vol. 22, pp. 1187–1196, Apr. 2020.

[36] G. S. Gulsin, D. J. Swarbrick, L. Athithan, E. M. Brady, J. Henson, E. Baldry, S. Argyridou, N. B. Jaicim, G. Squire, Y. Walters, A.-M. Marsh, J. McAdam, K. S. Parke, J. D. Biglands, T. Yates, K. Khunti, M. J. Davies, and G. P. McCann, "Effects of low-energy diet or exercise on cardiovascular function in working-age adults with type 2 diabetes: A prospective, randomized, open-label, blinded end point trial," *Diabetes Care*, vol. 43, pp. 1300–1310, Mar. 2020.

[37] M. P. M. Graham-Brown, S. F. Adenwalla, F. Y. Lai, W. H. Hunt, K. Parke, G. Gulsin, J. O. Burton, and G. P. McCann, "The reproducibility of cardiac magnetic resonance imaging measures of aortic stiffness and their relationship to cardiac structure in prevalent haemodialysis patients," *Clinical Kidney Journal*, vol. 11, pp. 864–873, June 2018.

[38] Xinapse Systems Ltd, "Jim 9 software."

[39] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," in *arXiv preprint arXiv:1804.02767*, 2018.

[40] S. Sengupta, A. Singh, H. A. Leopold, T. Gulati, and V. Lakshminarayanan, "An automatic deep learning approach for detection of diabetic retinopathy from fundus images," *Journal of Imaging*, vol. 5, no. 12, p. 89, 2019.

[41] A. F. Agarap, "Deep learning using rectified linear units (relu)," 2018.

[42] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[44] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, IEEE, 2013.

[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[46] Y. Tang, W. Zhang, Y. Tang, and W. Zhu, "Facial expression recognition with deeper convolutional neural networks and inception-resnet-v2," *PloS one*, vol. 13, no. 12, p. e0207742, 2018.

[47] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. W. van Uden, C. I. Sanchez, G. Litjens, F.-E. de Leeuw, B. van Ginneken, E. Marchiori, and B. Platel, "Deep multi-scale location-aware 3D convolutional neural networks for automated detection of lacunes of presumed vascular origin," *NeuroImage: Clinical*, vol. 14, pp. 391–399, 2017.

[48] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[49] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[50] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010* (Y. Lechevallier and G. Saporta, eds.), (Heidelberg), pp. 177–186, Physica-Verlag HD, 2010.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[52] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.), vol. 28 of *Proceedings of Machine Learning Research*, (Atlanta, Georgia, USA), pp. 1139–1147, PMLR, 17–19 Jun 2013.

[53] T. Tieleman and G. Hinton, "Lecture 6.5—rmsprop: Divide the gradient by a running average of its recent magnitude," 2012.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," 2015.

[55] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in *International Workshop on Machine Learning in Medical Imaging*, pp. 379–387, Springer, 2017.

[56] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *The Lancet*, vol. 327, no. 8476, pp. 307–310, 1986.

[57] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[58] J. C. Hsu, *Multiple Comparisons: Theory and Methods.* Chapman and Hall/CRC, 1996.

[59] M. Fréchet, "Sur quelques points du calcul fonctionnel," *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, vol. 22, no. 1, pp. 1–74, 1922.

[60] F. Hausdorff, *Grundzüge der Mengenlehre.* Veit, 1914.

[61] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.

[62] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 1, pp. 307–310, Feb. 1986.

[63] L. F. W. Anthony, B. Kanding, and R. Selvan, "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models." ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems, July 2020. arXiv:2007.03051.

[64] "Average CO2 emissions from newly registered motor vehicles in Europe — eea.europa.eu." `https://www.eea.europa.eu/data-and-maps/indicators/average-co2-emissions-from-motor-vehicles/assessment-1`. [Accessed 02-08-2023].

[65] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017.

[66] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, (Los Alamitos, CA, USA), pp. 4489–4497, IEEE Computer Society, dec 2015.

[67] W. Bai, H. Suzuki, J. Huang, C. Francis, S. Wang, G. Tarroni, F. Guitton, N. Aung, K. Fung, S. E. Petersen, S. K. Piechnik, S. Neubauer, E. Evangelou, A. Dehghan, D. P. O’Regan, M. R. Wilkins, Y. Guo, P. M. Matthews, and D. Rueckert, “A population-based phenome-wide association study of cardiac and aortic structure and function,” *Nature Medicine*, vol. 26, pp. 1654–1662, Oct. 2020.

[68] C. M. Francis, M. E. Futschik, J. Huang, W. Bai, M. Sargurupremraj, E. Petretto, A. S. Ho, P. Amouyel, S. T. Engelter, J. S. Ware, S. Debette, P. Elliott, A. Dehghan, and P. M. Matthews, “Genome-wide associations of aortic distensibility suggest causality for aortic aneurysms and brain white matter hyperintensities.,” *medRxiv*, 2021.

[69] D. Amodei and D. Hernandez, “AI and compute,” 2018.

[70] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learning in NLP,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 3645–3650, Association for Computational Linguistics, July 2019.

[71] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, “Quantifying the carbon emissions of machine learning,” *Clinical Orthopaedics and Related Research*, vol. abs/1910.09700, 2019.

[72] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, “Towards the systematic reporting of the energy and carbon footprints of machine learning,” *Journal of Machine Learning Research*, vol. 21, no. 248, pp. 1–43, 2020.

[73] D. A. Patterson, J. Gonzalez, Q. V. Le, C. Liang, L. Munguia, D. Rothchild, D. R. So, M. Texier, and J. Dean, “Carbon emissions and large neural network training,” *Clinical Orthopaedics and Related Research*, vol. abs/2104.10350, 2021.

[74] A. Siouras, S. Moustakidis, A. Giannakidis, G. Chalatsis, I. Liampas, M. Vlychou, M. Hantes, S. Tasoulis, and D. Tsaopoulos, “Knee injury detection using deep learning on MRI studies: A systematic review,” *Diagnostics (Basel)*, vol. 12, Feb. 2022.

[75] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, “Adversarial attacks on medical machine learning,” *Science*, vol. 363, pp. 1287–1289, Mar. 2019.

[76] H. Zhang, Z. Gao, D. Zhang, W. K. Hau, and H. Zhang, “Progressive perception learning for main coronary segmentation in x-ray angiography,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 864–879, 2023.

[77] X. Liu, S. Li, B. Wang, L. Xu, Z. Gao, and G. Yang, “Motion estimation based on projective information disentanglement for 3D reconstruction of rotational coronary angiography,” *Computers in Biology and Medicine*, vol. 157, p. 106743, 2023.

[78] F. Haddad, S. Hunt, D. Rosenthal, and D. Murphy, “Right ventricular function in cardiovascular disease, part i: Anatomy, physiology, aging, and functional assessment of the right ventricle.,” *Circulation*, vol. 117, pp. 1436–48, 2008. 2008,3,18.

[79] F. Grothues, J. Moon, N. Bellenger, G. Smith, H. Klein, and D. Pennell, "Interstudy reproducibility of right ventricular volumes, function, and mass with cardiovascular magnetic resonance," *American Heart Journal*, vol. 147, pp. 218–223, 2004.

[80] T. Geva, "Is MRI the preferred method for evaluating right ventricular size and function in patients with congenital heart disease?: MRI is the preferred method for evaluating right ventricular size and function in patients with congenital heart disease," *Circulation Cardiovascular imaging*, vol. 7, no. 1, pp. 190–197, 2014.

[81] C. Mooij, C. de Wit, D. Graham, A. Powell, and T. Geva, "Reproducibility of MRI measurements of right ventricular size and function in patients with normal and dilated ventricles," *Journal of magnetic resonance imaging*, vol. 28, no. 1, pp. 67–73, 2008.

[82] N. Keenan, G. Captur, G. McCann, C. Berry, S. Myerson, T. Fairbairn, L. Hudsmith, *et al.*, "Regional variation in cardiovascular magnetic resonance service delivery across the UK," *Heart*, vol. 107, pp. 1974–1979, 2021.

[83] L. G. Rudski, W. W. Lai, J. Afilalo, and et al., "Guidelines for the echocardiographic assessment of the right heart in adults: a report from the american society of echocardiography endorsed by the european association of echocardiography, a registered branch of the european society of cardiology, and the canadian society of echocardiography," *Journal of the American Society of Echocardiography*, vol. 23, no. 7, pp. 685–713, 2010.

[84] C. Jenkins, J. Chan, K. Bricknell, M. Strudwick, and T. Marwick, "Reproducibility of right ventricular volumes and ejection fraction using real-time three-dimensional echocardiography: comparison with cardiac MRI," *Chest*, vol. 131, pp. 1844–1851, 2007.

[85] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," in *Advances In Neural Information Processing Systems*, vol. 35, pp. 507–520, 2022.

[86] A. Vaswani *et al.*, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[87] J. Brophy and D. Lowd, "Instance-based uncertainty estimation for gradient-boosted regression trees," in *Advances In Neural Information Processing Systems*, vol. 35, pp. 11145–11159, 2022.

[88] Y. Lin and Y. Jeon, "Random forests and adaptive nearest neighbors," *Journal of the American Statistical Association*, vol. 101, no. 474, pp. 578–590, 2006.

[89] T. Daghistani and R. Alshammari, "Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes," *Journal of Advances in Information Technology*, vol. 11, pp. 78–83, may 2020.

[90] J. Kochav, J. Chen, L. Nambiar, H. Mitlak, A. Kushman, R. Sultana, E. Horn, *et al.*, "Novel echocardiographic algorithm for right ventricular mass quantification: Cardiovascular magnetic resonance and clinical prognosis validation," *Journal Of The American Society Of Echocardiography*, vol. 34, pp. 839–850.e1, 2021.

[91] P. N. Kampaktsis, T. A. Bohoran, M. Lebehn, L. McLaughlin, J. Leb, Z. Liu, S. Moustakidis, A. Siouras, A. Singh, R. T. Hahn, G. P. McCann, and A. Giannakidis, "An attention-based deep learning method for right ventricular quantification using 2D echocardiography: Feasibility and accuracy," *Echocardiography*, vol. 41, no. 1, p. e15719, 2024.

[92] S. E. Petersen, N. Aung, M. M. Sanghvi, F. Zemrak, K. Fung, J. M. Paiva, and et al., "Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort," *Journal of Cardiovascular Magnetic Resonance*, vol. 19, no. 1, p. 18, 2017.

[93] D. Aha, D. Kibler, and M. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37–66, 1991.

[94] A. Davies and Z. Ghahramani, "The random forest kernel and other kernels for big data from random partitions," *arXiv preprint arXiv:1402.4293*, 2014.

[95] J. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals Of Statistics*, vol. 29, pp. 1189–1232, 2001.

[96] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[97] T. Bohoran, P. Kampaktsis, L. McLaughlin, J. Leb, S. Moustakidis, G. McCann, and A. Giannakidis, "Right ventricular volume prediction by feature tokenizer transformer-based regression of 2D echocardiography small-scale tabular data," in *Functional Imaging And Modeling Of The Heart*, pp. 292–300, 2023.

[98] Turing, "The ultimate guide to transformer deep learning," Feb 2022.

[99] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances In Neural Information Processing Systems*, vol. 34, pp. 18932–18943, 2021.

[100] J. Devlin, M. Chang, *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805, 2018.

[101] Q. L. Wang, T. Xiao, J. Zhu, C. Li, D. Wong, and L. S. Chao, "Learning deep transformer models for machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[102] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, (San Diega, CA, USA), 2015.

[103] T. Chen and C. Guestrin, "XGBoost," in *Proceedings Of The 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*, 2016.

[104] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances In Neural Information Processing Systems*, vol. 30, 2017.

[105] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Advances In Neural Information Processing Systems*, vol. 31, 2018.

[106] T. Duan, A. Anand, D. Ding, K. Thai, S. Basu, A. Ng, and A. Schuler, "NGBoost: Natural gradient boosting for probabilistic prediction," in *Proceedings Of The 37th International Conference On Machine Learning*, vol. 119, pp. 2690–2700, 2020. 2020,7,13.

[107] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International journal of forecasting*, vol. 22, no. 4, pp. 679–688, 2006.

[108] C. J. Willmott, "On the evaluation of model performance in physical geography," *Progress in Physical Geography*, vol. 5, no. 2, pp. 184–202, 1981.

[109] T. F. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[110] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.

[111] J. S. Armstrong and F. Collopy, *Error measures for generalizing about forecasting methods: Empirical comparisons.* International Institute of Forecasters, 1992.

[112] S. Makridakis, "Accuracy measures: theoretical and practical concerns," *International Journal of Forecasting*, vol. 9, no. 4, pp. 527–529, 1993.

[113] N. R. Draper and H. Smith, *Applied Regression Analysis.* Wiley-Interscience, 3rd ed., 1998.

[114] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning.* Springer, 2013.

[115] K. Pearson, "Note on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.

[116] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[117] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938.

[118] J. E. Matheson and R. L. Winkler, "Scoring rules for continuous probability distributions," *Management Science*, vol. 22, no. 10, pp. 1087–1095, 1976.

[119] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[120] K. P. Murphy, *Machine learning: a probabilistic perspective.* MIT press, 2012.

[121] H. Hersbach, "Decomposition of the continuous ranked probability score for ensemble prediction systems," *Weather and Forecasting*, vol. 15, no. 5, pp. 559–570, 2000.

[122] J. Bracher, E. L. Ray, T. Gneiting, and N. G. Reich, "Evaluating epidemic forecasts in an interval format," *PLOS Computational Biology*, vol. 17, no. 2, p. e1008618, 2021.

[123] T. Gneiting and A. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal Of The American Statistical Association*, vol. 102, pp. 359–378, 2007.

[124] W. Conover, *Practical Nonparametric Statistics.* John Wiley & Sons, 3 ed., 1999.

[125] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference.* CRC Press, 5th ed., 2011.

[126] J. L. Fleiss, *The Design and Analysis of Clinical Experiments.* New York: Wiley, 1986.

[127] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychological Methods*, vol. 1, no. 1, pp. 30–46, 1996.

[128] P. E. Shrout and J. L. Fleiss, “Intraclass correlations: uses in assessing rater reliability,” *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979.

[129] O. Sprangers, S. Schelter, and M. Rijke, “Probabilistic gradient boosting machines for large-scale probabilistic regression,” in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.

[130] A. Malinin, L. Prokhorenkova, and A. Ustimenko, “Uncertainty in gradient boosting via ensembles,” in *International Conference On Learning Representations*, 2021.

[131] T. Bohoran, P. McLaughlin, L. Leb, J. Moustakidis, S. McCann, and A. Giannakidis, “Embracing uncertainty flexibility: Harnessing a supervised tree kernel to empower ensemble modelling for 2D echocardiography-based prediction of right ventricular volume,” in *The 16th International Conference on Machine Vision (ICMV 2023)*, 2023.

[132] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, “TabTransformer: Tabular data modeling using contextual embeddings,” 2020.

[133] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, “Trustworthy AI: From principles to practices,” *ACM computing surveys*, vol. 55, 2023. 2023,1.

[134] M. Tokodi, B. Magyar, A. Soós, M. Takeuchi, M. Tolvaj, B. K. Lakatos, T. Kitano, Y. Nabeshima, A. Fábián, M. B. Szigeti, A. Horváth, B. Merkely, and A. Kovács, “Deep learning-based prediction of right ventricular ejection fraction using 2D echocardiograms,” *JACC: Cardiovascular Imaging*, vol. 16, pp. 1005–1018, August 2023. Epub 2023 May 10.

[135] N. Srinivasan and R. Schilling, “Sudden cardiac death and arrhythmias,” *Arrhythmia and Electrophysiology Review*, vol. 7, pp. 111–117, Jun 2018.

[136] M. Di Cesare, H. Bixby, T. Gaziano, L. Hadeed, C. Kabudula, D. V. McGhie, J. Mwangi, B. Pervan, P. Perel, D. Piñeiro, S. Taylor, and F. Pinto, “World heart report 2023: Confronting the world’s number one killer,” 2023.

[137] F. Alarsan and M. Younes, “Analysis and classification of heart diseases using heartbeat features and machine learning algorithms,” *Journal of Big Data*, vol. 6, 2019. [Online]. Available: `https://doi.org/10.1186/s40537-019-0244-x`.

[138] N. Urushibata, K. Murata, H. Endo, A. Yoshiyuki, and Y. Otomo, “Evaluation of manual chest compressions according to the updated cardiopulmonary resuscitation guidelines and the impact of feedback devices in an educational resuscitation course,” *BMC Emergency Medicine*, vol. 20, no. 49, 2020.

[139] S. Irfan, N. Anjum, T. Althobaiti, A. A. Alotaibi, A. B. Siddiqui, and N. Ramzan, “Heartbeat classification and arrhythmia detection using a multi-model deep-learning technique,” *Sensors*, vol. 22, no. 15, p. 5606, 2022.

[140] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[141] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, “ECG heartbeat classification: A deep transferable representation,” in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pp. 443–444, 2018.

[142] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations (ICLR)*, 2015.

[143] W. E, "A proposal on machine learning via dynamical systems," *Communications in Mathematics and Statistics*, vol. 5, pp. 1–11, Mar 2017. [Online]. Available: https://doi.org/10.1007/s40304-017-0103-z.

[144] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Weinberger, "Deep networks with stochastic depth," in *Computer Vision – ECCV 2016*, pp. 646–661, 2016.

[145] G. Moody and R. Mark, "The impact of the MIT-BIH Arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, pp. 45–50, 2001.

[146] G. Moody and R. Mark, "MIT–BIH Arrhythmia database," 1992. [Online]. Available: https://physionet.org/content/mitdb/.

[147] A. Khamparia, D. Gupta, N. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, "Sound classification using convolutional neural network and tensor deep stacking network," *IEEE Access*, vol. 7, pp. 7717–7727, 2019.

[148] A. for the Advancement of Medical Instrumentation, "Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms." ANSI/AAMI EC38, 1998.

[149] C. Dong, L. Liu, Z. Li, and J. Shang, "Towards adaptive residual network training: A neural-ODE perspective," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, pp. 2616–2626, Jul 2020. [Online]. Available: https://proceedings.mlr.press/v119/dong20c.html.

[150] Y. Tsuzuku, I. Sato, and M. Sugiyama, "Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 31, 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/485843481a7edacbfce101ecb1e4d2a8-Paper.pdf.

[151] H. Gouk, E. Frank, B. Pfahringer, and M. Cree, "Regularisation of neural networks by enforcing Lipschitz continuity," 2020.

[152] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, p. 665–673, Nov. 2020.

[153] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. Earnshaw, I. Haque, S. M. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang, "Wilds: A benchmark of in-the-wild distribution shifts," in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664, PMLR, 18–24 Jul 2021.

[154] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*, pp. 201–205. Neural Information Processing Series, Yale University Press, 2009.

[155] S. Kumari and P. Singh, "Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives," *Computers in Biology and Medicine*, vol. 170, p. 107912, 2024.

[156] F. Liu, W. Xu, J. Lu, G. Zhang, A. Gretton, and D. J. Sutherland, "Learning deep kernels for non-parametric two-sample tests," in *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. Singh, eds.), vol. 119 of *Proceedings of Machine Learning Research*, pp. 6316–6326, PMLR, 13–18 Jul 2020.

[157] S. Zhao, A. Sinha, Y. He, A. Perreault, J. Song, and S. Ermon, “Comparing distributions by measuring differences that affect decision making,” in *International Conference on Learning Representations*, 2022.

[158] J. J. Ren, S. Fort, J. Z. Liu, A. G. Roy, S. Padhy, and B. Lakshminarayanan, “A simple fix to Mahalanobis distance for improving near-OOD detection,” *ArXiv*, vol. abs/2106.09022, 2021.

[159] S. Jang, S. Park, I. Lee, and O. Bastani, “Sequential covariate shift detection using classifier two-sample tests,” in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 9845–9880, PMLR, 17–23 Jul 2022.

[160] P. de Jorge, R. Volpi, P. H. Torr, and G. Rogez, “Reliability in semantic segmentation: Are we on the right track?,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7173–7182, 2023.

[161] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, p. 2481–2495, Dec. 2017.

[162] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “SegFormer: Simple and efficient design for semantic segmentation with transformers,” in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 12077–12090, Curran Associates, Inc., 2021.

[163] D. Zhou, Z. Yu, E. Xie, C. Xiao, A. Anandkumar, J. Feng, and J. M. Alvarez, “Understanding the robustness in vision transformers,” in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 27378–27394, PMLR, 17–23 Jul 2022.

[164] C. Kamann and C. Rother, “Benchmarking the robustness of semantic segmentation models with respect to common corruptions,” *International Journal of Computer Vision*, vol. 129, p. 462–483, Sept. 2020.

[165] J. Liang, R. He, and T. Tan, “A comprehensive survey on test-time adaptation under distribution shifts,” *International Journal of Computer Vision*, July 2024.

[166] S. Goldwasser, A. T. Kalai, Y. T. Kalai, and O. Montasser, “Beyond perturbations: Learning guarantees with arbitrary adversarial test examples,” *Clinical Orthopaedics and Related Research*, vol. abs/2007.05145, 2020.

[167] T. Ginsberg, Z. Liang, and R. G. Krishnan, “A learning based hypothesis test for harmful covariate shift,” in *The Eleventh International Conference on Learning Representations*, 2023.

[168] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016.

[169] M. D. Ernst, “Permutation methods: A basis for exact inference,” *Statistical Science*, vol. 19, no. 4, pp. 676–685, 2004.

[170] T. A. Bohoran, K. S. Parke, M. P. M. Graham-Brown, M. Meisuria, A. Singh, J. Wormleighton, D. Adlam, D. Gopalan, M. J. Davies, B. Williams, M. Brown, G. P. McCann, and A. Giannakidis, "Resource efficient aortic distensibility calculation by end to end spatiotemporal learning of aortic lumen from multicentre multivendor multidisease CMR images," *Scientific Reports*, vol. 13, Dec. 2023.

[171] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M.-M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P.-M. Jodoin, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[172] X. Zhuang, "Multivariate mixture model for myocardial segmentation combining multisource images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 2933–2946, 2019.

[173] C. F. Baumgartner, L. M. Koch, M. Pollefeys, and E. Konukoglu, "An exploration of 2d and 3d deep learning techniques for cardiac MR image segmentation," *Clinical Orthopaedics and Related Research*, vol. abs/1709.04496, 2017.

[174] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. DePristo, J. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," in *Advances in Neural Information Processing Systems*, vol. 32, pp. 14680–14691, 2019.

[175] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola, "A kernel two-sample test," in *Journal of Machine Learning Research*, vol. 13, pp. 723–773, 2012.

[176] Y. Li, K. Swersky, and R. Zemel, "Mmd nets: Maximum mean discrepancy learning for distributional regression," in *International Conference on Learning Representations (ICLR)*, 2015.

[177] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Advances in neural information processing systems*, pp. 137–144, 2007.

[178] A. N. Kolmogorov, "Sulla determinazione empirica di una legge di distribuzione," *Giornale dell'Istituto Italiano degli Attuari*, vol. 4, pp. 83–91, 1933.

[179] N. Smirnov, "Table for estimating the goodness of fit of empirical distributions," *Annals of Mathematical Statistics*, vol. 19, no. 2, pp. 279–281, 1948.

[180] G. Qi, "Loss-sensitive generative adversarial networks on Lipschitz densities," *International Journal of Computer Vision*, vol. 128, pp. 1118–1140, 2020.

[181] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier, "Parseval networks: Improving robustness to adversarial examples," in *International Conference on Machine Learning (ICML)*, pp. 854–863, 2017.