UNIVERSITÀ
DEGLI STUDI
FIRENZE

PHD PROGRAM IN SMART COMPUTING
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE (DINFO)

Unsupervised and Contextual Anomaly Detection with Application to Cardiology

Luca Bindini

Dissertation presented in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Smart Computing

PhD Program in Smart Computing
University of Florence, University of Pisa, University of Siena

Unsupervised and Contextual Anomaly Detection with Application to Cardiology

Luca Bindini

Advisor:

Prof. Paolo Frasconi

Head of the PhD Program:

Prof. Stefano Berretti

Evaluation Committee:
Prof. Cesare Alippi, *Università della Svizzera italiana*
Prof. Irwin King, *The Chinese University of Hong Kong*

XXXVIII ciclo — October 2025

To everyone who wishes to study but has no opportunity

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Paolo Frasconi, who has been an invaluable guide throughout these three years of my Ph.D. His mentorship, constant support, and insightful advice have played a fundamental role in my development as a researcher.

I am sincerely grateful to the members of my Supervisory Committee, Prof. Franco Scarselli and Prof. Jesse Davis, for their precious feedback and encouragement over the years.

I owe a special thanks to Prof. Jesse Davis for hosting me for three months at KU Leuven during my research stay abroad, an experience that greatly enriched me both scientifically and personally. During that period, I had the pleasure of working with Lorenzo Perini, who closely followed my work and taught me a great deal.

My heartfelt thanks also go to Prof. Simone Marinai and Prof. Marco Lippi from my lab, for their continuous support and helpful discussions throughout these years.

I would also like to thank Prof. Stefano Pagani from Politecnico di Milano for introducing me to the fascinating world of electrophysiology, which greatly broadened my scientific horizons.

A very special thank you goes to my girlfriend Valeria, my safe place, the one who has been able to lift me up in the darkest moments of these past years. She knows exactly what I mean. It is thanks to her support that I am here today.

I am profoundly grateful to my sister Alessia, for always being there for me, for her affection, and for the countless moments of laughter and lightness that helped me through these years.

I am deeply thankful to my mother Barbara and Iacopo, for always supporting me in my studies and for making me feel their pride every step of the way.

I thank my father Fabio for believing in me and for the support he provided throughout this journey.

A warm thought goes to my grandparents Paolo, Adele, and Gloria, who raised me with love, made me feel special, and gave me everything I needed. I also want to remember my grandfather Mario; wherever you are now, I know you are proud of me.

My gratitude extends to Claudia, Leonardo, Enrico, Elisa, Marta, and Laura, for the affection you have shown me over the years.

I also wish to thank Sandro, Fiorenza, and Mariateresa for believing in me and for their constant support.

I am truly lucky to have friends like Emanuele, Simone, and Athos who have always been there when I needed to relax or simply have a chat.

I would also like to thank all my university lunch companions: the LESC group (Daniele, Dasara, Chiara, Giulia, Francesca, and Peng Peng), Giacomo, and Roberto, the best “disturber” I have ever met.

Finally, a heartfelt thank you goes to my lab mates, Valeria (again), Eleonora, Kimiyia, Simone, Vittorio, and Lorenzo, for being the most fantastic group one could ask for. A special mention also goes to the former members of the AI Lab, Curzio and Andrea, who kept me company during my first year of the Ph.D., when I felt completely lost.

I am sure I have forgotten someone, but to all of you whom I have met along this journey and who made it brighter, thank you, truly.

Abstract

Reliable clinical decision support requires quantitative tools that can assess whether a measurement is "normal" for a given patient and, crucially, whether such an assessment is trustworthy in the presence of heterogeneous populations, acquisition variability, and limited reference data. In many cardiology applications, obtaining unambiguous labels is inherently difficult. Intracardiac electrograms (EGMs) acquired during ablation procedures exhibit complex and patient-specific morphologies that are not easily mapped to discrete classes. At the same time, echocardiographic reference equations for aortic diameters may become unreliable in sparsely sampled or distribution-shifted patient contexts. These challenges motivate the investigation of data-driven approaches based on unsupervised learning, anomaly detection, and explicit uncertainty quantification.

This thesis explores complementary methodological directions for anomaly detection with applications to cardiology. First, we propose an unsupervised deep anomaly detection framework to characterize atrial EGM morphology directly from raw waveforms. The framework produces robust anomaly scores that correlate with established electrophysiological indicators, including voltage, fractionation, and duration, while yielding coherent electroanatomical maps without the need for manually tuned thresholds or ad hoc combinations of handcrafted features. This provides a more synthetic and morphology-oriented description of atrial substrate.

Second, we introduce the *normalcy score* (NS), a probabilistic generalization of Z-score reasoning for contextual anomaly detection, in which the score itself is treated as a random variable rather than as a deterministic quantity. NS leverages heteroscedastic Gaussian processes to model context-dependent mean and variance, thereby distinguishing between aleatoric and epistemic uncertainty. This formulation enables uncertainty-aware and interpretable assessments, particularly in borderline or poorly supported regions of the input space, where overconfident point estimates may be misleading.

Third, we instantiate the same uncertainty-aware principle in echocardiography by reformulating the classical aortic Z-score. The resulting score provides clinicians with both an expected score and a highest-density interval, explicitly signaling when limited reference support makes the assessment less reliable. In this way, the proposed approach extends a familiar clinical tool while improving transparency and supporting a more cautious interpretation of abnormality.

Finally, we include a complementary methodological study on few-shot source attribution of AI-generated images. By training compact tiny autoencoders, we show that reconstruction residuals can be exploited as lightweight and discriminative signatures, while remaining compatible with class-incremental updates under severe data constraints. Although this chapter is not centered on a cardiology application, it reinforces a broader

methodological message of the thesis: reconstruction-based representation pipelines can provide effective and practical solutions in settings where labels are scarce and compact models are desirable.

Overall, the thesis shows how unsupervised scoring, contextual probabilistic modeling, and uncertainty-aware inference can support more reliable quantitative assessment across heterogeneous biomedical settings, while also highlighting the importance of interpretable scores and explicit reliability estimates in high-stakes decision-making.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Modern clinical decision support increasingly relies on quantitative measurements extracted from heterogeneous data sources, yet the path from measurement to action is often hindered by three recurring issues: scarcity of reliable labels, heterogeneity across patients and acquisition protocols, and explicit and trustworthy uncertainty estimates (Litjens et al., 2017; Esteva et al., 2019; Feeny et al., 2020; Kelly et al., 2019). In biomedical practice, machine-learning systems are rarely judged only by raw predictive accuracy: they are expected to produce outputs that remain interpretable under population variability, acquisition artifacts, and imperfect reference cohorts. This thesis is motivated by two cardiology scenarios in which these difficulties are particularly evident and in which score-based decision support is often more appropriate than hard supervised labeling.

Intracardiac electrograms (EGMs) in atrial fibrillation. Atrial fibrillation (AF) is the most common sustained arrhythmia in clinical practice and remains associated with stroke, heart failure, hospitalizations, and reduced quality of life (Hindricks et al., 2021). Catheter ablation is an important rhythm-control strategy for selected patients with symptomatic AF. In first-time procedures, the standard approach is usually pulmonary vein isolation, especially in patients with paroxysmal AF, i.e., AF characterized by self-terminating episodes (Haïssaguerre et al., 1998; Piccini et al., 2009). In persistent AF, however, outcomes are less satisfactory, and the search for patient-specific substrate targets beyond pulmonary vein isolation remains an active and controversial line of research (Ramirez et al., 2017; Parameswaran et al., 2021; Frontera et al., 2021). High-density electroanatomical mapping enables clinicians to record thousands of bipolar EGMs across the atrial surface, turning local waveforms into a spatial description of the arrhythmogenic substrate. Yet mapping EGM morphology to a discrete label (normal vs. abnormal, or more refined classes) is intrinsically difficult: morphology is complex, strongly affected by acquisition conditions, and still interpreted through partially debated biomarkers

such as voltage, fractionation, and duration (Konings et al., 1994; Jadidi et al., 2012; Anter and Josephson, 2016; Sim et al., 2019; Wong et al., 2019). Consequently, methods that require supervised labels are often impractical, and there is a need for unsupervised, morphology-aware tools that provide consistent indicators without relying on hand-crafted thresholds alone.

Normalcy assessment of echocardiographic aortic diameters. Thoracic aortic dilatation is associated with potentially severe outcomes, and transthoracic echocardiography is routinely used to measure aortic diameters at specific anatomical levels (Kim et al., 2016; Isselbacher et al., 2022). In practice, a measured diameter y is interpreted relative to a patient context vector x (e.g., age, sex, body size) using reference equations and Z-scores (Campens et al., 2014; Colan, 2013; Frasconi et al., 2021). While Z-scores are interpretable and widely adopted, classical implementations typically assume a linear mean relationship and constant (homoscedastic) residual variance; more importantly, they provide no explicit indication of reliability when the patient context lies in a region that is poorly represented by the reference cohort (Dallaire et al., 2015; Patel et al., 2022a). For borderline cases, pediatric or syndromic populations, and under-represented combinations of age and body size, this limitation is not merely academic: it directly affects how aggressively measurements are monitored and interpreted.

1.1 A score-based anomaly-detection perspective

These two case studies involve different data modalities and different forms of clinical reasoning, but they share a structural difficulty. In both cases, the clinically useful notion of abnormality is gradual rather than categorical. An EGM may be more or less morphologically irregular; an aortic diameter may be more or less unexpectedly large for a given patient profile. A continuous score is often a better interface than a hard label. Scores can be mapped onto anatomy, compared across patients, related to existing biomarkers, thresholded at different operating points, or accompanied by uncertainty summaries. This thesis adopts that score-centric view and studies how such scores can be learned, interpreted, and made trustworthy when labels are scarce or unreliable.

More broadly, the thesis sits at the intersection of AI for medicine and anomaly detection. Much of the impact of machine learning in healthcare has come from supervised settings in which the target label is reasonably stable and large labeled datasets can be assembled (Litjens et al., 2017; Esteva et al., 2019; Feeny et al., 2020; Kelly et al., 2019). Many clinically relevant tasks, however, are not naturally closed-set classification problems. The target concept may be gradual, observer-dependent, or defined only relative to patient covariates. In these cases, anomaly detection

provides a more natural framing because they learn a reference notion of regularity and quantify deviation from it rather than forcing a classification (Schölkopf et al., 2001; Chandola et al., 2009; Chalapathy and Chawla, 2019; Pang et al., 2021).

This perspective is already present in the biomedical literature. Anomaly detection has been used for clinical event detection, physiological monitoring, and medical imaging (Hauskrecht et al., 2007; Wolleb et al., 2022). Yet two limitations remain recurring. First, many methods return a score that is difficult to relate to established clinical reasoning or to integrate into workflow-friendly representations. Second, when abnormality is contextual rather than marginal, the methodological picture is still comparatively fragmented: existing approaches can model context-dependent deviation, but they rarely communicate how reliable that assessment is when the reference data are sparse or unevenly distributed (Valko et al., 2011; Li and van Leeuwen, 2023; Hüllermeier and Waegeman, 2021). This limitation is particularly consequential in medicine, where a score close to a threshold may influence follow-up intensity, treatment escalation, or reassurance.

These are the issues addressed in the remainder of the thesis. Rather than asking only whether an observation should be assigned to a class, the dissertation studies how to construct scores that remain interpretable, anatomically or clinically meaningful, and explicit about their own reliability. In the EGM setting, this means learning morphology-aware anomaly scores that can be projected back onto electroanatomical maps. In the contextual setting, it means extending familiar Z-score reasoning into an uncertainty-aware normalcy assessment. The complementary chapter on tiny autoencoders then serves as a methodological stress test outside medicine: it asks whether compact reconstruction-based score pipelines remain useful when labels are scarce, new classes arrive over time, and full retraining is undesirable.

1.2 Contributions

The thesis is organized around one methodological contribution and three applications. The first three items form the main cardiology storyline of the dissertation, whereas the fourth broadens the methodological perspective by testing a related reconstruction-based scoring idea under a different form of data scarcity.

1. **Unsupervised deep anomaly detection for EGM morphology.** We investigate representative deep anomaly detection algorithms, including reconstruction-based and one-class objectives, to derive morphology-aware anomaly scores from intracardiac EGMs without requiring labels. We validate the learned scores by correlation with established electrophysiological indicators and by spatial coherence on electroanatomical maps (Bindini et al., 2024b). The result-

ing anomaly maps condense several aspects of waveform morphology into a single continuous quantity that can be read alongside classical descriptors.

2. **Normalcy score for contextual anomaly detection with uncertainty.** We propose a probabilistic generalization of Z-score reasoning in which both the conditional mean and variance are learned as functions of context, and the resulting normalcy score is treated as a random variable (Bindini et al., 2026b). NS uses heteroscedastic Gaussian process regression (Rasmussen and Williams, 2006; Titsias, 2009) and yields posterior summaries such as highest-density intervals that quantify epistemic uncertainty.

3. **Reliable assessment of aortic diameters.** We instantiate the normalcy score framework in a clinically established workflow and obtain a Bayesian reformulation of the classical aortic Z-score. In this sense, the Bayesian Z-score chapter is the direct cardiology-oriented application of the methodological contribution developed in the NS chapter. The resulting score provides clinicians with an expected value and a credible interval, explicitly warning when the assessment straddles clinical thresholds due to limited reference coverage (Campens et al., 2014; Frasconi et al., 2021; Bindini et al., 2026a).

4. **Tiny autoencoder representations for few-shot source attribution.** We report a complementary representation-learning study in multimedia forensics: a modular attribution system based on tiny autoencoders trained from very few samples per generative model. Although the application domain differs from cardiology, the chapter is not an isolated add-on. It stress-tests one recurrent idea of the thesis, namely that compact reconstruction-based representation-and-score pipelines can remain useful when data are scarce, and demonstrates how this idea behaves in a class-incremental setting (Bindini et al., 2024a).

1.3 Thesis structure

After a general background chapter that reviews anomaly detection, contextual anomaly detection, and the clinical foundations needed for the cardiology applications, the dissertation develops the cardiology storyline in Chapters 3–5 and then broadens the methodological discussion with the complementary tiny-autoencoder chapter. The remainder of the dissertation is organized as follows.

- Chapter 2 provides the background of the thesis, covering anomaly detection, contextual anomaly detection, source attribution of generative models, and the clinical foundations needed for the cardiology applications.

- Chapter 3 presents the first application, namely unsupervised anomaly detection for intracardiac EGM characterization and electroanatomical mapping.

- Chapter 4 introduces the main methodological contribution of the thesis, the normalcy score.
- Chapter 5 presents a cardiology-oriented application of the normalcy score framework.
- Chapter 6 presents the complementary application based on tiny autoencoders for few-shot source attribution of AI-generated images.
- Chapter 7 summarizes the main findings, discusses methodological and clinical implications, and outlines future research directions.

1.4 Reproducibility

A dissertation that emphasizes score reliability should also make the underlying research process as reproducible as possible. For this reason, the main methodological contributions of the thesis are accompanied by public code repositories that document the implementation details needed to reproduce the experiments, inspect the models, and adapt the pipelines to related datasets. Rather than embedding extensive pseudocode in the manuscript, the thesis points to these repositories as the primary research artifacts and treats the written text as the conceptual and methodological specification.

The repositories associated with the thesis are the following:

- **Deep Atrial Anomaly Detection**: `https://github.com/lucabindini/DeepAtrialAnomalyDetection`
- **Normalcy Score**: `https://github.com/lucabindini/NormalcyScore`
- **Tiny Autoencoders**: `https://github.com/lucabindini/TinyAutoencoders`

A final point concerns data availability. Some of the datasets used in the thesis are clinical and therefore subject to privacy constraints, governance policies, or local institutional agreements. In such cases, open-sourcing raw data may be impossible or only partially feasible. In particular, the atrial EGM dataset and the aortic dataset used in the clinical chapters are private datasets. They may nevertheless be made available upon reasonable request and subject to the approval of the corresponding institutional and ethical constraints. By contrast, the benchmark datasets used in the normalcy score chapter and in the complementary chapter on tiny autoencoders are public and can be accessed online, which facilitates independent verification of the methodological results reported in those parts of the thesis.

Chapter 2

Background

This chapter reviews the literature that motivates the methodological choices made in the thesis. Since the dissertation does not revolve around a single algorithmic family, the purpose of the review is not merely to list papers, but to provide a roadmap through a set of connected questions: how normality can be learned from data when explicit labels are scarce, how this notion can be converted into a quantitative score, how such a score can be interpreted in biomedical applications, and when uncertainty must become part of the output rather than an afterthought.

We begin from the general anomaly detection problem and the semantics of score-based decision support; we then discuss the principal methodological families, from classical geometric and density-based detectors to deep representation-learning approaches; next we turn to contextual anomaly detection and uncertainty-aware models, which provide the conceptual bridge to the normalcy score and to its application; we then position the complementary chapter on tiny autoencoders within the broader literature on generative models and source attribution; finally, we close with the medical background needed to read the atrial-electrogram and aortic-diameter applications in their clinical context.

2.1 Conceptual overview

Anomaly detection is commonly described as the task of identifying observations that deviate from what is considered normal. In the classical literature, this is often formalized as support estimation, novelty detection, or outlier detection: one observes samples from an unknown data-generating distribution and attempts to characterize the region of input space where nominal observations concentrate, so that observations outside that region receive higher anomaly scores (Schölkopf et al., 2001; Chandola et al., 2009; Markou and Singh, 2003). This perspective remains useful because it emphasizes an aspect that is central throughout the thesis: the learner does not observe normality directly, but only a finite, noisy, and sometimes

contaminated sample from which normal structure must be inferred.

A score-based view is particularly important in high-stakes domains. In many practical applications, and especially in medicine, the immediate objective is not a hard binary decision but a ranking or a continuous deviation measure. A real-valued score can be thresholded differently depending on the operating point, compared against established biomarkers, mapped across an anatomical structure, or interpreted jointly with confidence information. For this reason, the present thesis consistently treats anomaly detection as a problem of learning useful scores, rather than as purely categorical prediction tasks. This position is consistent with the broader medical-AI literature, which increasingly stresses that decision support systems should be evaluated not only by predictive discrimination but also by calibration, robustness, and fitness for clinical use (Litjens et al., 2017; Esteva et al., 2019; Feeny et al., 2020).

A minimal taxonomy helps clarify the field. From the point of view of the anomaly itself, one usually distinguishes:

- **Point anomalies**, namely individual observations that appear unusual in isolation.
- **Collective anomalies**, where a segment, sequence, or group is abnormal even if single elements are not individually extreme.
- **Contextual anomalies**, where abnormality can only be defined relative to covariates.

From the point of view of supervision, one commonly distinguishes:

- **Supervised anomaly detection**, which assumes access to labeled anomalies and is uncommon in the biomedical scenarios addressed here.
- **One-class or semi-supervised learning**, where training data are assumed to be mostly nominal, and the goal is to estimate the support of normality.
- **Fully unsupervised learning**, where the training set may be contaminated by anomalies, but the latter are assumed to be comparatively rare or structurally distinct.

These distinctions already anticipate the two principal application settings of the thesis. Intracardiac EGMs are treated as waveform-level objects for which discrete ground-truth pathology labels are not realistically available; by contrast, aortic-diameter assessment is a contextual problem, since the same numerical diameter may be normal or abnormal depending on age, sex, body size, and related covariates.

A simple but useful contamination model writes the training distribution as

$$P_{\text{train}} = (1 - \varepsilon)P_0 + \varepsilon P_1, \tag{2.1}$$

where P_0 denotes the nominal distribution, P_1 the anomalous component, and ε the contamination rate. Even if this model is simplistic, it conveys an important lesson: in unsupervised anomaly detection, the learner never sees the nominal distribution in isolation. In clinical deployment, the problem is compounded by dataset heterogeneity and by shifts between the population used to build the model and the population in which the model is eventually interpreted. This is precisely why score reliability matters as much as score magnitude.

The score-centric perspective adopted in this thesis suggests four practical criteria by which methods should be judged:

- **Discrimination**: nominal and clearly abnormal cases should receive systematically different scores.
- **Stability**: nuisance perturbations, preprocessing choices, and acquisition artifacts should not dominate the ranking.
- **Interpretability**: the score should be relatable to known biomarkers, clinically meaningful signal characteristics, or familiar statistical concepts.
- **Reliability**: the method should indicate when the score depends on extrapolation or weak support in the reference data.

These criteria map naturally to the directions developed later in the thesis. The EGM chapter focuses on morphology-oriented discrimination and map-level stability, while normalcy score chapter emphasizes contextual interpretation and explicit uncertainty decomposition.

Evaluation must be read in the same spirit. In problems with clean labels, ranking metrics such as ROC-AUC and PR-AUC remain standard, and the statistical comparison of ROC curves can be performed, for example, with the DeLong test (DeLong et al., 1988). Yet many clinical anomaly detection tasks do not offer unambiguous anomaly labels, so it's important to rely on richer validation strategies, including agreement with accepted biomarkers, subgroup analyses, spatial or temporal plausibility, and robustness to preprocessing or hyperparameter changes.

Finally, it is useful to distinguish between a score and a decision. A score is a quantitative summary of deviation from a learned reference; a decision attaches an operating point and a downstream action to that summary. In some chapters of the thesis, the score is primarily used for visualization and ranking, as in electroanatomical mapping; in others, it is interpreted against clinically established thresholds, as in Z-score-based assessment.

2.2 Anomaly detection methods

The methodological literature can be read as a collection of answers to a single question: how should normality be represented so that deviations from it can be scored? Classical methods answer this through geometry, neighborhoods, density, or support boundaries; modern deep methods typically learn a representation in which those same ideas become more informative.

Distance-based scoring. A first family of methods treats anomalous points as observations that are isolated from their neighbors. Given a distance d on $\mathcal{X}$ and the set $\mathcal{N}_k(x)$ of k nearest neighbors of x, a simple score is the average neighbor distance

$$s_{\mathrm{kNN}}(x) = \frac{1}{k} \sum_{x_j \in \mathcal{N}_k(x)} d(x, x_j). \tag{2.2}$$

Radius-based and robust variants are also common (Chandola et al., 2009). The appeal of this family is its simplicity: no explicit density model is required, and the score is intuitive. Its main weakness is that distance concentration can make neighborhoods uninformative in high dimensions. This limitation is one of the conceptual motivations for learning compact embeddings before applying any distance-based scoring. The method family also depends strongly on feature scaling and on the choice of metric, which means that strong performance in a low-dimensional handcrafted representation does not automatically transfer to raw waveform or image spaces.

Density-based and local-density methods. A second family scores observations by low probability under an estimated distribution, typically through a quantity such as $s(x) = -\log \hat{p}(x)$. Parametric variants include Gaussian mixtures; non-parametric variants include kernel density estimation and histogram-based models (Goldstein and Dengel, 2012). In a Gaussian-mixture model,

$$s_{\mathrm{GMM}}(x) = -\log \left(\sum_{m=1}^{M} \pi_m \, \mathcal{N}(x; \mu_m, \Sigma_m) \right). \tag{2.3}$$

Density modeling can work well when the representation is low-dimensional and well aligned with the true structure of normal data, but it becomes fragile in raw signal spaces. Local-density methods attempt to correct the limitations of global densities by comparing the density around a point to the density around its neighbors. The Local Outlier Factor (LOF) is the most widely used example: it compares the local reachability density of a point to that of its neighborhood, thereby flagging points that are sparse relative to their immediate surroundings (Breunig et al., 2000). In practical terms, LOF is useful when anomalies are local

deviations rather than global extremes, but its behavior depends on neighborhood size and on the metric used to define neighborhoods.

Isolation-based methods. Isolation Forest takes a different route. Instead of estimating density or distances explicitly, it recursively partitions the space and exploits the fact that anomalous points tend to be isolated by fewer random splits than typical points (Liu et al., 2008). If $h(x)$ denotes the path length needed to isolate x across a forest of random trees, then a normalized anomaly score can be written as

$$s_{\mathrm{IF}}(x) = 2^{-\mathbb{E}[h(x)]/c(N)}, \tag{2.4}$$

where $c(N)$ is a normalization constant depending on the sample size. Isolation-based methods are strong practical baselines because they avoid explicit density estimation and often scale well. Their limitation, again, is that raw features may not be the right space in which to isolate the relevant notion of abnormality. In contextual tasks, moreover, applying Isolation Forest directly to concatenated context–measurement vectors may conflate rarity of the context with abnormality of the measurement conditional on that context.

Support estimation and one-class learning. A more principled view is that anomaly detection aims to estimate the support of the nominal distribution (Schölkopf et al., 2001). One-Class SVM learns a boundary that encloses most data points while allowing a fraction of outliers controlled by a parameter ν. In its primal form, it solves

$$\begin{aligned} \min_{w,\rho,\xi} \quad & \frac{1}{2}\|w\|^2 + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i - \rho \\ \text{s.t.} \quad & \langle w, \Phi(x_i)\rangle \geq \rho - \xi_i, \qquad \xi_i \geq 0, \end{aligned} \tag{2.5}$$

where Φ is a feature map. Support Vector Data Description (SVDD) expresses a closely related idea in terms of a minimal-volume hypersphere enclosing nominal data (Tax and Duin, 2004). The resulting anomaly score is the distance of a point from the center relative to the learned radius. These methods are important not only as baselines but also as conceptual ancestors of deep one-class learning, where the feature map is no longer fixed but learned jointly with the scoring objective.

The broader lesson of one-class methods is that anomaly detection is not always about reconstructing the input or estimating its density. Sometimes it is more useful to learn a compact region in representation space and measure how far a new observation lies from it.

Subspace and reconstruction views. Classical anomaly detection also includes a reconstruction perspective. In principal-component analysis, normal data are

assumed to lie near a low-dimensional linear subspace. If $U \in \mathbb{R}^{D \times r}$ contains the principal directions, then the residual

$$s_{\mathrm{PCA}}(x) = \|x - UU^\top x\|^2 \tag{2.6}$$

acts as an anomaly score. This is conceptually important because it shows that reconstruction error is not specific to neural autoencoders. Rather, it is a general scoring primitive: learn a compressed description of typical structure, then measure how much of the input cannot be explained by that description.

Thresholding and operating points. Anomaly detection methods naturally produce rankings, but practical use often requires a threshold. In fully unsupervised settings, thresholds are commonly based on score quantiles or on parametric fits to the tail of the score distribution. In biomedical deployment, however, a single universal threshold is seldom realistic. The cost of false positives and false negatives depends on the clinical action attached to the decision, and the same score may be used either as a map-level visualization, as a triage cue, or as a standardized deviation measure. For this reason, later chapters of the thesis emphasize the interpretation of the score itself and, when possible, uncertainty intervals around it rather than a single frozen cutoff.

From classical methods to deep anomaly detection. Deep anomaly detection can be understood as augmenting the previous families with representation learning (Chalapathy and Chawla, 2019; Pang et al., 2021; Ruff et al., 2021). Instead of applying density, distance, support, or residual calculations directly in the input space, one first learns an encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$ and then constructs the anomaly score in the learned space. This shift is decisive in high-dimensional inputs such as images and biomedical waveforms, because raw Euclidean geometry is often poorly aligned with the structure of normality that matters for the task.

The deep literature is broad, but the principal methodological families that matter for this thesis can be summarized as follows:

- **Reconstruction-based models**, in which anomaly is measured through residual error after compression and reconstruction.
- **Latent-density models**, in which an embedding is learned, and a simple density model is fitted in latent space.
- **Deep one-class methods**, in which the network is trained to map nominal samples to a compact region.
- **Self-supervised representation learning**, in which the representation is learned from surrogate tasks and can later support anomaly scoring.

The following discussion details these families.

Autoencoders and reconstruction residuals. A standard autoencoder learns an encoder f_θ and decoder g_ϕ by minimizing a reconstruction loss,

$$\mathcal{L}_{\mathrm{AE}}(\theta, \phi) = \frac{1}{N} \sum_{i=1}^{N} \|x_i - g_\phi(f_\theta(x_i))\|^2. \tag{2.7}$$

At test time, the reconstruction residual

$$s_{\mathrm{rec}}(x) = \|x - g_\phi(f_\theta(x))\|^2 \tag{2.8}$$

becomes the anomaly score. This family is attractive because the residual often has an immediate visual interpretation: it highlights the parts of the input that the learned model of normality cannot explain well. Yet the literature also documents a well-known limitation: sufficiently expressive decoders may reconstruct anomalous inputs surprisingly well, especially when those anomalies share low-level structure with the nominal data (Chalapathy and Chawla, 2019; Pang et al., 2021). Architectural choices, regularization, and the training distribution matter greatly.

Several variants seek to address these weaknesses. Denoising autoencoders encourage invariance to noise by reconstructing a clean input from a corrupted version; sparse and contractive autoencoders constrain the latent representation; memory-augmented autoencoders introduce an explicit bank of normal prototypes and reconstruct through sparse access to that memory, thereby making unusual patterns harder to reproduce faithfully (Gong et al., 2019). The key conceptual point is the same across these variants: they do not merely compress data; they bias the model toward the recurrent structure of the nominal class, making residuals more useful as abnormality indicators.

Variational autoencoders and probabilistic reconstruction. Variational autoencoders (VAEs) reinterpret autoencoding within a probabilistic latent-variable model (Kingma and Welling, 2014). Instead of a deterministic code, the encoder outputs an approximate posterior $q_\phi(z \mid x)$, while the decoder defines a likelihood $p_\theta(x \mid z)$. Training maximizes the evidence lower bound,

$$\mathrm{ELBO}(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - D_{\mathrm{KL}}\big(q_\phi(z \mid x) \| p(z)\big). \tag{2.9}$$

The probabilistic formulation enables several score choices, including negative ELBO, reconstruction likelihood, or latent posterior discrepancies. VAEs offer a more explicitly probabilistic language than plain autoencoders, but in anomaly detection practice, they still require careful interpretation, since likelihood terms may be influenced by nuisance factors and posterior uncertainty does not automatically coincide with clinically meaningful reliability.

Latent-density and hybrid models. Another influential idea is to combine representation learning with explicit density estimation in the latent space. The Deep Autoencoding Gaussian Mixture Model (DAGMM) is a representative example: it jointly learns an autoencoder and a Gaussian-mixture density model over latent features, then scores anomalies via an energy term that mixes reconstruction and density information (Zong et al., 2018). AnoGAN and related generative-adversarial approaches follow a different route: they learn to generate nominal samples and then define an anomaly score from the discrepancy between a query and its closest generated counterpart, often mixing pixel-space residuals and feature-space mismatch (Schlegl et al., 2017). These models are historically important because they show that reconstruction, density, and representation can be combined in many ways.

Deep one-class objectives. Deep one-class learning replaces the fixed feature map of classical support-estimation methods with a trainable encoder. Deep SVDD is the canonical example: it learns f_θ so that nominal samples map close to a center c in latent space by minimizing

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \|f_\theta(x_i) - c\|^2. \tag{2.10}$$

The anomaly score is then simply the squared distance from that center. This family is appealing because it endows the score with a direct semantic interpretation: anomaly is distance from the learned normal region. In waveform applications, this can be advantageous, because it avoids some of the pathologies of reconstruction-based models while still exploiting the power of deep representation learning (Ruff et al., 2018).

Self-supervised and contrastive representations. Recent representation-learning literature has shown that useful embeddings can be learned from surrogate tasks without explicit anomaly labels. Pretext tasks such as solving jigsaw puzzles, predicting rotations, masked reconstruction, or contrastive discrimination among augmented views have produced features that transfer well across downstream tasks (Noroozi and Favaro, 2016; Gidaris et al., 2018; Chen et al., 2020; He et al., 2020; Grill et al., 2020; He et al., 2022; Radford et al., 2021). From the perspective of anomaly detection, these methods matter because they decouple representation learning from the final score. One may first learn invariances and salient structure through self-supervision, then apply a simpler anomaly detector in the learned space.

A compact way to summarize the methodological landscape is the following. Classical methods supply the basic score semantics: distance, density, isolation,

support, or residual. Deep methods do not replace these semantics so much as relocate them into a learned representation where they become more effective.

2.3 Contextual anomaly detection and uncertainty

Not all clinically relevant abnormality is waveform- or image-based. Many measurements are anomalous only relative to patient context. This is the domain of *contextual anomaly detection*, where each observation is naturally represented as a pair (x, y) consisting of context variables x and a measured behavior or outcome y (Song et al., 2007; Hauskrecht et al., 2007; Valko et al., 2011; Tang et al., 2013). In this setting, the question is not whether y is unusual in marginal terms, but whether it is unusual given the patient profile encoded in x.

The classical statistical prototype of this idea is the Z-score:

$$Z(x, y) = \frac{y - m(x)}{\sigma}. \tag{2.11}$$

Here, y denotes the observed value, $m(x)$ is the expected or reference value associated with the covariates x, and σ is the standard deviation, that is, the measure of variability around the reference. Therefore, $Z(x, y)$ quantifies how many standard deviations the observation lies above or below the expected value. Depending on the reference model, the mean relationship may be oversimplified, variance may be treated as constant even when it depends on context, and the score may be reported as if it were equally trustworthy for well-supported and poorly supported regions of covariate space.

These limitations are well known in the literature on reference standards and growth modeling. Classical reference-interval methodology, pediatric growth curves, and medical nomograms all rest on the idea that measurements should be interpreted relative to age, body size, or developmental stage, but they differ substantially in how they model conditional variability and in how they communicate uncertainty (Goldstein, 1972; Healy, 1978; Royston, 1991; Lee et al., 2010; Ceriotti, 2012; Zierk et al., 2021). This broader literature is relevant to the thesis because it shows that the problem of contextual normalcy is older and richer than the recent machine-learning formulation alone: clinicians have long needed ways to compare an observed measurement with a context-dependent expectation, but the statistical tools used in practice often hide strong assumptions about both mean and variance.

These limitations motivate a more explicit probabilistic perspective. The literature on uncertainty distinguishes between *aleatoric uncertainty*, which reflects irreducible biological variability or measurement noise, and *epistemic uncertainty*, which reflects lack of knowledge due to limited or biased reference data (Hüllermeier and Waegeman, 2021; Kiureghian and Ditlevsen, 2009; Kendall and Gal, 2017;

Zhou et al., 2022). In a clinical reference model, aleatoric uncertainty helps define what range of variation is physiologically plausible, whereas epistemic uncertainty answers a different question: how confident should we be that the estimated reference itself is accurate for a patient with those covariates? This distinction is conceptually central to the methodological contribution of the thesis and to its clinical instantiation.

Gaussian-process regression provides a particularly transparent language for this problem (Rasmussen and Williams, 2006). Without entering into kernel formulas, the key idea is that a Gaussian process yields a posterior predictive distribution whose mean and uncertainty depend on the support provided by nearby reference data. In dense regions of covariate space, predictive uncertainty narrows; in sparse or under-represented regions, it broadens. Heteroscedastic extensions further allow the scale of variability to change with context, while sparse variational approximations make inference practical on larger datasets (Snelson and Ghahramani, 2005; Titsias, 2009; Kersting et al., 2007; Saul et al., 2016).

The broader uncertainty-estimation literature offers alternative strategies. Monte Carlo dropout interprets dropout at test time as an approximate Bayesian inference scheme (Gal and Ghahramani, 2016). Deep ensembles estimate uncertainty through variability across independently trained models (Lakshminarayanan et al., 2017). Bayesian neural networks place distributions over weights and approximate the posterior through variational methods (Blundell et al., 2015). These approaches are influential, particularly in high-dimensional deep learning, but in low- to moderate-dimensional contextual reference modeling, Gaussian-process methods remain especially attractive because they preserve an interpretable link between predictive uncertainty and covariate support.

Alternative contextual baselines are also important. Quantile-regression approaches can model conditional tails without assuming Gaussian residuals (Meinshausen and Ridgeway, 2006); reliable-learning frameworks emphasize abstention or selective prediction in regions where the model is uncertain (Senge et al., 2014); and explainable contextual methods attempt to decompose why a point is abnormal given its context (Li and van Leeuwen, 2023).

2.4 Generative models and source attribution

The rapid diffusion of modern generative models has changed the scope of digital forensics. Earlier work on manipulated media was often concerned with detecting hand-crafted edits or image splicing, whereas current forensic pipelines must increasingly deal with images produced entirely by learned generators. Generative adversarial networks and diffusion models can now synthesize highly realistic content at scale (Goodfellow et al., 2014; Ho et al., 2020; Rombach et al., 2022). This

has made the classical distinction between "real" and "fake" only a first step: in many practical scenarios, it is equally important to determine which generative model most likely produced a suspicious image, or at least to narrow the range of plausible candidate generators.

From a methodological perspective, source attribution is more demanding than standard detection. A binary detector only needs to separate authentic content from synthetic content, whereas attribution requires the system to discriminate among multiple generators that may share similar architectures, training corpora, and visual signatures. The difficulty is compounded by the fact that many contemporary image generators belong to closely related families, especially in the diffusion setting, and therefore leave artifacts that are weaker and less visually obvious than those exploited by earlier forensic detectors. In addition, the number of candidate models is not fixed: new generators appear continuously, existing ones are updated, and closed commercial systems often expose only a black-box interface. This makes source attribution a moving-target problem in which scalability and adaptability are at least as important as raw classification accuracy.

The literature distinguishes between two broad settings:

- **White-box attribution**, where the forensic analyst assumes access to the architecture or weights of the generator. In this setting, attribution may exploit model-specific internals, inversion procedures, or direct comparisons with known generation pipelines (Albright et al., 2019; Hirofumi et al., 2022; Ricker et al., 2024). While this assumption can be useful for controlled experiments, it is often unrealistic in practical scenarios involving proprietary or remotely accessed generators.

- **Black-box attribution**, where only output images are available. This is the more realistic setting for passive forensics, but it is also substantially harder because the attribution system must infer source-specific traces solely from generated samples (Wang et al., 2023; Jeon et al., 2020). In many cases, black-box methods are designed primarily for real-vs-fake detection rather than for multiclass source attribution, and do not directly address the problem considered in the complementary chapter of this thesis.

A second important distinction concerns the data regime. Many recent forensic methods rely on large pretrained backbones and achieve strong performance when hundreds or thousands of examples per class are available. A representative case is CLIP-based detection (Radford et al., 2021), which motivates approaches such as DE-FAKE (Sha et al., 2023). These methods benefit from large-scale vision-language pretraining and can be highly effective in standard supervised settings. However, they are not naturally optimized for scenarios in which a new generator becomes available with only a handful of samples, or in which the set of candidate generators

evolves over time. In such conditions, the computational footprint of the detector, the number of trainable parameters, and the need to retrain the whole system become critical practical constraints rather than secondary implementation details.

These considerations connect source attribution with the broader themes of this thesis. In the cardiology chapters, the central question is how to derive reliable scores from scarce or weakly labeled data; in the complementary chapter, the question becomes how to discriminate among generators when only a few examples per class are available. Although the downstream task is different, the methodological pressure is similar: models should remain informative under low-data conditions, and their outputs should be based on representations that are stable enough to generalize beyond the training examples. This explains the appeal of compact autoencoder-based strategies. Autoencoders have long been used for unsupervised representation learning and dimensionality reduction (Hinton and Salakhutdinov, 2006; Vincent et al., 2008; He et al., 2022), and reconstruction error has repeatedly proved useful as a score in anomaly detection and related tasks (An and Cho, 2015; Zhou and Paffenroth, 2017). More broadly, reconstruction-based learning occupies an intermediate position between classical feature engineering and fully end-to-end discriminative models: it preserves an interpretable notion of fit to a learned manifold while remaining lightweight enough to be adapted in resource-constrained settings.

This is especially relevant in few-shot and class-incremental forensic scenarios. When a new generator appears, retraining a large detector from scratch may be computationally expensive and operationally slow. For this reason, recent forensic research has also explored continual-learning strategies for synthetic-media detection (Marra et al., 2019; Magistri et al., 2023, 2024; Tassone et al., 2024). These works address an important problem, namely, how to keep forensic systems updated as generators evolve. At the same time, they are usually studied in regimes with relatively abundant data and often involve a trade-off between adaptation and performance retention. Large-scale continual-learning benchmarks such as those considered in (Pan et al., 2023; Gao et al., 2024) are valuable from a systems perspective, but they do not directly solve the few-shot attribution problem in which only a very limited number of source examples are available.

The contribution of the complementary chapter on attribution is not to propose yet another large-scale detector, but to examine whether a bank of compact reconstruction models can serve as a modular attribution mechanism when data are scarce, and new classes need to be incorporated quickly. In this sense, the chapter extends the reconstruction-based viewpoint developed elsewhere in the thesis into a different application domain. The target task is no longer anomaly detection in a biomedical setting, but the underlying methodological intuition is closely related: if a model can learn a compact representation of what is typical

for a source, then deviations in reconstruction behavior may become informative signals for downstream decision making.

2.5 Medical background: atrium and aorta

The cardiology applications developed in this thesis require a minimal clinical background on two measurement settings: intracardiac electrograms for atrial substrate mapping, and echocardiographic measurements for aortic-diameter assessment. More generally, biomedical anomaly detection poses challenges that standard benchmarks often understate. Signals are noisy, non-stationary, and affected by acquisition artifacts; inter-patient variability is often substantial; and the boundary between pathology and normal variation is frequently gradual rather than categorical. As a result, the validity of an anomaly score cannot be judged solely by benchmark metrics. In biomedical applications, preprocessing, representation, and evaluation protocols are integral parts of the methodological design.

Several recurrent issues motivate caution:

- **Noise and artifacts** can dominate residual-based scores if the learned representation is too sensitive to nuisance fluctuations.
- **Segmentation and alignment** determine what the model sees as a meaningful unit of analysis.
- **Inter-subject heterogeneity** can make a method confuse legitimate physiological diversity with abnormality.
- **Label ambiguity** often makes direct supervised learning less informative than expected, even when some labels are available.

These issues are especially relevant for cardiac signals, where conduction, activation sequence, electrode placement, and preprocessing all affect waveform appearance.

Intracardiac electrograms acquired during mapping procedures for atrial fibrillation are a particularly demanding case. EGMs reflect local electrophysiological properties of atrial tissue, but their morphology depends not only on local tissue properties, but also on electrode spacing, catheter orientation, contact force, wavefront direction, filtering, and rhythm (Anter and Josephson, 2016; Sim et al., 2019; Wong et al., 2019). Clinically, the arrhythmogenic substrate is often characterized through handcrafted biomarkers such as bipolar voltage, electrogram duration, and measures of fractionation or deflection complexity (Konings et al., 1994; Nademanee et al., 2004; Jadidi et al., 2012; Frontera et al., 2021, 2022b). These descriptors are meaningful and widely used, but none of them is a complete representation

of morphology, and their interpretation may vary across acquisition settings and anatomical locations (La Rosa et al., 2021; Williams et al., 2021).

This is precisely the setting in which unsupervised waveform-level scoring becomes attractive. Rather than choosing a small set of handcrafted descriptors and then defining hard pathological classes, one may attempt to learn a representation directly from EGM waveforms and derive from it a morphology-oriented anomaly score. The resulting score is not meant to replace established biomarkers, but to synthesize aspects of waveform irregularity in a way that can then be compared with voltage, fractionation, and duration maps.

A short clinical remark helps frame the importance of this application. Catheter ablation has become a central rhythm-control strategy in symptomatic atrial fibrillation, and pulmonary vein isolation remains the standard lesion set for first procedures, especially in paroxysmal AF. Yet recurrence after ablation remains common in persistent forms of the disease, where arrhythmogenic mechanisms are more diffuse, and the substrate extends beyond pulmonary vein triggers. This is why substrate mapping continues to play a major role in translational electrophysiology: clinicians are not only interested in whether arrhythmia can be terminated, but also in whether local electrogram morphology can reveal regions of slow conduction, conduction block, or structural remodeling that deserve further interpretation (Nademanee et al., 2004; Jadidi et al., 2012; Frontera et al., 2022b). In this setting, a score that is stable, morphology-aware, and comparable across regions of the atrium is clinically attractive because it can complement standard voltage-based maps rather than compete with them.

The cardiology literature on aortic diameters provides a clear example of the importance of having a reliable score. Several studies have established normal values and reference equations for the aortic root and ascending aorta across populations of different ages and body sizes (Campens et al., 2014; Frasconi et al., 2021; Patel et al., 2022a; Vriz et al., 2014; Cantinotti et al., 2017). However, different cohorts and equations may produce different classifications, particularly for borderline cases or for patients whose characteristics are rare in the reference sample (Curtis et al., 2016; Dallaire et al., 2015; van Kimmenade et al., 2013). The classical Z-score remains useful as a simple standardized index, but it does not by itself tell the clinician whether the patient lies in a region where the reference equation is extrapolating. Aortic assessment is also a good example of why contextual modeling is unavoidable in clinical measurement: the same raw diameter can carry very different implications in a young child, a small adult woman, or a tall adult man, and the uncertainty of the reference model becomes most consequential precisely in those borderline cases where clinical follow-up, additional imaging, or longitudinal surveillance may depend on the interpretation of a single standardized value. For this reason, the problem is not only to estimate an expected diameter curve, but also to quantify

how trustworthy that reference is in sparsely populated regions of the covariate space (Curtis et al., 2016; Dallaire et al., 2015).

Chapter 3

Deep anomaly detection for intracardiac electrograms

3.1 Motivation

Atrial fibrillation (AF) is the most common sustained arrhythmia, and its prevalence increases markedly with age, making it a major clinical and socioeconomic burden. Beyond the direct symptoms caused by the arrhythmia itself, AF is associated with an increased risk of stroke, heart failure, hospitalization, and reduced quality of life. For symptomatic patients, catheter ablation (CA) has become a cornerstone rhythm-control therapy, either as a first-line strategy in selected cases or after failure or intolerance of antiarrhythmic drug therapy. In current clinical practice, pulmonary vein isolation (PVI) represents the standard lesion set for initial procedures, particularly in paroxysmal AF, because ectopic triggers originating from the pulmonary veins are well-established initiators of atrial arrhythmias. However, although PVI is highly effective in many patients, success rates remain more limited in advanced forms of the disease, especially in persistent AF, where the sustaining substrate is often more complex and extends beyond pulmonary vein triggers.

This limitation has motivated a long-standing search for patient-specific substrate markers capable of identifying regions that contribute to arrhythmia maintenance and recurrence. Several complementary strategies have been explored for this purpose, including invasive mapping during sinus rhythm or fibrillation, imaging-based assessment of atrial remodeling, and computational modeling of patient-specific conduction patterns (Konings et al., 1994; Rodrigo et al., 2014; Zahid et al., 2016; Boyle et al., 2019). Among these options, high-density electroanatomical mapping remains particularly attractive because it provides direct access to local electrical activity with high spatial resolution and can be integrated naturally into routine ablation workflows. During a mapping procedure, thousands of intracardiac electrograms (EGMs) can be collected across the atrial surface, thereby

offering a rich and spatially resolved description of conduction properties. The challenge, however, is that this wealth of information is difficult to summarize in a way that is both physiologically meaningful and robust across patients, rhythms, and acquisition conditions.

Machine learning techniques have also been explored in atrial fibrillation beyond procedures that directly target the substrate through catheter ablation. For instance, Bernardini et al. (2024) employed supervised models to predict clinical outcomes in anticoagulated AF patients, showing that data-driven methods can complement guideline-based scores in risk stratification and patient management. These studies highlight the broader potential of artificial intelligence in arrhythmia care, but they also underscore a key methodological gap: when the goal is substrate characterization, predictive performance alone is not sufficient, and there remains a strong need for descriptors that preserve a clear link with the underlying electrophysiology. This is one of the reasons why morphology-oriented and interpretable representations of intracardiac signals are particularly valuable in the context of electroanatomical mapping.

In sinus rhythm (SR), several electrophysiological indicators have been proposed to characterize arrhythmic substrate, such as bipolar voltage, the number of deflections (fractionation), and the duration of the fragmented component (Anter and Josephson, 2016; Jadidi et al., 2012; Frontera et al., 2021). These biomarkers are intuitively appealing because they are simple to compute and can be visualized directly on electroanatomical maps. Nevertheless, each of them captures only one facet of a much more complex morphology. Voltage reflects signal amplitude but is influenced by catheter orientation, contact quality, inter-electrode spacing, and wavefront direction (Anter and Josephson, 2016; Sim et al., 2019; Wong et al., 2019). Fractionation is related to signal complexity, but similar patterns may arise from distinct electrophysiological phenomena (Konings et al., 1994; Lau et al., 2015). Duration is linked to delayed or heterogeneous conduction, but it may not, in isolation, uniquely identify the pathological substrate. Despite efforts to design more specific descriptors, such as amplitude-normalized area (Mendonca Costa et al., 2020), there is still no consensus on the most effective way to synthesize EGM morphology into a single informative marker.

This lack of consensus is not only a technical issue but also a conceptual one. Atrial substrate is heterogeneous, and abnormal conduction does not necessarily manifest through a single stereotyped pattern. Signals recorded in low-voltage areas may correspond to diseased tissue, but complex morphologies can also appear in regions with preserved amplitude because of local conduction slowing, anisotropy, or wavefront collision. Conversely, not all fractionated or prolonged signals should be interpreted as direct evidence of fibrosis or as targets suitable for ablation (La Rosa et al., 2021). In other words, the mapping problem is intrin-

sically multivariate: clinically relevant information is distributed across several morphological dimensions, and reducing it to one handcrafted descriptor at a time may lead to unstable or oversimplified interpretations. In practice, this means that electrophysiologists often need to compare several maps simultaneously and to combine partial indicators that may highlight overlapping, but not identical, substrate regions.

These observations motivate the search for approaches that can learn morphology directly from data rather than relying exclusively on predefined features. Deep learning is particularly appealing in this setting because it can process high-dimensional waveforms and infer compact representations of their structure without requiring manual feature engineering. Within this broader family of methods, unsupervised deep anomaly detection is especially attractive for atrial EGMs: instead of requiring exhaustive labels for all possible abnormal morphologies, it learns the dominant structure of the available signals and assigns each waveform a score reflecting its deviation from that learned support (Schölkopf et al., 2001; Markou and Singh, 2003). In practice, this means that signals with uncommon or highly irregular morphology can be highlighted automatically, potentially revealing regions associated with abnormal conduction or structural remodeling. This is particularly valuable in electrophysiology, where labels are difficult to define unambiguously and where the clinical notion of abnormality often depends on a continuum rather than on a clean categorical boundary. Unlike supervised approaches, which require predefined class labels and carefully curated annotations (Liao et al., 2021), anomaly detection can instead exploit the natural structure of the acquired waveforms and identify deviations in a more exploratory and morphology-oriented manner.

Motivated by these considerations, this chapter investigates the application of state-of-the-art deep anomaly detection algorithms as an all-in-one tool for characterizing EGM morphology in atrial substrate mapping. We show that the resulting anomaly scores correlate strongly with traditional indicators, while also providing a single, threshold-free, and morphology-oriented metric that can support a more synthetic and potentially more robust interpretation of electroanatomical maps.

Voltage

The peak-to-peak voltage measures the maximum difference between the positive and negative peaks of a bipolar EGM. Clinically, voltages greater than 1mV are considered to reflect healthy tissue, values between 0.5 and 1.0mV are attributed to a border zone, and voltages below 0.5mV suggest low-voltage regions or scar. There is no universally accepted gold standard for voltage thresholds, and this biomarker has low specificity because it depends on many factors such as electrode spacing, contact force, catheter orientation, and stimulation rate (Wong et al., 2019; Sim

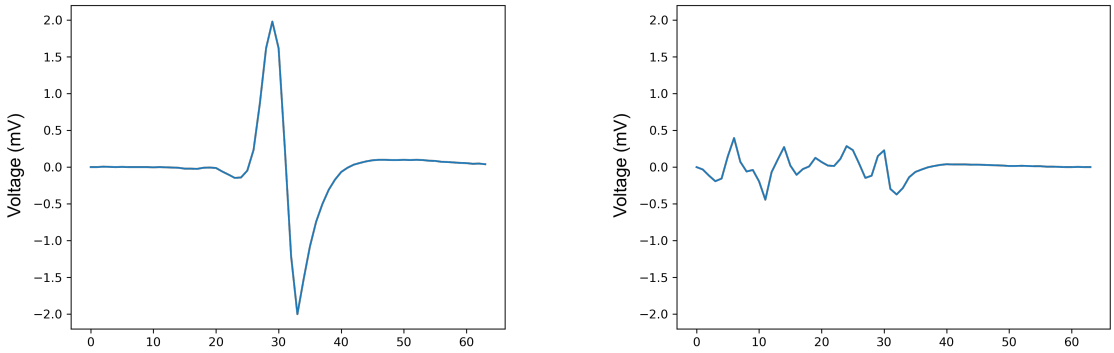
et al., 2019; La Rosa et al., 2021). Moreover, abnormal conduction can occur even in high-voltage areas due to wavefront collisions and slow conduction corridors (Frontera et al., 2022a).

Fractionation index

A physiological EGM recorded in SR typically exhibits one main deflection followed by a single negative peak. In contrast, fractionated EGMs contain multiple deflections, reflecting complex local activation patterns. Fractionation has long been associated with arrhythmic substrate, and early studies during AF proposed complex fractionated atrial electrograms (CFAEs) as ablation targets (Konings et al., 1994; Nademanee et al., 2004). However, CFAEs are highly dependent on the rhythm at the time of recording, and subsequent trials questioned their therapeutic value (Lau et al., 2015; Vogler et al., 2015). During SR, fractionation often arises from functional phenomena such as wavefront collisions or conduction delay near slow conducting areas (Jadidi et al., 2012; Frontera et al., 2022a). A simple fractionation index F can be defined from the signal $s(t)$ as follows (Williams et al., 2021):

$$F = \left| \left\{ t \in \mathcal{P} : s^2(t) > \tfrac{1}{3} M \right\} \right|, \quad M = \max_t s^2(t), \tag{3.1}$$

where $\mathcal{P}$ denotes the set of peak positions in $s(t)$. In healthy tissue, one typically observes at most three peaks ($F \leq 3$), whereas higher values indicate fractionation. This definition depends on a threshold relative to the maximal energy M and thus may not capture all forms of complex morphology; nevertheless, it provides a quantitative baseline for comparison.



(a) Physiological EGM

(b) Fractionated EGM

Figure 3.1: Comparison between a physiological EGM and a fractionated EGM recorded during sinus rhythm. Fractionation manifests as multiple deflections and a longer duration.

Duration index

Low amplitude yet long-duration EGMs are often associated with slow conduction corridors and fibrotic tissue; these signals may mask localized re-entrant drivers (Frontera et al., 2019; Pagani et al., 2021; Rossi et al., 2021). A duration index D can be defined analogously to fractionation by measuring the proportion of samples where the squared signal exceeds a fraction of its maximum amplitude:

$$D = \left|\{t : s^2(t) > \tfrac{1}{10}M\}\right|, \tag{3.2}$$

with M as in (3.1). Higher values of D suggest prolonged activation and may reflect slow conduction or conduction block.

Anomaly detection

The goal of anomaly detection is to estimate the support of a probability distribution given a set of samples assumed to be normal. A deep anomaly detector learns a representation of the data and outputs a score $s(x)$ that is large when the point x has a low probability of belonging to the distribution (Schölkopf et al., 2001; Chandola et al., 2009). Unlike density estimation, anomaly detection focuses on the boundary of the distribution and is more tractable in high dimensions. In this study, we employ three deep anomaly detection algorithms that differ in how they compute the anomaly score: (i) reconstruction error of an autoencoder, (ii) reconstruction error of a memory-augmented autoencoder, and (iii) distance to the center of a learned hypersphere in a one-class classification model.

3.2 Methodology

Study design and data acquisition

The study included eight patients (see Table 3.1) referred for CA due to symptomatic paroxysmal or persistent AF. Inclusion criteria followed the latest ESC guidelines (Hindricks et al., 2021). Exclusion criteria were previous ablation, structural heart disease, or the presence of left atrial thrombus. For each patient, a high-density sinus rhythm map of the left atrium was acquired using the CARTO3 mapping system (Biosense Webster), with a PentaRay catheter providing 20 electrodes. Maps consisted of at least 3000 EGMs distributed over the atrial surface. Catheter contact was assessed using the Tissue Proximity Index to avoid misleading voltage values. Signals were bandpass filtered between 30 and 300Hz before export.

Table 3.1: Baseline characteristics of the study population (N=8).

Age (years)	65.5 ± 7.7
Male, n (%)	5 (62.5%)
Paroxysmal AF, n (%)	3 (37.5%)
Persistent AF, n (%)	5 (62.5%)
LVEF (%)	55.6 ± 6.8
LA area (cm^2)	25.6 ± 3.7
Dyslipidemia, n (%)	7 (87.5%)
Hypertension, n (%)	5 (62.5%)
Diabetes, n (%)	1 (12.5%)
Mild mitral regurgitation, n (%)	6 (75%)

Preprocessing

Unipolar and bipolar EGMs were annotated by centering each bipolar EGM around the time of maximal negative slope of the corresponding unipolar signal, estimated via finite differences:

$$t_0 = \operatorname*{argmin}_t \frac{u(t+1) - u(t-1)}{2\Delta t}, \tag{3.3}$$

where $u(t)$ is the unipolar EGM and Δt is the sampling period ($\Delta t = 1\,\text{ms}$). A window of 64ms centered at t_0 was extracted to capture the morphology of the bipolar signal (see Figure 3.2). Signals were smoothed by convolving with a Gaussian kernel ($\sigma = 0.8$) to attenuate acquisition noise and then normalized to lie in $[0, 1]$ in order to remove amplitude scaling differences across patients and regions. Figure 3.2 shows an example of a processed EGM.

Deep anomaly detection models

Variational autoencoder (VAE). A VAE is a generative model parameterized by an encoder $q(z \mid x)$ and a decoder $p(x \mid z)$, where z denotes a latent variable drawn from a prior $p(z) = \mathcal{N}(0, I)$ and x denotes the input signal. The model is trained by maximizing the evidence lower bound (ELBO)

$$\mathcal{L}(x; \theta, \phi) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - D_{\mathrm{KL}}\big(q_\phi(z \mid x) \| p(z)\big), \tag{3.4}$$

which encourages the approximate posterior $q_\phi(z \mid x)$ to match the prior while reconstructing x via $p_\theta(x \mid z)$ (Kingma and Welling, 2014).

Memory-augmented autoencoder (MemAE). MemAE augments the autoencoder architecture with an external memory $M = [m_1, \ldots, m_N]$ of learnable basis vectors

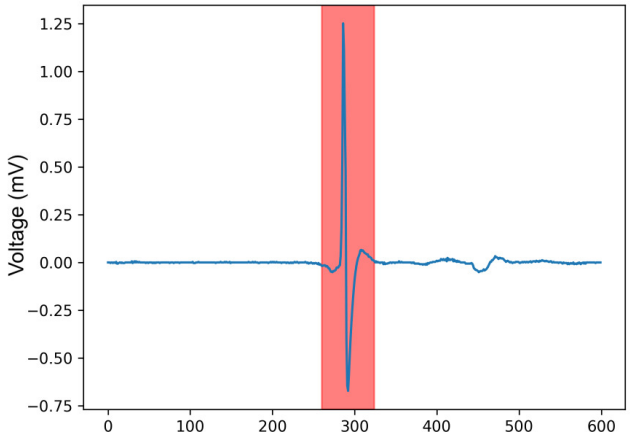


Figure 3.2: Region of interest of a bipolar EGM after alignment and smoothing. A 64 ms window centered on the steepest unipolar slope is extracted to capture the morphology.

(Gong et al., 2019). The encoder produces an embedding $z = f_e(x)$ that is used to query the memory via a soft addressing vector

$$\alpha_i = \frac{\exp(z^\top m_i)}{\sum_{j=1}^{N} \exp(z^\top m_j)}, \quad i = 1, \ldots, N, \tag{3.5}$$

followed by hard shrinkage

$$\widehat{\alpha}_i = \frac{\max(\alpha_i - \lambda, 0) \cdot \alpha_i}{|\alpha_i - \lambda| + \epsilon}, \tag{3.6}$$

where $\lambda \in [1/N, 3/N]$ controls sparsity. The shrunk addressing vector selects a small subset of memory elements whose linear combination yields the final embedding $\widehat{z}$. The decoder reconstructs the input via $\widehat{x} = f_d(\widehat{z})$. The anomaly score is the squared reconstruction error $s(x) = \|x - \widehat{x}\|^2$. By leveraging a learned dictionary, MemAE can adaptively reconstruct only normal patterns and highlight deviations.

Deep Support Vector Data Description (Deep-SVDD). Deep-SVDD extends the classical one-class support vector data description (Tax and Duin, 2004) to deep representations (Ruff et al., 2018). It learns a feature map $\phi(x; w)$ via a neural network and seeks the smallest hypersphere of radius R and center c that encloses most training data:

$$\min_{R,c,w} \; R^2 + \frac{1}{\nu n} \sum_{i=1}^{n} \xi_i \quad \text{subject to} \quad \|\phi(x_i; w) - c\|^2 \le R^2 + \xi_i, \; \xi_i \ge 0, \tag{3.7}$$

where $\nu \in (0, 1)$ controls the trade-off between sphere volume and margin violations. Once trained, the anomaly score for a new x is its squared distance from the center: $s(x) = \|\phi(x; w) - c\|^2$. In contrast to reconstruction-based methods, Deep-SVDD employs a discriminative criterion and captures global signal features.

Training procedure

All models were implemented using one-dimensional convolutional networks with six convolutional layers in the encoder and six upsampling layers in the decoder (for autoencoder variants). The latent space dimension was set to 16. Hyperparameters such as latent size, kernel size (2-5) and number of filters (8-32) were varied in preliminary experiments; results were found to be highly correlated across architectures (weighted Kendall's $\tau > 0.8$), indicating robustness to architecture choices. Models were trained with the Adam optimizer (learning rate 5×10^{-3}) for a maximum of 100 epochs, using early stopping with patience 3 on a 20% validation subset. To evaluate generalization across patients, a leave-one-patient-out cross-validation scheme was employed: at each iteration, seven patients were used for training and one for testing. Inference times on an NVIDIA RTX A6000 GPU were approximately 2ms per batch of 512 signals; on CPU, inference took about 10ms per batch.

Weighted ranking correlation

To quantify the agreement between different anomaly detectors and traditional indicators, we used the weighted Kendall's τ_w statistic (Shieh, 1998). Given two score vectors R and S, the weighted inner product is

$$\langle R, S \rangle_w = \sum_{i<j} w(i, j)\ \mathrm{sgn}(R_i - R_j)\ \mathrm{sgn}(S_i - S_j), \tag{3.8}$$

where the weights $w(i, j)$ emphasize pairs with high scores via

$$w(i, j) = \frac{1}{\rho(R_i)} + \frac{1}{\rho(S_j)}, \tag{3.9}$$

and ρ returns the rank of its argument. The weighted Kendall coefficient is

$$\tau_w(R, S) = \frac{\langle R, S \rangle_w}{\|R\|_w \|S\|_w}, \tag{3.10}$$

which ranges between -1 and 1, with larger values indicating stronger agreement for high-ranked entries. This statistic is appropriate here because we care more about how methods rank anomalous signals than normal ones.

3.3 Experimental evaluation

Correlation with traditional indicators

Figure 3.3 reports the pairwise weighted Kendall coefficients between the anomaly scores produced by VAE, MemAE, and Deep-SVDD and the traditional indicators (voltage, fractionation and duration). Almost all coefficients exceeded 0.7 (unweighted $p < 10^{-10}$), indicating a strong correlation among deep models and between anomaly scores and standard biomarkers. Although Deep-SVDD yielded slightly higher correlations with fractionation and duration, reconstruction-based methods (VAE and MemAE) exhibited comparable trends. These results confirm that deep anomaly detection captures largely the same patterns as conventional indicators, but with the advantage of a single unified score.

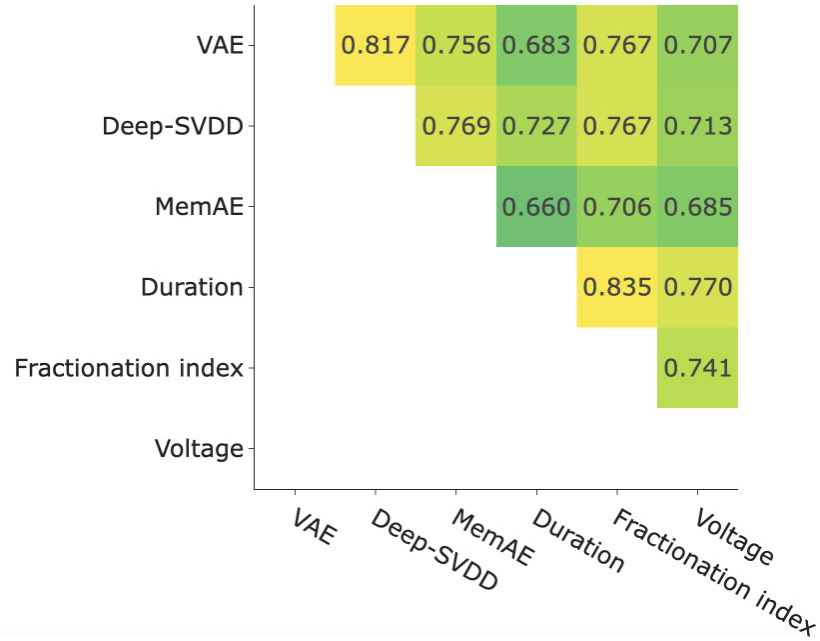
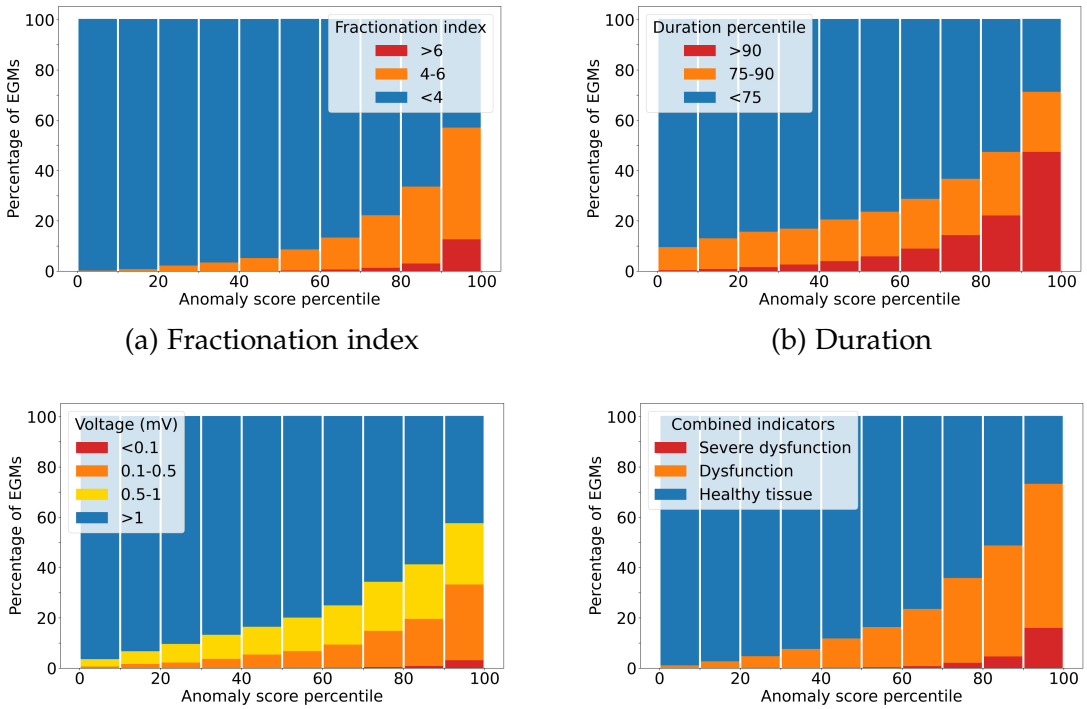


Figure 3.3: Matrix of weighted Kendall coefficients τ_w computed over all eight patients. The upper-left triangle shows agreements among deep anomaly detectors; the lower-right triangle shows agreements among traditional indicators; and the upper-right block shows cross-correlations between anomaly scores and indicators.

Thresholding based on anomaly percentiles

One difficulty in using multiple indicators is deciding threshold values for each and how to combine them. In contrast, anomaly scores can be thresholded by percentiles without calibrating individual scales. For a given percentile q, we computed the mean indicator value over the EGMs whose anomaly score exceeds the qth

percentile. The average fractionation and duration increased monotonically with q, while the average voltage decreased, as shown in Figure 3.5. Histograms of indicator distributions for different anomaly percentiles (see Figure 3.4) further illustrate that the fraction of EGMs with high fractionation or long duration grows with the anomaly score. Voltage alone had weaker discriminative power, highlighting the benefit of incorporating morphology through anomaly detection.



(a) Fractionation index

(b) Duration

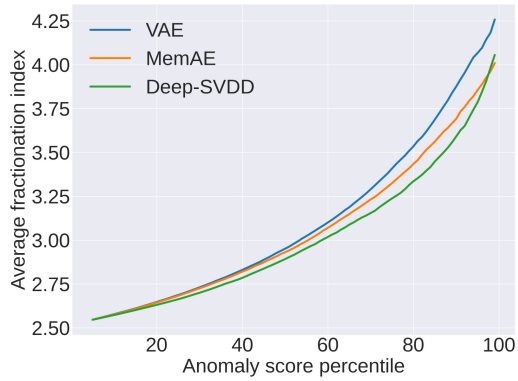
(c) Voltage

(d) Combined indicators

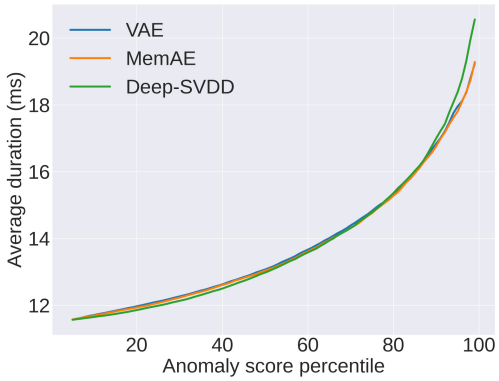
Figure 3.4: Distribution of fractionation index, signal duration, voltage categories, and combined indicator thresholds as functions of the anomaly score percentile. The fractionation index was divided into three ranges: <4, 4-6, >6; duration was divided according to its percentiles into three ranges: <75th percentile, 75th-90th percentile, and >90th percentile; the voltage was divided into four common categories. Figure 3.4(d) combines the indicators. A fractionation index $\geq$4, >90th percentile of duration and a voltage <0.5mV are considered as thresholds where red (severe dysfunction) represents a signal that has all of them, orange (dysfunction) one or two and blue (healthy tissue) none of the three.

Analysis of low-voltage EGMs

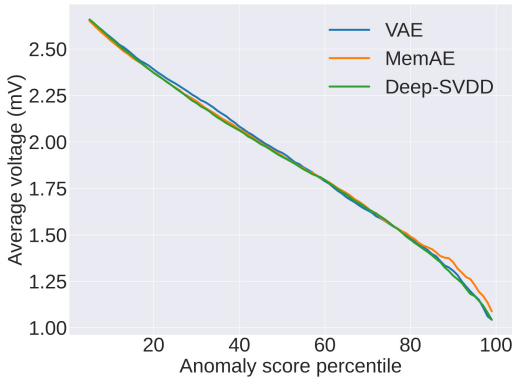
Low-voltage EGMs are often considered markers of fibrotic or scar tissue. We normalized signals before inputting them into the deep models, so the anomaly detectors relied solely on morphology. Most low-voltage signals had high anomaly



(a) Fractionation vs. anomaly percentile



(b) Duration vs. anomaly percentile



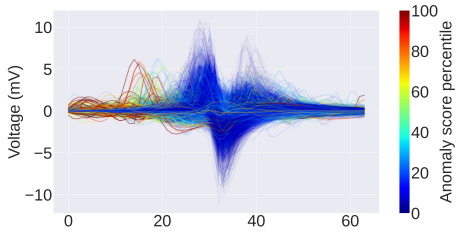
(c) Voltage vs. anomaly percentile

Figure 3.5: Average values of traditional indicators as a function of the anomaly score percentile. Fractionation and duration increase monotonically with the anomaly percentile, while voltage decreases.

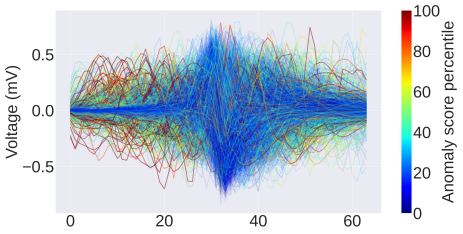
scores, but a non-negligible fraction of border-zone (0.5-1mV) and low-voltage (0.1-0.5mV) EGMs exhibited low anomaly scores and benign morphology (Figure 3.6). This suggests that voltage reduction can stem from catheter positioning or wavefront orientation rather than pathologic substrate. Conversely, some high-voltage EGMs were classified as anomalous due to subtle morphological features (Figure 3.7), potentially corresponding to pivot points or other conduction phenomena. Thus, anomaly detection is more robust to measurement artefacts and can identify morphological abnormalities even when the voltage remains relatively high.

Electroanatomical mapping

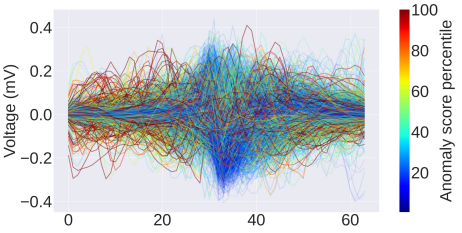
By interpolating anomaly scores onto the three-dimensional atrial mesh, we obtained electroanatomical maps highlighting regions of abnormal electrical behavior



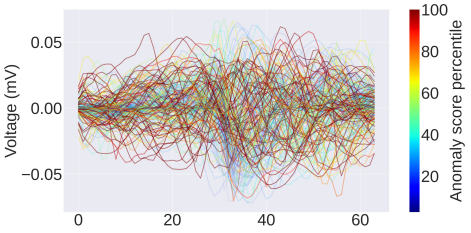
(a) > 1 mV



(b) 0.5-1 mV

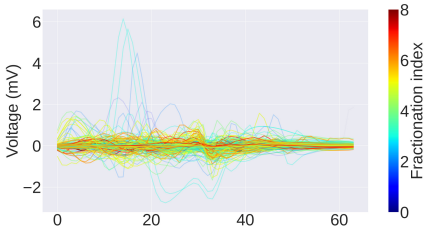


(c) 0.1-0.5 mV

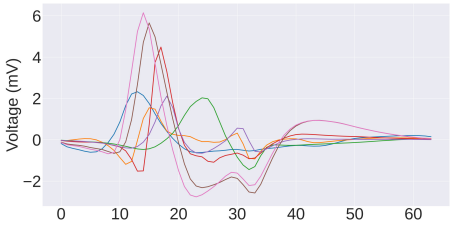


(d) < 0.1 mV

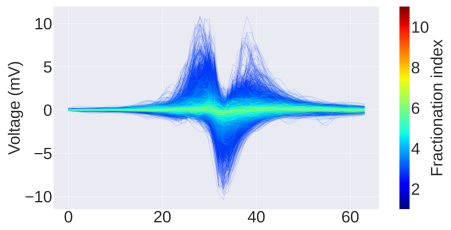
Figure 3.6: Examples of EGMs from all patients colored by Deep-SVDD anomaly score percentile for different voltage ranges. Note the morphological similarity among signals with low anomaly in the border-zone and low-voltage ranges, indicating that voltage alone may be misleading.



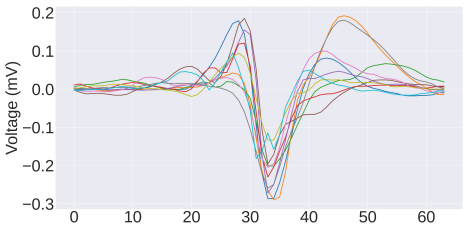
(a) High anomaly scores



(b) low fractionation, high voltage



(c) Low anomaly scores



(d) high fractionation, low voltage

Figure 3.7: EGMs with extreme anomaly scores. Top: signals with anomaly scores above the 95th percentile, colored by fractionation index, with the right panel highlighting high-voltage, low-fractionation signals. Bottom: signals with anomaly scores below the 50th percentile, with the right panel highlighting low-voltage, high-fractionation signals. These examples illustrate how anomaly detection can uncover subtle morphological patterns beyond simple voltage thresholds.

(see Figures 3.8, 3.9). For visualization, scores were normalized by the 90th percentile within each patient. Deep models produced clear, spatially coherent regions with elevated scores. Compared to traditional maps (voltage, fractionation, and duration), anomaly maps condensed the information into a single layer and were less susceptible to low-voltage noise. Among the deep methods, Deep-SVDD yielded slightly sharper delineations, possibly due to its global feature representation. These maps may assist electrophysiologists in targeting ablation more accurately and efficiently.

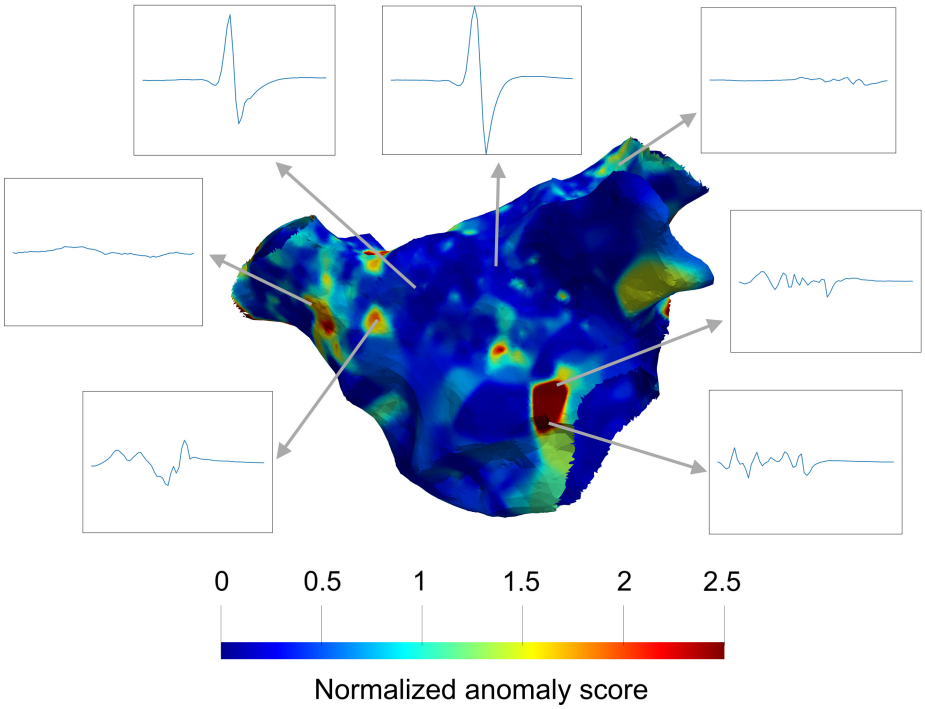
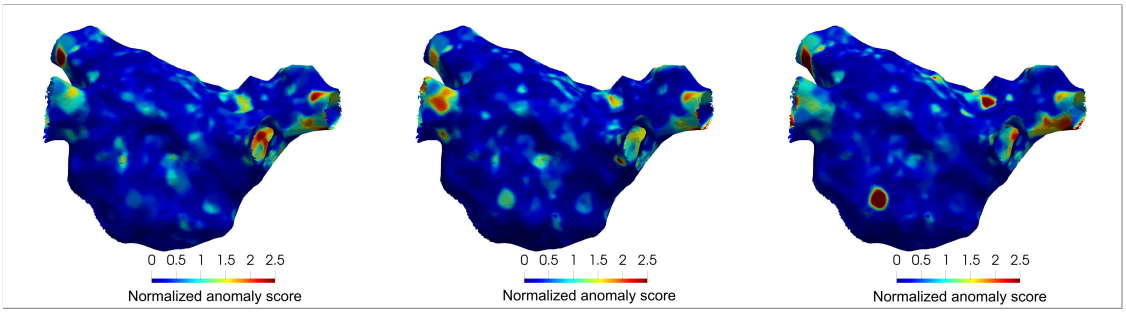


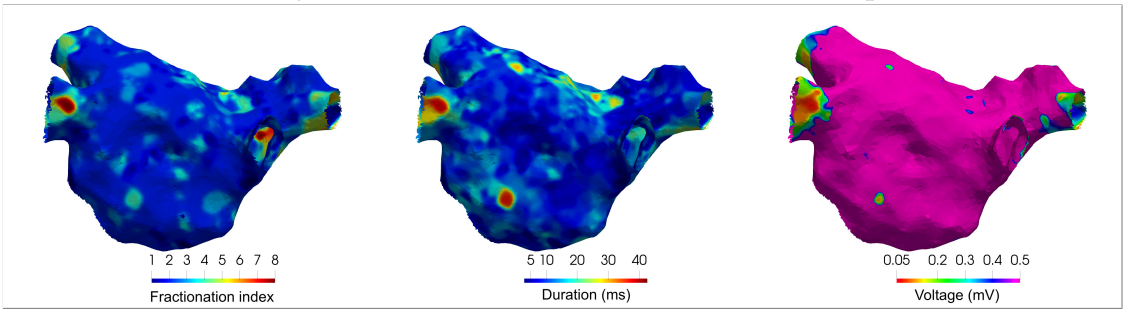
Figure 3.8: Example electroanatomical map of the left atrium for a single patient, colored according to the Deep-SVDD anomaly score. Values are normalized by the 90th percentile of anomaly scores for that patient. Warmer colors indicate regions with higher anomaly scores, corresponding to potential abnormal substrate.

3.4 Discussion

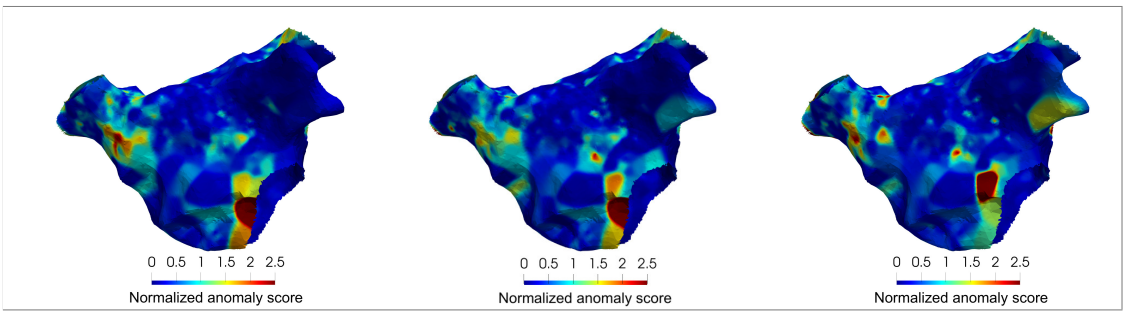
The results presented in this chapter demonstrate that unsupervised deep anomaly detection can characterize intracardiac EGMs in a way that aligns with, and in some cases improves upon, traditional voltage, fractionation, and duration measures. Three distinct models (VAE, MemAE, and Deep-SVDD) yield highly correlated rankings of signals and exhibit strong agreement with conventional biomarkers. Unlike standard indicators that rely on arbitrary amplitude thresholds and are sensitive



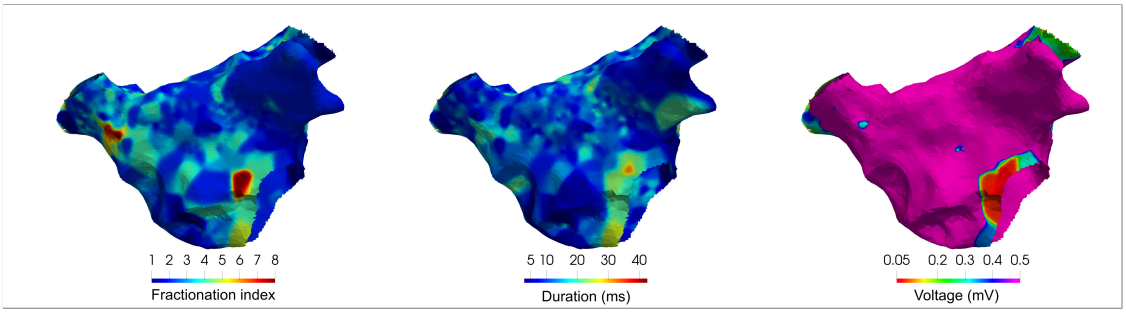
(a) Anomaly detectors (front view): VAE, MemAE, Deep-SVDD



(b) Traditional indicators (front view): number of peaks (fractionation), duration, and voltage



(c) Anomaly detectors (back view)



(d) Traditional indicators (back view)

Figure 3.9: Front and back views of the left atrium of a single patient colored according to deep anomaly scores (VAE, MemAE, Deep-SVDD) and traditional indicators (number of peaks, duration, voltage). Anomaly maps condense information from multiple indicators into a single map and appear more specific and robust to noise.

to electrode orientation and contact, anomaly scores offer a unified, threshold-free assessment of abnormal morphology that is robust across patients. The networks used are lightweight ($\approx 150\,\mathrm{k}$ parameters), train quickly, and generalize well in leave-one-patient-out evaluation, making them suitable for integration into clinical mapping workflows.

From a practical standpoint, these findings suggest that anomaly scores could serve as an all-in-one surrogate marker of atrial substrate abnormalities. The monotonic relationship between anomaly percentiles and traditional indicators means that clinicians could employ percentile thresholds to delineate candidate ablation targets without tuning multiple parameters. In particular, the identification of high anomaly scores in signals with otherwise normal voltage may uncover slow conduction corridors or pivot points missed by amplitude-based criteria. Conversely, the observation of low anomaly scores in border-zone or low-voltage EGMs cautions against over-interpreting voltage reduction as evidence of fibrotic tissue. Electroanatomical maps colored by anomaly scores provide intuitive visual summaries of substrate complexity and could inform personalized ablation strategies.

These benefits should be weighed against the limitations of the present study. The dataset comprised only eight patients, and there is no histological ground truth or long-term follow-up to determine whether high anomaly scores correspond to arrhythmic drivers. Models were trained exclusively on sinus rhythm recordings, leaving their behavior during AF or other rhythms untested. Future research should include larger, multi-center cohorts, integrate imaging and mechanistic modeling, and evaluate whether targeting high-anomaly regions improves clinical outcomes. Extending the unsupervised framework to other cardiac signals and modalities may further strengthen its clinical utility.

Chapter 4

Normalcy score for contextual anomaly detection

4.1 Motivation and problem setting

This chapter develops the *normalcy score* (NS), a probabilistic framework for contextual anomaly detection. The central problem is simple to state but subtle to solve: given an observation described by a set of contextual variables and a behavioral variable, when should that behavioral value be regarded as anomalous for that specific context? In many scientific and clinical applications, this question cannot be answered by looking at the joint rarity of all variables. Contextual factors may be uncommon, but uncommon does not necessarily mean abnormal. A correct assessment requires distinguishing between an unusual context and an unusual behavior conditional on that context.

Formally, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a dataset where $x_i \in \mathbb{R}^P$ denotes the contextual variables and $y_i \in \mathbb{R}$ denotes the behavioral variable. Contextual anomaly detection (CAD) aims to identify observations for which y is unlikely under the conditional distribution $p(y \mid x)$, rather than observations that are rare under the joint distribution $p(x, y)$. This formulation was introduced explicitly by Song et al. (2007) and subsequently extended in several directions, including dependency-based formulations (Valko et al., 2011; Tang et al., 2013), robust contextual modeling (Liang and Parthasarathy, 2016), and explainable quantile-based methods (Li and van Leeuwen, 2023). The distinction between contextual and behavioral variables is conceptually important because modeling $p(x, y)$ directly tends to overestimate the outlierness of points that simply occur in sparse regions of the contextual space. The anomaly detector then confounds rarity of context with abnormality of behavior.

This same logic has long been present in the medical-statistics literature, although typically under different names. Reference intervals, centile curves, and

age-adjusted standards all aim to characterize how a response variable behaves after conditioning on covariates such as age, sex, or body size (Goldstein, 1972; Healy, 1978; Royston, 1991; Altman, 1993; Zierk et al., 2021; Ammer et al., 2023). Child growth standards (WHO Multicentre Growth Reference Study Group, 2006) are a canonical example: height or weight is not interpreted in isolation, but relative to age and sex. In cardiovascular imaging, analogous ideas appear in the construction of Z-scores and nomograms for aortic diameters (Roman et al., 1989; Lee et al., 2010; Colan, 2013; Campens et al., 2014). However, much of this literature relies on deterministic regression models, often after hand-crafted transformations such as Box–Cox mappings (Box and Cox, 1964), and typically summarizes deviation from normalcy with a point-valued Z-score. The resulting tools are interpretable, but they struggle when conditional variability depends on context, when the reference population is unevenly sampled, or when the model is forced to extrapolate (Chubb and Simpson, 2012; Mawad et al., 2013; Curtis et al., 2016; Ceriotti, 2012; Asch et al., 2019; Lancellotti et al., 2013; Ricci et al., 2021).

These limitations are particularly relevant in high-stakes domains. In cardiology, for example, normalcy assessment is often used to support decisions about further monitoring or intervention, and the consequences of overconfident errors can be substantial (Isselbacher et al., 2022). The domain-specific clinical instantiation is developed in the next chapter; here, the goal is instead methodological. We seek a CAD framework that preserves the intuitive semantics of Z-score reasoning while explicitly distinguishing between two sources of uncertainty. The first is *aleatoric uncertainty* (AU), namely the intrinsic variability of y within the subpopulation sharing the same context. The second is *epistemic uncertainty* (EU), namely the uncertainty induced by finite, biased, or locally sparse data (Senge et al., 2014; Depeweg et al., 2018; Hüllermeier and Waegeman, 2021). In practical terms, AU tells us how much spread is normal at a given context, whereas EU tells us how much the model should be trusted when estimating that spread. Existing CAD methods usually focus on the first aspect but provide little or no access to the second.

The chapter asks a broader question than standard anomaly detection: not only is this observation anomalous given its context, but also, how reliable is that judgment? This perspective is motivated by the same demand for reliability-aware automation that has recently emerged in anomaly detection with abstention or reject options (Perini and Davis, 2023). It also connects to recent work on disentangling uncertainty in heteroscedastic Bayesian models (Patel et al., 2022b). Our contribution is to bring these ideas into contextual anomaly detection by constructing a score that is itself random, so that one can report not only a point estimate but also an interval describing the confidence of the assessment.

Figure 4.1 illustrates the intuition on a synthetic example inspired by WHO

growth curves (WHO Multicentre Growth Reference Study Group, 2006). The behavioral variable is height, the contextual variable is age, and the sampling distribution is intentionally denser for younger subjects than for older ones. In this setting, a non-contextual detector such as Isolation Forest tends to mark observations with rare ages as suspicious simply because they lie in low-density regions of the joint space. Dedicated contextual methods already improve on this behavior, but they still struggle when the variance of the behavioral variable changes across the contextual domain. NS addresses both issues by modeling the conditional mean and the conditional variance, and by attaching a high-density interval to the resulting anomaly score.

4.2 Normalcy score

The starting point for NS is the classical contextual Z-score. If $f(x)$ is a regression model that predicts the expected value of the behavioral variable from the context, and if S is an estimate of the residual standard deviation, then the conventional score is

$$Z(x, y) = \frac{y - f(x)}{S}. \tag{4.1}$$

This formulation is attractive because it standardizes deviations and is easy to interpret. In practice, a value such as $|Z| > 2$ is often used as an operational threshold for flagging abnormality (Roman et al., 1989; Lee et al., 2010; Colan, 2013). Nevertheless, Equation (4.1) is based on strong assumptions. In its simplest form, it uses a deterministic predictor and a single global residual scale. If the variance of $y \mid x$ depends on the context, the denominator averages out that variability, which leads to systematic overestimation of abnormality in low-variance regions and underestimation in high-variance regions (Chubb and Simpson, 2012; Mawad et al., 2013).

A classical response to this problem is to model the conditional variance separately. Early approaches stratified populations into groups (Goldstein, 1972). A more flexible strategy was proposed by Altman (1993), who estimated spread through an additional regression on the absolute residuals. The key limitation, however, remains that the resulting score is still treated as deterministic once the regressions have been fitted. This means that the final normalcy assessment ignores the uncertainty of the regression functions themselves. When data are sparse in part of the contextual space, the model can become epistemically uncertain while still outputting a precise point score.

NS generalizes this reasoning by replacing the two deterministic regressions with two Gaussian-process models. Let f_1 model the conditional mean and f_2

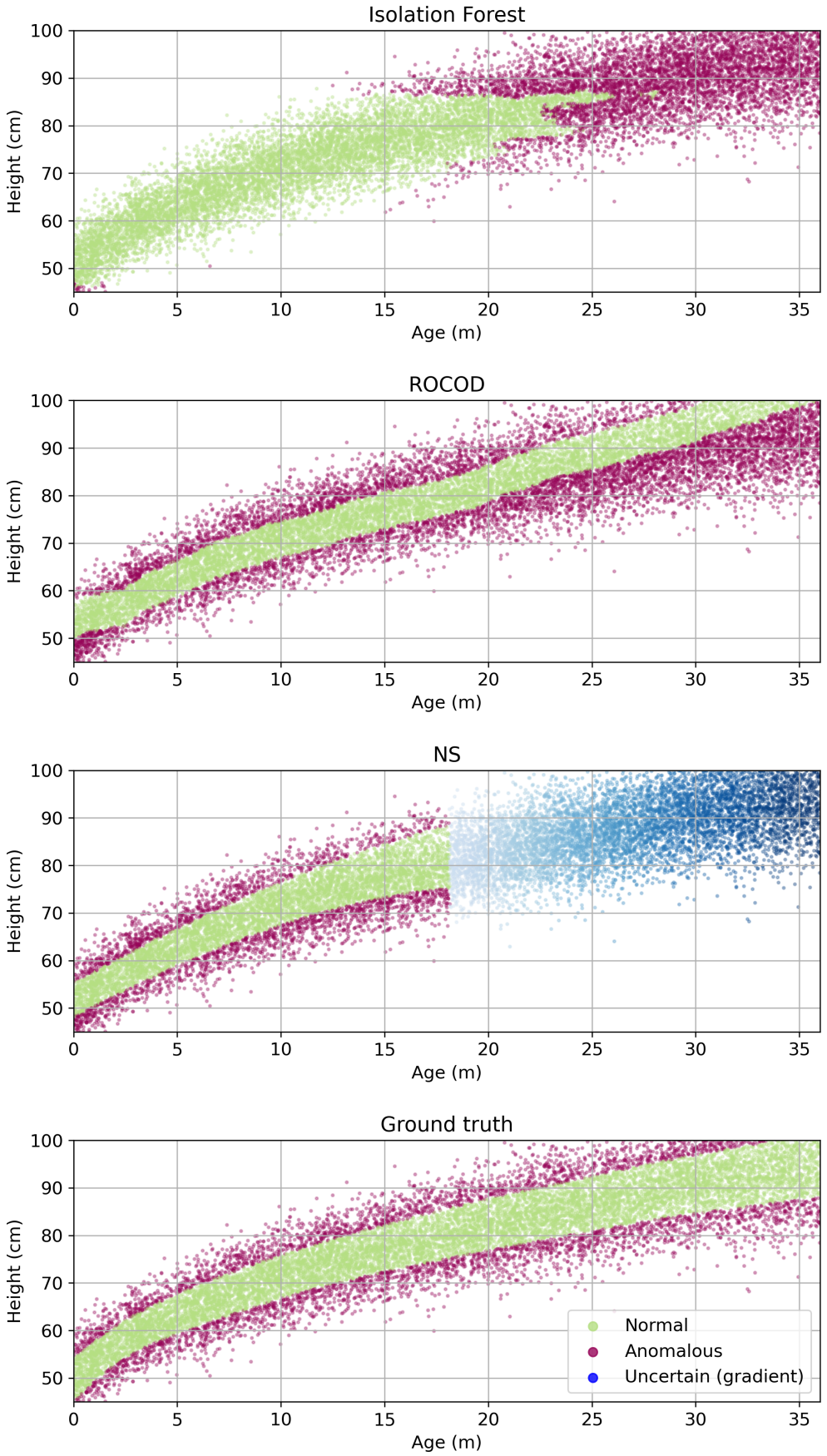


Figure 4.1: Synthetic contextual-anomaly example inspired by WHO growth curves. The objective is to detect observations that are unusual given the context, rather than observations that simply occur in low-density contextual regions. NS models both the context-dependent expected behavior and the uncertainty of the assessment.

model the log-standard deviation:

$$f_1(x) \sim \mathcal{GP}(m_1(x), k_1(x, x')), \tag{4.2}$$
$$f_2(x) \sim \mathcal{GP}(m_2(x), k_2(x, x')). \tag{4.3}$$

The use of a Gaussian process (GP) provides a flexible Bayesian prior over functions (Rasmussen and Williams, 2006). Modeling the log-standard deviation rather than the standard deviation directly guarantees positivity and follows the heteroscedastic GP formulation of Goldberg et al. (1997), later developed in more scalable and practical ways by Le et al. (2005). NS is then defined as

$$\mathrm{NS}(x, y) = \frac{y - f_1(x)}{e^{f_2(x)}}. \tag{4.4}$$

This expression mirrors the Z-score but differs in one fundamental respect: because both $f_1(x)$ and $f_2(x)$ have posterior distributions, $\mathrm{NS}(x, y)$ is itself a random variable.

This probabilistic reinterpretation is the key conceptual move of the chapter. Once the score is random, anomaly assessment becomes richer than a single scalar ranking. A natural point estimate is the posterior expectation

$$s(x, y) \triangleq \mathbb{E}[\mathrm{NS}(x, y)], \tag{4.5}$$

which preserves the semantics of a signed standardized deviation. Under the independence of the two GPs, one obtains

$$s(x, y) = \big(y - m_1(x)\big)\, e^{-m_2(x) + \sigma_2^2(x)/2}, \tag{4.6}$$

where $m_1(x)$ is the posterior mean of the mean process, and $m_2(x), \sigma_2^2(x)$ are the posterior mean and variance of the log-standard-deviation process. In the limit where posterior uncertainty on f_2 vanishes, the correction term $e^{\sigma_2^2(x)/2}$ tends to 1, and the expression collapses to the familiar standardized residual. Thus, NS does not abandon Z-score reasoning; rather, it embeds it in a probabilistic setting where uncertainty can be propagated all the way to the final score.

This propagation is important because it enables a principled distinction between AU and EU. The context-dependent spread modeled by f_2 represents AU, i.e., the variability that would remain even with arbitrarily many samples. By contrast, the posterior variances of f_1 and f_2 represent EU, which shrinks as more observations accumulate near the test context. The two types of uncertainty play different conceptual roles. AU determines what should count as normal variation in a well-sampled region, while EU determines how trustworthy the estimated notion of normality is at the test point. This distinction is one of the reasons why NS is more informative than existing point-valued contextual detectors such as ROCOD (Liang and Parthasarathy, 2016) or quantile-based methods such as QCAD (Li and van Leeuwen, 2023), even when the raw detection performance is similar.

Bayesian estimation and uncertainty intervals

The practical implementation of NS relies on heteroscedastic Gaussian process regression. Exact GP inference scales cubically with the number of observations, which is impractical for repeated cross-validation and for the larger benchmark datasets. We therefore adopt the sparse variational approximation based on inducing variables introduced by Snelson and Ghahramani (2005) and placed on firm variational grounds by Titsias (2009). In this approximation, a set of $M \ll N$ inducing points summarizes the posterior process, reducing the dominant computational cost from $O(N^3)$ to approximately $O(M^2N)$. The model is implemented in GPflow (Matthews et al., 2017), which makes it possible to optimize the variational parameters with natural gradients while updating the remaining model parameters with Adam.

The relevance of this approximation is not merely computational. It also makes the probabilistic formulation usable in contexts where repeated model fitting is required, such as hyperparameter sensitivity studies or repeated anomaly injection experiments. Although the present chapter focuses on a scalar behavioral variable, the same basic logic can be extended. Multi-output or multi-task Gaussian processes (Bonilla et al., 2007) would allow the joint modeling of multiple related behavioral variables, which could capture context-dependent dependencies among them. These extensions are left for future work, but they highlight that the proposed framework is not tied to the scalar setting.

Because $\mathrm{NS}(x, y)$ is a random variable, one needs a summary not only of its central tendency but also of its concentration. We use a highest-density interval (HDI) (Kruschke, 2015) to summarize the posterior uncertainty of the score. Specifically, for a given confidence level, the HDI is the smallest interval containing the specified posterior mass. Since there is no general closed-form expression for the HDI of Equation (4.4), we estimate it numerically by sampling from the posterior of $f_1(x)$ and $f_2(x)$, computing the induced samples of $\mathrm{NS}(x, y)$, estimating the corresponding density, and then numerically extracting the 95% HDI. We denote the length of this interval by $i(x, y)$. The longer the interval, the less concentrated the posterior score, and hence the less reliable the anomaly assessment.

This interval has a natural operational meaning. Suppose that two observations receive similar point estimates $|s(x, y)|$, but one is associated with a narrow interval and the other with a wide interval. The first can be interpreted as a relatively confident anomaly assessment; the second should be regarded with caution because the model has insufficient support to decide sharply. This is especially relevant in CAD, where sparse contexts are often exactly the regions in which deterministic approaches are most brittle. In this sense, NS does not merely rank observations by abnormality; it also exposes where the ranking itself is uncertain.

The use of an interval-valued anomaly score also connects NS to the broader

literature on reliability-aware machine learning. Quantile-based models such as QCAD (Li and van Leeuwen, 2023), based on quantile regression forests (Meinshausen and Ridgeway, 2006), provide a richer description of $p(y \mid x)$ than a mean-only regressor, but they still do not separate irreducible variability from uncertainty due to sparse data support. By contrast, NS is expressly designed to make that distinction visible. The same distinction has become increasingly important in uncertainty-aware learning and decision support (Senge et al., 2014; Hüllermeier and Waegeman, 2021; Patel et al., 2022b). Here it is turned into an anomaly detection tool by asking whether a score should be trusted, not only whether it is large.

4.3 Experimental protocol

The empirical evaluation addresses three questions.

First, how does NS compare with existing contextual and non-contextual anomaly detection methods on standard benchmarks?

Second, does the proposed uncertainty interval behave in a way that is consistent with the intended semantics of epistemic uncertainty?

Third, how sensitive is the framework to standard modeling choices such as the covariance family and the number of inducing points? The corresponding clinical instantiation, which in the paper was used as an additional real-world case study, is intentionally deferred to the next chapter so that the present one remains methodological.

We follow the benchmark protocol introduced by Li and van Leeuwen (2023). Five UCI datasets are considered: Abalone (Nash et al., 1994), Concrete Compressive Strength (Yeh, 1998), Synchronous Machine (UCI, 2021), QSAR Fish Toxicity (Ballabio et al., 2015), and Yacht Hydrodynamics (Gerritsma et al., 1981). None of these datasets contains native contextual anomalies, so anomalies are injected synthetically before train–test splitting by perturbing the behavioral variable while leaving the contextual variables unchanged. This design ensures that the anomaly is genuinely contextual: the same behavioral value may be normal elsewhere in the contextual space, but not for the modified context. The protocol improves over earlier formulations in which some injected anomalies could remain conditionally plausible (Song et al., 2007). Table 4.1 summarizes the benchmark datasets and anomaly ratios.

NS is compared with three classes of baselines. The first class contains contextual methods designed specifically for CAD: ROCOD (Liang and Parthasarathy, 2016) and QCAD (Li and van Leeuwen, 2023). The second class contains regression-based contextual baselines: a standard linear Z-score, the two-stage variance-estimation approach of Altman (1993), and a homoscedastic version of NS, denoted $\mathrm{NS}_{\mathrm{hom}}$,

Table 4.1: Summary of UCI benchmark datasets with injected contextual anomalies.

Dataset	#Samples	Anomaly Ratio	Behavioral Variable
Abalone	4177	2.4%	Rings
Concrete	1030	4.8%	Strength
SynMachine	557	8.9%	IF
Toxicity	908	5.5%	Toxicity
Yacht	308	9.7%	Resistance

which replaces the heteroscedastic variance model with a single GP. The third class contains standard non-contextual anomaly detectors: Isolation Forest (Liu et al., 2008), Local Outlier Factor (Breunig et al., 2000), and Histogram-Based Outlier Score (Goldstein and Dengel, 2012), all implemented via PyOD (Zhao et al., 2019). Their purpose is not to provide competitive contextual baselines, but to demonstrate that the crucial modeling choice is to estimate $p(y \mid x)$ rather than $p(x, y)$.

The GP-based models are trained with a sparse variational implementation in GPflow (Matthews et al., 2017). Rational Quadratic covariance family is used in the default configuration, inducing variables are initialized from a small fraction of the training set, and the optimization alternates natural-gradient updates for the variational parameters with Adam steps for the remaining parameters. This setup provides a computationally manageable approximation while retaining the benefits of Bayesian inference. For evaluation we use five-fold cross-validation and repeat the anomaly injection procedure with five different random seeds. Following common practice in anomaly detection (Aggarwal, 2017; Kuo et al., 2018; Liang and Parthasarathy, 2016; Li and van Leeuwen, 2023), performance is summarized with ROC AUC and PR AUC.

4.4 Results

The main quantitative comparison is reported in Table 4.2. Across the five benchmark datasets, NS achieves the best ROC AUC and PR AUC on four datasets and remains competitive on the remaining one, where QCAD has a narrow advantage. The pattern is consistent across domains with very different sample sizes, anomaly ratios, and forms of contextual dependency. On SynMachine, both NS and strong contextual baselines reach essentially perfect performance, indicating that the dataset is relatively easy once the conditional structure is modeled correctly. The more interesting cases are those in which heteroscedasticity and sparse contexts matter simultaneously: there NS shows its clearest advantage, especially in PR AUC, which is the more demanding metric when anomalies are rare.

A second result is equally important conceptually. The plain Z-score, although

absent from much of the CAD benchmarking literature, turns out to be a remarkably strong baseline. This is one of the most informative findings of the chapter, because it shows that modern CAD methods should not be compared only against highly engineered contextual detectors but also against simple, interpretable reference scores. In the present benchmark, the classical Z-score is often competitive, and the Altman-style heteroscedastic correction also performs strongly on several datasets. NS improves on these baselines by being more flexible and by quantifying reliability, but the results make clear that contextual anomaly detection should not ignore strong statistical baselines from the reference-interval literature.

The non-contextual methods perform substantially worse. Isolation Forest, LOF, and HBOS model the joint structure of the data and confound unusual contexts with genuinely unusual behaviors. This is not a mere implementation artifact but a consequence of the problem formulation itself. It is also the reason why non-contextual methods are still useful in this chapter: they act as context-only probes in later analyses of epistemic uncertainty. Their weak anomaly detection performance does not make them irrelevant; rather, it highlights that the contextual structure is the essence of the problem.

The interval-valued nature of NS becomes most useful when the score is interpreted together with abstention. If $i(x, y)$ is genuinely measuring epistemic uncertainty, then abstaining on the examples with the widest intervals should increase performance on the retained examples. To test this, we compare two abstention strategies. The first abstains on the 5% of test observations with largest HDI width. The second fits an Isolation Forest on the contextual variables only and abstains on the 5% most contextually atypical points. The latter is a strong context-only proxy because it deliberately ignores the behavioral variable and focuses only on contextual sparsity. Table 4.3 shows that HDI-based abstention is equal or superior across all datasets. This supports the interpretation of $i(x, y)$ as a meaningful reliability signal rather than a mere artifact of posterior sampling. In applications where anomalous decisions are costly, the model could defer high-uncertainty cases for expert inspection, an idea closely related to the reject-option perspective studied in anomaly detection (Perini and Davis, 2023).

A complementary analysis studies whether the width of the HDI indeed tracks contextual sparsity. We compute context-only anomaly scores using IForest, LOF, and HBOS, and then measure their ordinal association with $i(x, y)$ through the weighted Kendall's tau statistic of Shieh (1998). Weighted Kendall's tau is useful here because it emphasizes agreement near the top of the ranking, where uncertainty-driven abstention would matter most. The positive correlations reported in Table 4.4 show that HDI width increases in sparse contextual regions, as intended. At the same time, the correlations are not perfect, indicating that the interval does not simply collapse to a context-rarity detector: it remains tied to the

Table 4.2: Average ($\pm$ std) ROC AUC and PR AUC across five independent anomaly injections, each evaluated with 5-fold cross-validation.

	Abalone		Concrete		SynMachine		Toxicity		Yacht	
Method	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC
NS	**0.96 ± 0.01**	**0.65 ± 0.04**	0.89 ± 0.02	0.60 ± 0.01	**1.00 ± 0.00**	**1.00 ± 0.00**	**0.92 ± 0.02**	**0.67 ± 0.04**	**0.97 ± 0.02**	**0.88 ± 0.06**
NS_{hom}	0.95 ± 0.03	0.64 ± 0.06	0.86 ± 0.02	0.52 ± 0.03	**1.00 ± 0.00**	**1.00 ± 0.00**	**0.92 ± 0.02**	0.63 ± 0.05	0.95 ± 0.01	0.57 ± 0.06
Z-score	0.95 ± 0.01	0.57 ± 0.06	0.86 ± 0.03	0.55 ± 0.04	**1.00 ± 0.00**	**1.00 ± 0.00**	0.91 ± 0.02	0.57 ± 0.05	0.82 ± 0.04	0.53 ± 0.09
Altman	0.95 ± 0.01	0.48 ± 0.05	0.87 ± 0.03	0.58 ± 0.06	0.99 ± 0.01	0.96 ± 0.03	0.91 ± 0.02	0.61 ± 0.03	0.81 ± 0.05	0.59 ± 0.11
QCAD	0.90 ± 0.01	0.28 ± 0.04	**0.93 ± 0.02**	**0.64 ± 0.04**	0.98 ± 0.01	0.96 ± 0.02	0.86 ± 0.01	0.45 ± 0.07	0.96 ± 0.03	0.85 ± 0.11
ROCOD	0.93 ± 0.01	0.40 ± 0.05	0.79 ± 0.02	0.34 ± 0.05	0.90 ± 0.04	0.79 ± 0.08	0.84 ± 0.03	0.46 ± 0.07	0.78 ± 0.03	0.31 ± 0.05
IForest	0.77 ± 0.01	0.05 ± 0.01	0.62 ± 0.02	0.08 ± 0.02	0.83 ± 0.03	0.33 ± 0.06	0.71 ± 0.05	0.10 ± 0.02	0.79 ± 0.06	0.32 ± 0.10
LOF	0.92 ± 0.00	0.39 ± 0.03	0.49 ± 0.06	0.06 ± 0.01	0.92 ± 0.02	0.78 ± 0.04	0.61 ± 0.05	0.08 ± 0.02	0.71 ± 0.04	0.20 ± 0.04
HBOS	0.67 ± 0.01	0.04 ± 0.00	0.74 ± 0.05	0.27 ± 0.05	0.67 ± 0.03	0.26 ± 0.04	0.69 ± 0.04	0.10 ± 0.02	0.76 ± 0.05	0.34 ± 0.08

Table 4.3: ROC AUC and PR AUC after abstaining on the 5% most uncertain test instances, either using the HDI width $i(x, y)$ or a contextual Isolation Forest.

Dataset	Abstain using $i(x, y)$		Abstain using IForest	
	ROC AUC	PR AUC	ROC AUC	PR AUC
Abalone	**0.97**	**0.71**	0.96	0.66
Concrete	**0.92**	**0.65**	0.86	0.55
SynMachine	**1.00**	**1.00**	**1.00**	**1.00**
Toxicity	**0.95**	**0.74**	0.92	0.70
Yacht	**1.00**	**0.95**	0.97	0.87

conditional modeling problem.

$$\tau_w = \frac{\sum_{i<j} w_{ij} \operatorname{sgn}\big[(x_i - x_j)(y_i - y_j)\big]}{\sum_{i<j} w_{ij}}. \tag{4.7}$$

Table 4.4: Weighted Kendall's tau between HDI width $i(x, y)$ and context-only anomaly scores. Positive values indicate that uncertainty concentrates in sparse contextual regions.

Dataset	IForest	LOF	HBOS
Abalone	0.71	0.66	0.62
Concrete	0.66	0.66	0.64
SynMachine	0.65	0.62	0.60
Toxicity	0.73	0.68	0.63
Yacht	0.68	0.70	0.65

We also test the sensitivity of the framework to two practical modeling choices: the covariance family and the number of inducing points. Replacing the default Rational Quadratic covariance with Matérn 5/2 or RBF produces only minor changes in the expected score and the interval width. Likewise, increasing the inducing-point ratio beyond 5% yields only modest changes in the score statistics while substantially increasing training time. These experiments do not prove that kernel choice is irrelevant in general, but they do show that the main benefits of NS come from the probabilistic formulation itself rather than from delicate covariance tuning. In that sense, the framework is robust in practice.

Table 4.5: Impact of covariance-family choice on NS statistics.

Kernel comparison	$\Delta\mathbb{E}[\mathrm{NS}]$	$\Delta i(x, y)$
Matérn 5/2 vs RQ	0.07	0.02
RBF vs RQ	0.04	0.01

Table 4.6: Effect of the inducing-point ratio on score stability and training time.

Inducing-point ratio	$\Delta\mathbb{E}[\mathrm{NS}]$ (vs. 5%)	Training time
5%	—	∼10 min
10%	0.06	∼40 min
20%	0.10	∼70 min

4.5 Discussion

Several conclusions follow from these experiments. First, contextual anomaly detection should be understood as conditional normalcy assessment rather than joint outlier detection. This may sound like a semantic distinction, but the empirical comparison shows that it has immediate practical consequences. Methods that ignore the distinction between contextual and behavioral variables underperform consistently, not because they are poorly engineered, but because they solve a different problem. Second, simple statistical baselines remain unexpectedly strong. The competitiveness of the classical Z-score and of Altman's heteroscedastic correction suggests that CAD research benefits from maintaining a close connection with the older literature on reference values and standardization. Third, and most importantly, uncertainty matters. NS provides not only a score, but also a reliability signal that can support abstention and more cautious decision-making.

The chapter is methodological by design, but it is motivated by domains where these properties are especially valuable. One such domain is the assessment of aortic diameters using echocardiographic reference equations (Frasconi et al., 2021), where the stakes of miscalibrated normalcy judgments are clinically relevant (Isselbacher et al., 2022). Rather than presenting that application here, the next chapter specializes the same probabilistic logic to a Bayesian reformulation of the aortic Z-score. The separation is intentional: the present chapter establishes the general CAD methodology, while the next chapter develops the domain-specific clinical tool.

There are also several technical directions for further development. Extending NS to vector-valued behavioral variables would enable the detection of anomalies that are only visible at the multivariate level, even when each scalar component appears individually normal. Multi-output Gaussian processes (Bonilla et al., 2007) provide one principled route, while Wishart-process constructions (Heaukulani and van der Wilk, 2019) offer a way to model context-dependent covariance structure more explicitly.

Overall, NS can be read as a Bayesian extension of Z-score reasoning to contextual anomaly detection. It preserves the interpretability of standardized residuals, models heteroscedasticity directly, and explicitly separates variability in the data from uncertainty in the model. By doing so, it turns anomaly scoring into a

reliability-aware process rather than a purely point-valued ranking.

Chapter 5

More reliable assessment of aortic diameters

5.1 Clinical background and motivation

Thoracic aortic dilatation is clinically relevant because it is associated with adverse outcomes, including dissection, rupture, progressive valvular dysfunction, and the need for prophylactic surgery (Kim et al., 2016; Asch et al., 2019; Ricci et al., 2021). In routine care, the first-line modality for monitoring proximal aortic size is transthoracic echocardiography (TTE), thanks to its wide availability, favorable safety profile, and relatively low cost (Lancellotti et al., 2013; Campens et al., 2014). Yet the interpretation of aortic diameters is not straightforward: a raw measurement can only be understood in relation to patient context, because age, sex, height, weight, body surface area (BSA), and body mass index (BMI) all affect what should be considered physiologically normal (Roman et al., 1989; Devereux et al., 2012; Campens et al., 2014; Lopez et al., 2017; Frasconi et al., 2021). The construction of age-related reference ranges is not a minor statistical detail, but a prerequisite for meaningful clinical use (Altman, 1993; Colan, 2013; Curtis et al., 2016).

An additional practical issue is that currently available reference tools are fragmented. Some studies report limits only at a single aortic level, others provide nomograms rather than patient-specific calculators, and several are restricted to pediatric or adult age ranges only (Campens et al., 2014; Cantinotti et al., 2017; Frasconi et al., 2021). As a consequence, clinicians may need to switch between partially overlapping tools across follow-up, especially in transitional age ranges, and this can introduce inconsistencies that are methodological rather than biological. A more flexible probabilistic formulation is attractive not only because it improves statistical modeling, but also because it offers a route toward a single, coherent framework for normalcy assessment across heterogeneous patient profiles.

In current practice, this contextual interpretation is typically summarized

through the Z-score, which measures how many standard deviations a diameter lies above or below its expected value for a patient with a given set of characteristics. This tool is attractive because it converts a heterogeneous clinical problem into an immediately interpretable number. However, the apparent simplicity of the Z-score conceals strong modeling assumptions. Most available calculators are derived from linear regressions estimated on healthy reference cohorts and use a single residual scale to standardize all patients (Campens et al., 2014; Cantinotti et al., 2017; Patel et al., 2022a). Such an approach is convenient, but it may become unreliable when the relationship between context and diameter is nonlinear, when the variability of normal diameters changes across the covariate space, or when the reference data provide weak support for specific patient profiles (Mawad et al., 2013; Colan, 2013; Curtis et al., 2016).

These issues are especially relevant in patients at increased risk of aortopathy, such as subjects with Marfan syndrome and related disorders or with bicuspid aortic valve (BAV) (Fernandes et al., 2012; Ricci et al., 2021). In such populations, management decisions often depend on repeated measurements near clinically meaningful thresholds. A point estimate alone may be insufficient; clinicians would benefit from knowing not only whether the score is above a threshold, but also whether that conclusion is well-supported by the available reference data. This observation connects the problem to the broader literature on contextual anomaly detection, where the question is not whether an observation is globally rare, but whether it is unusual given its context (Song et al., 2007; Chandola et al., 2009; Li and van Leeuwen, 2023). Within this view, the diameter is the behavioral variable, while age, sex, and body-size descriptors constitute the context.

The goal of this chapter is to reformulate the echocardiographic Z-score in Bayesian terms while preserving its familiar clinical semantics. The key idea is simple: rather than computing a deterministic number from a fixed mean prediction and a global residual scale, we model the conditional distribution of aortic diameters with a heteroscedastic Bayesian regressor and propagate uncertainty to the score itself. The result is a Bayesian Z-score characterized by two outputs: an expected Z-score, which maintains continuity with current clinical practice, and a highest-density interval (HDI), which quantifies how reliable that score is for the patient under examination.

5.2 From classical to Bayesian Z-score

Let $x^* \in \mathbb{R}^d$ denote the patient context and let $y^* \in \mathbb{R}$ be the aortic diameter measured at a fixed anatomical level. In the most general case, the classical Z-score

can be written as

$$Z_{\text{classic}} = \frac{y^* - m(x^*)}{s(x^*)}, \tag{5.1}$$

where $m(x^*)$ and $s(x^*)$ denote the expected diameter and the corresponding standard deviation for a patient with context x^*.

In current echocardiographic practice this quantity is often implemented through deterministic reference equations, typically linear regressions with transformed covariates and a single global residual scale (Roman et al., 1989; Devereux et al., 2012; Campens et al., 2014; Lopez et al., 2017; Martinez-Millana et al., 2018; Patel et al., 2022a). This is clinically convenient, but it hides three limitations that are particularly relevant for aortic assessment.

First, the dependence of aortic diameter on age and body size is not fully linear. Growth is rapid in childhood, then progressively saturates in adulthood, and related nonlinear effects have also been reported for anthropometric descriptors such as BSA and BMI (Daubeney et al., 1999; Chubb and Simpson, 2012; Campens et al., 2014; Sluysmans and Colan, 2016). Classical calculators partially address this issue through transformations or subgroup-specific models, but these remedies remain somewhat ad hoc and may create discontinuities across age ranges or model families (Box and Cox, 1964; Gautier et al., 2010; Campens et al., 2014).

Second, the variability of normal diameters is itself context-dependent. In other words, the spread of physiologic values is not constant across childhood, adulthood, sex, and body-size profiles. A homoscedastic score based on a single RMSE therefore averages together situations with genuinely different variability and may overestimate or underestimate the dispersion that should be expected for a given patient (Goldstein, 1972; Altman, 1993; Chubb and Simpson, 2012; Mawad et al., 2013).

Third, and most importantly for the present application, standard calculators do not communicate epistemic uncertainty. Even if the expected value and the residual variability were modeled correctly, the resulting score would still be less reliable for patient profiles that are weakly represented in the healthy reference cohort. This point matters in practice because reference datasets are rarely uniform across the whole age–body-size space: young children, very large or very small subjects, and atypical anthropometric combinations are typically less represented than middle-range adult profiles. When a patient falls in one of these regions, a point score alone can give a false sense of precision (Kiureghian and Ditlevsen, 2009; Kendall and Gal, 2017; Zhou et al., 2022).

For this reason, in this chapter we do not introduce a new score from scratch, but specialize the normalcy score framework described in Chapter 4 to the clinically established problem of aortic normalcy assessment. The general mathematical construction has already been presented there. Here, the important point is how

that framework is instantiated for echocardiography.

From the viewpoint of contextual anomaly detection, age, sex, height, and weight play the role of contextual variables, whereas the aortic diameter is the behavioral variable. A patient should not be considered abnormal because they are tall, young, or have an unusual body size; the question is whether the measured diameter is unexpectedly large given that context. This is exactly the setting in which the normalcy score perspective is preferable to standard anomaly detection on the joint space.

The Bayesian Z-score is obtained by modeling the conditional distribution of healthy aortic diameters with the same heteroscedastic Gaussian-process machinery introduced for NS, and then propagating posterior uncertainty to the score. In this application, the score remains fully aligned with the clinical language of Z-scores: it still quantifies how unexpected a diameter is relative to a patient-specific reference, but it additionally reports whether that judgment is well supported by the available healthy data.

To keep the notation consistent with the original paper, we denote by Z the patient-specific random score induced by posterior uncertainty in the predictive model, by $\mathrm{BZ} \doteq \mathbb{E}[Z]$ its posterior expectation, and by $I[Z]$ its 95% highest-density interval (HDI) (Kruschke, 2015). In practical terms, BZ is the number that most directly replaces the conventional Z-score, whereas $I[Z]$ provides the clinically missing information about reliability. A narrow interval indicates that the reference model is well supported around the patient's context; a wide interval indicates that the score depends on regions of the covariate space where healthy examples are sparse.

This distinction is especially useful near the standard decision threshold of $Z = 2$. Two patients may have a similar expected score, yet require different interpretations because one lies in a well-covered region of the reference population while the other lies in a poorly supported one. The first case may justify a confident classification; the second should instead be regarded as borderline, not because the diameter itself is necessarily less concerning, but because the supporting evidence from the reference cohort is weaker. In this sense, the Bayesian reformulation improves the clinical usability of the Z-score less by changing its semantic meaning than by clarifying when that meaning is trustworthy.

We use a heteroscedastic Gaussian-process regressor because it addresses the three shortcomings discussed above within a single probabilistic model: it can capture nonlinear trends without committing to rigid parametric transformations, it allows the dispersion of normal diameters to vary across the covariate space, and it exposes epistemic uncertainty through posterior intervals on the score. We also retain the same patient-level reporting strategy of the paper, namely the pair $(\mathrm{BZ}, I[Z])$, because it maps naturally onto the clinical workflow: clinicians can still

reason in terms of a familiar scalar Z-score, but they are also informed when the score is fragile.

A further application-specific advantage of this formulation is that it supports comparison between different input representations without changing the clinical output. In particular, one can either follow the traditional route and summarize body size through BSA or use height and weight as separate covariates. The latter option is attractive because two patients with the same BSA may still differ substantially in body habitus, and a flexible nonparametric model can, in principle, exploit this additional information. In contrast, classical calculators are usually tied to a single handcrafted representation of body size.

Finally, this formulation makes explicit why merging partially complementary healthy cohorts can be beneficial. Increasing the size of the reference population is not useful only because it may slightly improve predictive accuracy; more importantly, it reduces epistemic uncertainty in regions of the covariate space that would otherwise remain poorly supported. For an application such as aortic monitoring, where follow-up decisions are often driven by borderline measurements, this is arguably the main practical gain of the Bayesian approach.

5.3 Data and evaluation protocol

The model was trained on a merged healthy reference population obtained by unifying two previously published echocardiographic cohorts, here denoted by D_1 and D_2 (Campens et al., 2014; Frasconi et al., 2021). The first cohort recruited healthy individuals aged at least one year at Ghent University Hospital between 2008 and 2013; the second collected healthy subjects aged at least five years from pediatric and adult cardiology services in Florence, Trieste, and Altavilla Vicentina between 2013 and 2017. Recruitment, exclusion criteria, and measurement conventions followed the original studies. After harmonizing units, measurement sites, and covariate definitions, the two cohorts were treated as a single reference population. This pooling step is not merely convenient for sample size reasons. It is also methodologically important because the Bayesian formulation makes visible how reference coverage affects epistemic uncertainty: merging partially complementary cohorts enlarges the portion of the age–body-size space where the score is supported by nearby healthy subjects and reduces the number of contexts in which the model must effectively extrapolate.

Comprehensive transthoracic Doppler echocardiography was performed with commercially available phased-array systems according to a predefined protocol for acquisition, storage, and measurement. Parasternal long-axis views were optimized at four aortic levels: annulus, sinuses of Valsalva (SoV), sinotubular junction, and proximal ascending aorta (AA). The largest diameters were measured at end-

Table 5.1: Baseline characteristics of the healthy reference population ($N = 1{,}947$). Values are median [interquartile range].

Variable	Total	Male	Female
Age [years]	37 [17–53]	24 [16–48]	43 [24–56]
Height [cm]	167 [158–174]	173 [164–180]	163 [157–168]
Weight [kg]	64 [53–75]	70 [55–80]	60 [52–69]
BSA [m^2]	1.7 [1.5–1.9]	1.9 [1.6–2.0]	1.6 [1.5–1.8]
BMI [kg/m^2]	22.6 [19.7–25.7]	23.0 [19.6–25.9]	22.3 [19.8–25.5]
SoV diameter [mm]	29 [26–32]	30 [27–33]	28 [25–31]
AA diameter [mm]	28 [25–31]	28 [25–31]	28 [25–32]

diastole, perpendicular to the aortic axis, using the leading-edge to leading-edge technique except at the annulus. For the present study, we focused on SoV and AA because these levels are central to clinical monitoring and were available in both healthy and at-risk cohorts (Campens et al., 2014; Frasconi et al., 2021). Height and weight were recorded at the time of TTE, BMI was computed as weight/height^2, and BSA was calculated according to the Du Bois formula.

The merged reference population comprised $N = 1{,}947$ healthy subjects aged 1–89 years, thereby covering childhood, adolescence, adulthood, and older ages. Its baseline characteristics are summarized in Table 5.1. The value of the merged cohort is not merely statistical. It provides broader demographic coverage than each source alone, and it makes epistemic uncertainty clinically visible: even after pooling data, some combinations of age and body size remain underrepresented, which is precisely where a reliability-aware score is most useful.

For independent clinical validation, we evaluated two at-risk cohorts that had not undergone aortic surgery. The first cohort consisted of 117 subjects with Marfan syndrome and related disorders, diagnosed according to revised Ghent criteria and confirmed by the identification of a pathogenic or likely pathogenic variant in one of the main aortopathy-associated genes. These patients were prospectively evaluated at Ghent University Hospital between March 2024 and September 2025. The second cohort consisted of 351 subjects with BAV previously described by Frasconi et al. (2021). Both studies were conducted according to the Declaration of Helsinki, and participants provided written informed consent. These cohorts are clinically meaningful because they represent situations in which threshold-based decisions are common, and underestimation of dilatation may have direct implications for follow-up and management.

We considered four complementary analyses.

First, we compared the prevalence of dilatation defined as $Z > 2$ under the Campens calculator and under the Bayesian Z-score, explicitly reporting straddle

Table 5.2: Prevalence of aortic dilatation ($Z > 2$) in Marfan and BAV patients according to Campens' Z-score and the proposed Bayesian Z-score, for the sinuses of Valsalva (SoV) and ascending aorta (AA). "Straddle" denotes cases whose 95% HDI crosses the clinical threshold.

	SoV			AA		
Cohort	Campens	Bayesian	Straddle	Campens	Bayesian	Straddle
BAV	33.6%	35.6%	1.7%	64.7%	68.9%	0.9%
Marfan	53.0%	63.2%	7.7%	15.9%	16.8%	3.5%

cases, i.e., patients whose 95% HDI crossed the threshold.

Second, we assessed predictive accuracy on the healthy reference set by 5-fold cross-validation, comparing mean absolute error (MAE) and RMSE between the Campens model and Bayesian variants with different input sets.

Third, we visualized patient-level agreement and disagreement between the classical and Bayesian scores in the BAV and Marfan cohorts.

Fourth, we mapped the width of the Bayesian HDI across grids of age and BSA for fixed diameters, to localize where the reference model is reliable and where it is not.

5.4 Results

Prevalence of dilatation in at-risk cohorts. Table 5.2 reports the prevalence of aortic dilatation in the BAV and Marfan cohorts according to the Campens Z-score and the Bayesian Z-score, for SoV and AA. In both cohorts, the Bayesian score identifies at least as many positive cases as the classical calculator, and often more. The difference is especially evident for Marfan patients at the sinuses of Valsalva, where the Bayesian score yields a noticeably larger fraction of $Z > 2$ cases. Importantly, the table also reports the percentage of straddle cases, that is, patients whose 95% HDI spans the threshold and cannot be interpreted as confidently normal or confidently abnormal.

Figure 5.1 complements this result by showing how the positive sets of the Campens and Bayesian methods overlap. The Bayesian Z-score detects additional positives, but it also makes explicit which of these belong to the epistemically uncertain region through the straddle annotation. This is a practical distinction: the method does not simply "increase sensitivity" in the abstract, but highlights where the increase is supported by confident estimates and where it should instead trigger cautious interpretation.

Nonlinearity, heteroscedasticity, and predictive accuracy. Figure 5.2 compares the predicted mean diameter versus age obtained from the classical Campens model

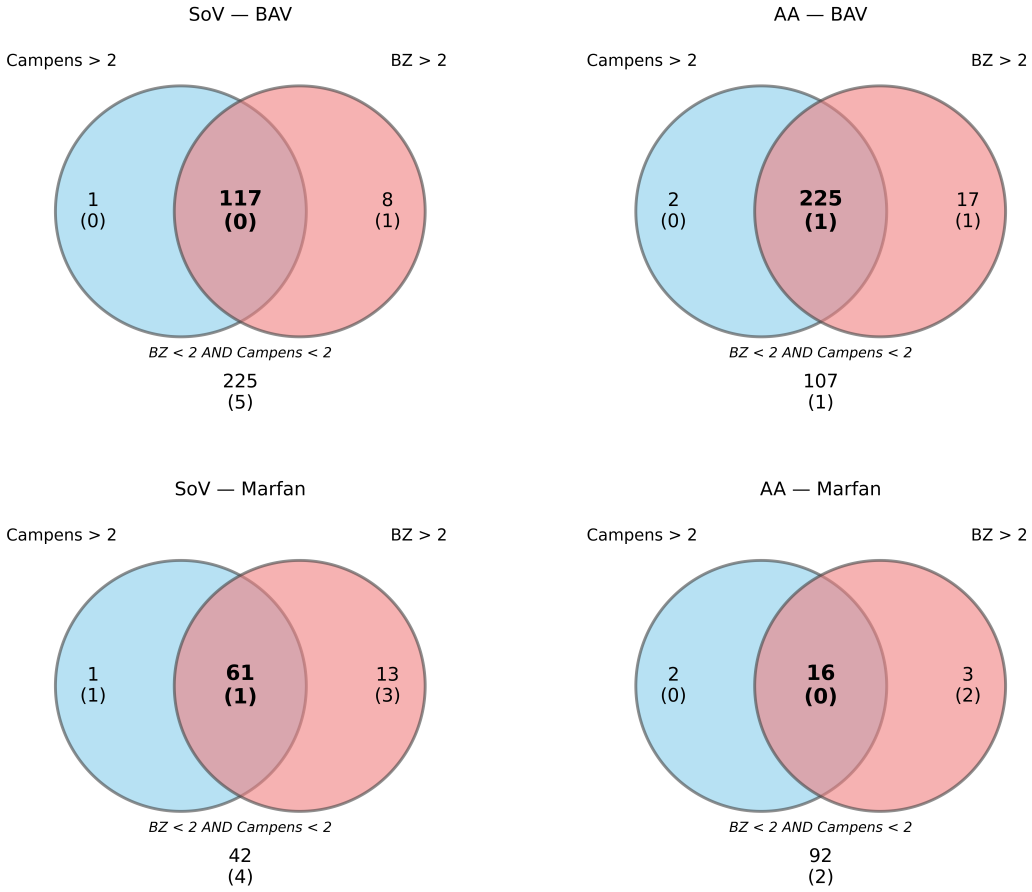


Figure 5.1: Overlap between patients classified as dilatated by the Campens calculator and by the Bayesian Z-score in the BAV and Marfan cohorts, for both SoV and AA. Parenthesized numbers denote straddle cases, i.e., patients whose 95% HDI crosses the decision threshold.

and from the heteroscedastic Gaussian-process regressor underlying the Bayesian Z-score. The two approaches recover a remarkably similar overall shape: rapid growth during childhood and adolescence followed by a plateau in adulthood. This provides empirical support for the long-standing use of logarithmic transformations in conventional equations. At the same time, however, the Bayesian model reveals a feature that the classical calculator cannot expose, namely the fact that predictive variability changes with age and sex. The shaded bands are narrower in some regions and wider in others, directly illustrating heteroscedasticity.

The corresponding cross-validated prediction errors are reported in Table 5.3. The Bayesian approach matches or slightly improves on the Campens baseline in terms of MAE and RMSE, and a small additional gain is observed when height and weight are used as separate covariates instead of being compressed into BSA. The numerical improvement is modest, which is itself an informative result: the main value of the Bayesian formulation is not only lower error, but the fact that

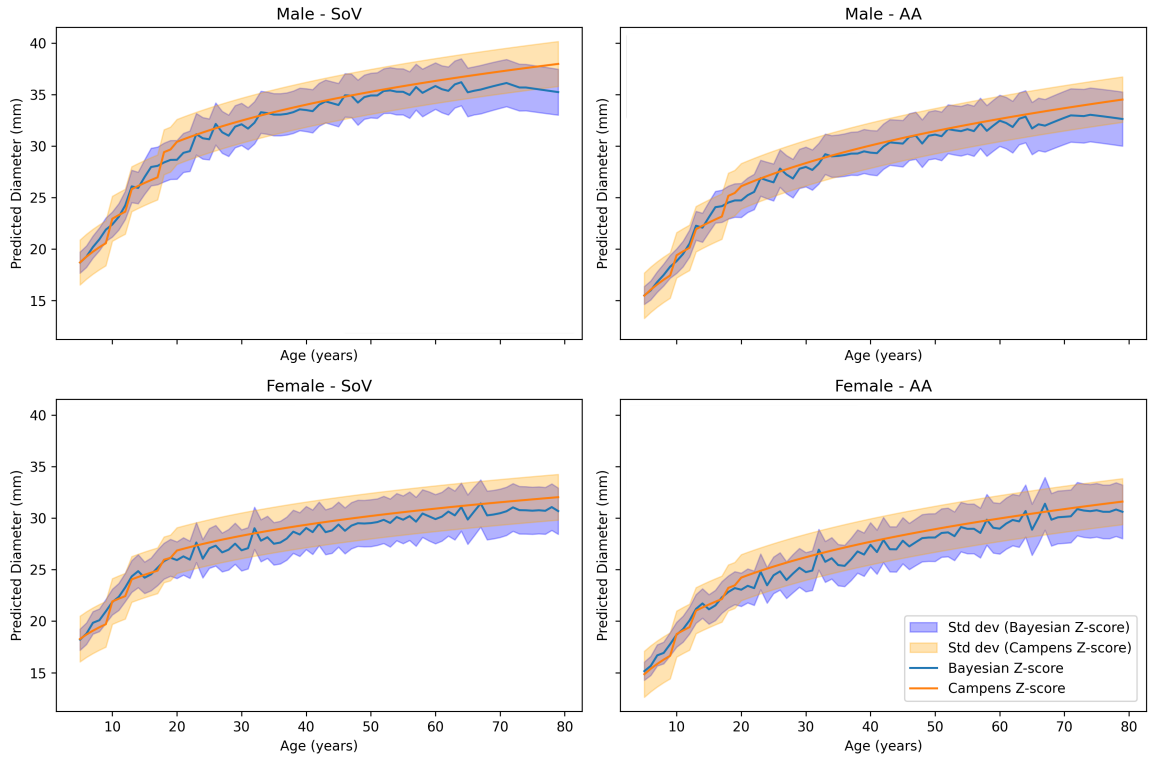


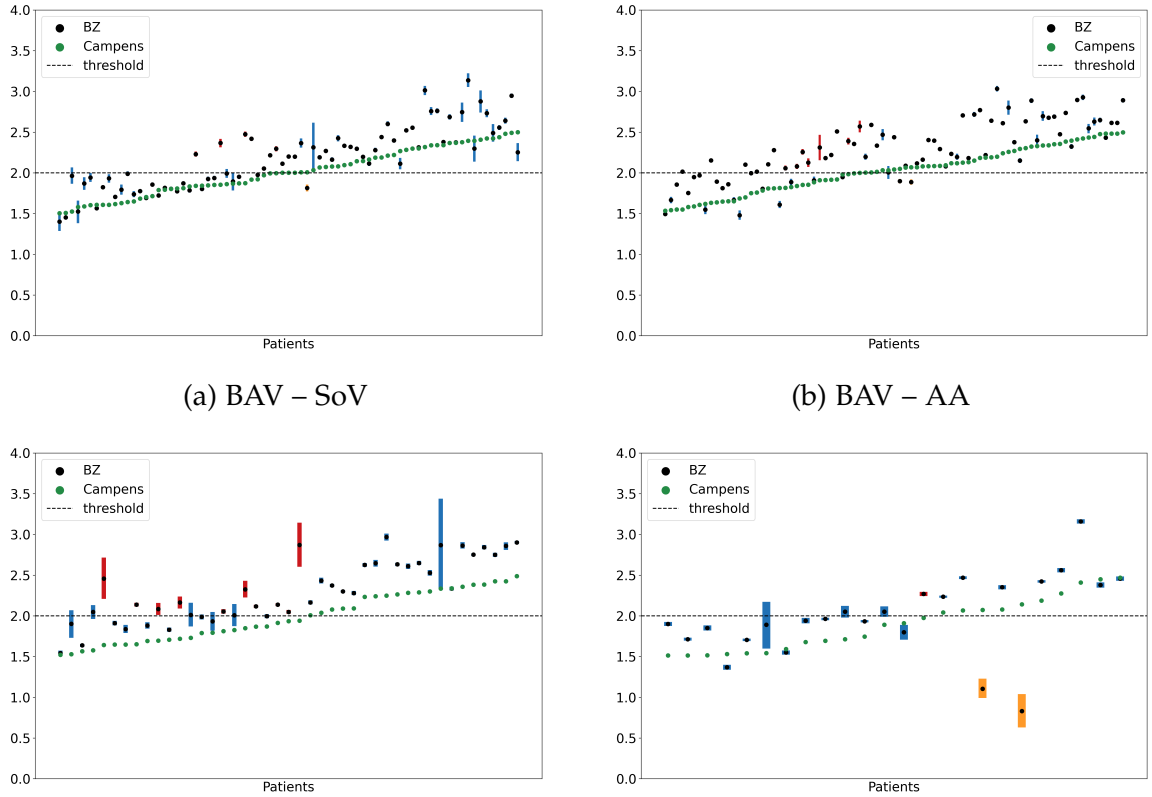
Figure 5.2: Predicted mean aortic diameter versus age in the healthy reference population, stratified by sex. Orange curves correspond to the Campens model, blue curves to the heteroscedastic Gaussian-process regressor underlying the Bayesian Z-score, and shaded areas denote one predictive standard deviation.

Table 5.3: Cross-validated prediction error on the healthy reference set for three models. Results are reported separately for SoV and AA. MAE = mean absolute error; RMSE = root-mean-square error, both in mm.

	SoV		AA	
Model	MAE [mm]	RMSE [mm]	MAE [mm]	RMSE [mm]
Campens (BSA)	2.15	2.75	2.16	2.81
Bayesian GP (BSA)	2.14	2.75	2.15	2.80
Bayesian GP (Height + Weight)	2.13	2.73	2.14	2.79

similar accuracy can be achieved while simultaneously providing patient-specific uncertainty quantification.

Patient-level relevance of epistemic uncertainty. Figure 5.3 shows expected Bayesian Z-scores and their HDIs for individual patients in the BAV and Marfan cohorts, ordered by increasing Campens score. HDI width varies substantially across patients, which means that the same point score can have very different



(a) BAV – SoV

(b) BAV – AA

(c) Marfan – SoV

(d) Marfan – AA

Figure 5.3: Expected Bayesian Z-score (black dots) and 95% HDI (blue bars) versus the Campens Z-score (green) across the two pathological cohorts and two anatomical levels. Bars colored in red correspond to patients with $BZ > 2$ and Campens< 2, whereas orange bars indicate the opposite case. Intervals crossing the threshold identify straddle cases.

reliability depending on the context. Straddle cases are clinically interpretable: they identify patients who sit close to the decision threshold but for whom the reference data are insufficiently informative to support a definitive call.

The same message is conveyed in a more explicit way by Figure 5.4: in one patient, the score is borderline, and the interval overlaps the threshold, whereas in the other the interval is narrow and the conclusion is stable. From the viewpoint of clinical decision support, this distinction is exactly the added value of the Bayesian reformulation.

Where the model is reliable. Finally, Figure 5.5 maps the length of the 95% HDI across a grid of age and BSA values for fixed diameters, separately by sex and anatomical level. Blue regions correspond to contexts where the score is strongly supported by healthy reference data; red regions indicate age and body-size combinations for which the reference population is sparse, and the posterior

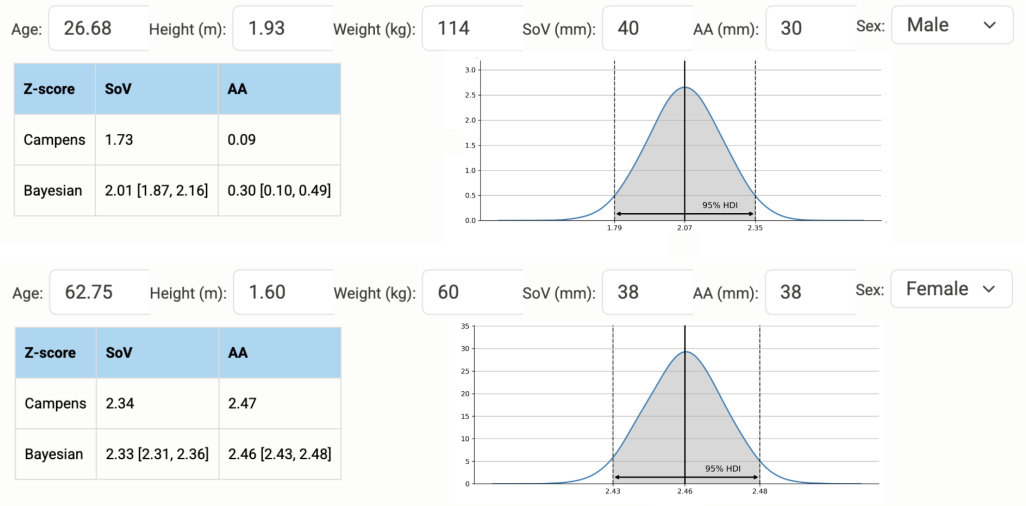


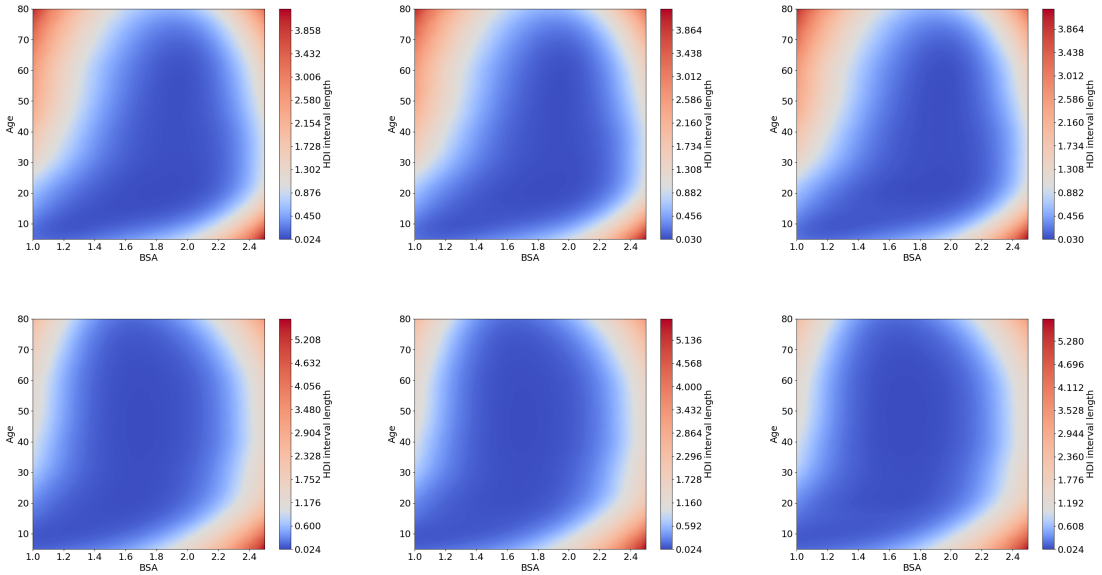
Figure 5.4: Examples of Bayesian Z-score outputs for two patients. Besides the expected score, the method reports a 95% HDI. The upper example illustrates a borderline case whose interval overlaps the conventional decision threshold; the lower example shows a confidently abnormal case with a narrow interval entirely above threshold.

over the score is correspondingly diffuse. The maps show that epistemic uncertainty concentrates at the boundaries of the design space, for example, in very young children, older adults, and extreme body-size profiles. This observation has two implications. Scientifically, it validates the intuition that uncertainty is tied to coverage. Clinically, it suggests where additional healthy data would most improve the reliability of future calculators.

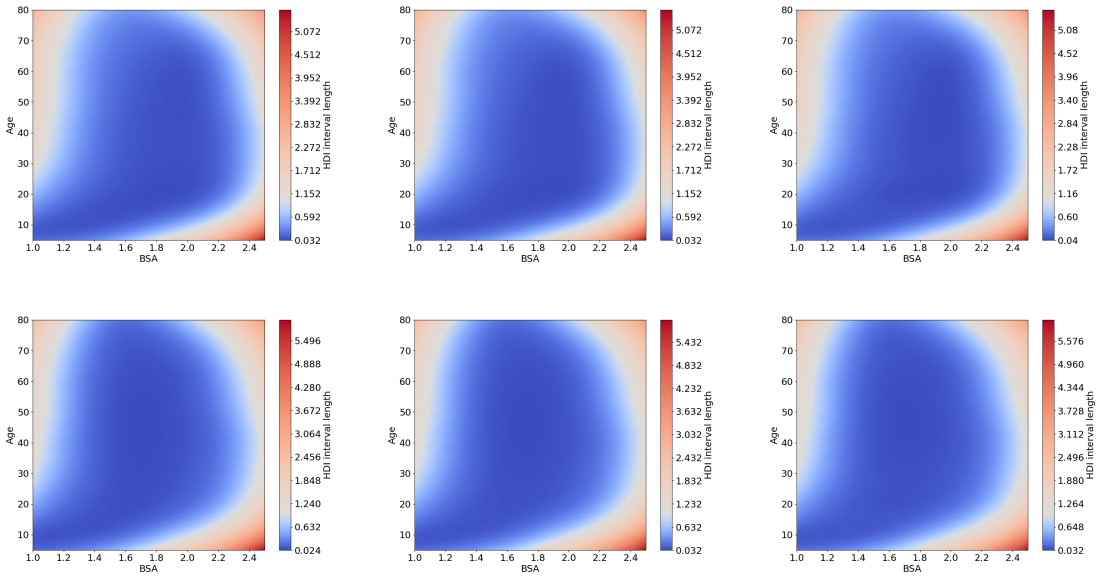
5.5 Discussion

This chapter translated the probabilistic normalcy-assessment logic introduced earlier in the thesis into a clinically familiar interface for aortic monitoring. The Bayesian Z-score preserves the intuitive interpretation of the conventional Z-score while addressing three limitations that have long been recognized in the echocardiographic literature: linearity assumptions, homoscedastic residual variance, and the lack of any explicit representation of model uncertainty (Mawad et al., 2013; Colan, 2013; Curtis et al., 2016). The empirical results indicate that this reformulation is not only statistically appealing but clinically meaningful. In the Marfan and BAV cohorts, the method identifies a slightly larger prevalence of dilatation and, more importantly, distinguishes confidently abnormal cases from borderline cases that should be interpreted with caution.

The relationship with the previous chapter is straightforward: the Bayesian Z-score is not a different principle from the normalcy score, but a domain-specific



(a) Sinuses of Valsalva (SoV). Top row: male subjects; bottom row: female subjects.



(b) Ascending aorta (AA). Top row: male subjects; bottom row: female subjects.

Figure 5.5: Heatmaps of the 95% HDI length of the Bayesian Z-score as a function of age and BSA for fixed diameters. Cooler colors indicate more concentrated posteriors and lower epistemic uncertainty; warmer colors indicate contexts where the score is less reliable because the reference data provide weaker support.

instantiation expressed in the language that clinicians already use. The expected value of the score preserves continuity with existing practice, whereas the HDI adds a reliability layer that standard calculators fail to provide. In this sense, the chapter illustrates how a general uncertainty-aware scoring principle can be specialized into a clinically deployable decision-support tool.

Chapter 6

Few-shot image source attribution using tiny autoencoders

6.1 Motivation

This chapter is intentionally positioned as a methodological complement to the core contributions of the thesis. Its application domain is multimedia forensics rather than clinical decision support, and the final goal is not anomaly detection in the strict sense but few-shot source attribution. Nevertheless, the chapter reuses a key modeling primitive that recurs throughout the thesis: learning a compact representation and turning reconstruction residuals into an informative score. In Chapter 3, reconstruction-based and one-class models are used to quantify deviation from physiological regularity; here, a bank of compact autoencoders is used to quantify how well an image patch conforms to the characteristic output of a candidate generator.

The methodological interest of this study is twofold. First, it shows that reconstruction-error-based pipelines remain effective even when the downstream task is no longer "normal versus abnormal" but multiclass attribution under severe data scarcity. Second, it highlights a deployment regime that is also relevant for biomedical machine learning: limited data per class, limited computational budget, and the need to extend a trained system when new classes appear. In this sense, the chapter does not broaden the clinical claims of the thesis; rather, it reinforces the more general lesson that compact reconstruction models can provide stable and interpretable scoring mechanisms across heterogeneous domains.

Generative models based on GANs and diffusion have made it trivially easy to produce photorealistic images from text prompts (Goodfellow et al., 2014; Ho et al., 2020; Rombach et al., 2022). While this democratization of content creation enables many positive applications, the same technology can be misused to disseminate misinformation and propaganda. In situations where the generating

model is unknown, the ability to attribute a synthetic image to its source can help investigators trace the origin of malicious content and understand its provenance. Existing attribution techniques generally fall into two categories: (i) white-box methods, which assume access to the generator's architecture and weights, and (ii) black-box approaches that learn to classify generated images without access to the model. Both approaches face practical obstacles. White-box assumptions rarely hold in the wild because most models are proprietary. Black-box methods, including state-of-the-art CLIP-based (Radford et al., 2021) detectors such as DE-FAKE (Sha et al., 2023), require hundreds of training examples per generator, consume hundreds of millions of parameters, and must be retrained from scratch when new models appear. As the number of private generative models grows, these limitations threaten the scalability and deployability of passive forensics systems.

At the same time, forensic analysts are often constrained by tight time and data budgets: newly observed generators may only produce a handful of accessible samples before they evolve or disappear, and any attribution pipeline must integrate smoothly with downstream review processes. In such low-resource conditions, architectural simplicity and modularity are not only desirable but necessary. A method that can be extended to accommodate new sources without catastrophic interference, while maintaining competitive accuracy from very few examples, would substantially improve the time-to-deployment of operational systems.

This chapter describes a modular framework for few-shot source attribution. The method pretrains a small decoder on natural images and, for each generator, learns a compact encoder from only a handful of synthetic samples. The reconstruction error produced by each encoder-decoder pair forms a discriminative feature vector, which is fed to a lightweight SVM (Cortes and Vapnik, 1995) for classification. Beyond improving accuracy, this strategy reduces memory footprint and simplifies maintenance: adding a new generator amounts to training a tiny encoder and updating the final classifier, leaving the shared decoder and previously learned encoders untouched.

This design yields three key advantages: (a) state-of-the-art accuracy in few-shot settings; (b) graceful scaling as the number of candidate models grows; and (c) robustness to common post-processing such as JPEG compression and resizing.

6.2 Methodology

A base autoencoder (E_N, D) is first trained on a large corpus of natural images to capture general image statistics. The encoder E_N maps non-overlapping 64×64 patches to a latent space, and the decoder D reconstructs the patches. After pretraining, the decoder D is frozen and reused for all subsequent tasks. For each generative model i in a given forensic scenario, a dedicated encoder E_i is trained on

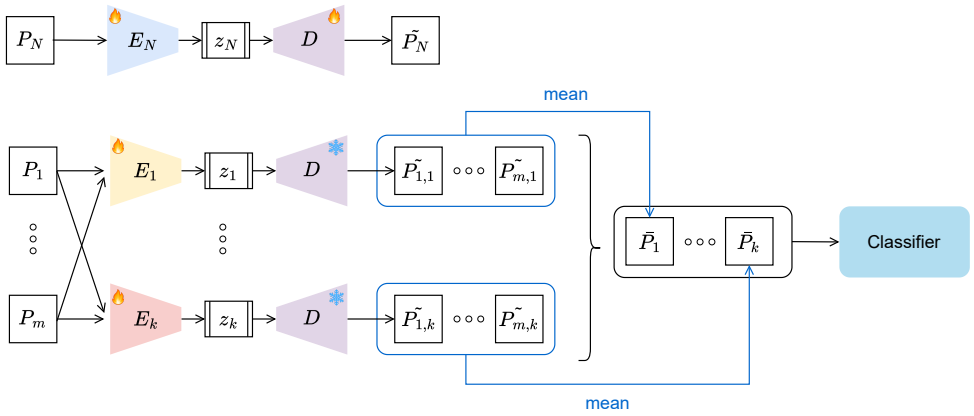


Figure 6.1: Schematic overview of our method. A base autoencoder (E_N, D) is trained on natural images. The decoder D is then frozen and shared by model-specific encoders E_i trained on a few images generated by model i. Reconstruction errors produced by each (E_i, D) form a feature vector for a linear SVM classifier.

a few generated images while keeping D fixed. The resulting autoencoder (E_i, D) learns to approximate the distribution of patches emitted by generator i with only a handful of examples, leveraging the strong prior encoded in D to avoid overfitting. Figure 6.1 illustrates this design.

While conceptually simple, this split between a shared decoder and model-specific encoders has two important consequences. First, it enforces a common reconstruction space across classes: all encoders must learn codes that the same D can decode well. This regularization discourages trivial memorization and reduces drift when only a few class-specific samples are available. Second, it enables class-incremental updates with negligible interference: new encoders are trained in isolation, and D is never altered once pretrained.

Given an input image x, we extract a set of non-overlapping patches $S(x)$ and select a subset of high-entropy patches $S_E(x)$. For each encoder E_i, the reconstruction error over $S_E(x)$ is

$$r_i(x) = \frac{1}{|S_E(x)|} \sum_{\bar{x} \in S_E(x)} \left\| \bar{x} - D(E_i(\bar{x})) \right\|^2. \tag{6.1}$$

Collecting the errors from k encoders yields a feature vector $r(x) = (r_1(x), \ldots, r_k(x))$. A linear SVM trained on these features performs the final attribution. The modularity of this architecture allows new classes to be incorporated by training only a new encoder E_{k+1}, while keeping D and existing encoders unchanged.

Entropy-based patch selection. Patches extracted from synthetic images exhibit varying degrees of discriminative content. Smooth backgrounds or flat regions

contain little information about the generator, whereas textured areas often reveal subtle artefacts arising from upsampling filters, denoisers, or color processing. Our method selects only the 16 patches per image with the highest entropy for training and inference. This simple preselection acts as a signal-to-noise booster: by excluding low-information regions, the autoencoders receive truly informative content and converge more stably in the few-shot regime.

Two entropy definitions are considered:

- **Shannon entropy** of the pixel intensity histogram of a patch. High Shannon entropy corresponds to large intensity variability and tends to identify complex textures and edges. It is inexpensive to compute and works well even when the number of training images is extremely small.
- **GLCM entropy**, defined as the average Shannon entropy of gray-level co-occurrence matrices computed over a set of spatial offsets. Given a quantized grayscale patch I and offsets Ω, the GLCM entropy is

$$H_{\mathrm{GLCM}}(I) = -\frac{1}{|\Omega|} \sum_{\Delta \in \Omega} \sum_{i,j=1}^{G} P^{(\Delta)}(i,j) \log P^{(\Delta)}(i,j), \tag{6.2}$$

  where $P^{(\Delta)}(i,j)$ is the probability that gray levels i and j occur at relative offset Δ. This measure emphasizes second-order texture patterns and has been used in the image analysis literature (Haralick et al., 2007). In practice, it tends to prioritize micro-textures that survive mild downsampling and compression, which is advantageous for robustness.

Training details. The base autoencoder (E_N, D) is pretrained on the Flickr30k dataset (Plummer et al., 2016) using the mean squared reconstruction loss. Model-specific encoders are trained with the Adam optimizer (learning rate 10^{-3}) for 1000 epochs on the selected patches of a few generated images. At test time, each image is divided into non-overlapping 64×64 patches, the top-entropy patches are scored by each encoder, and the resulting feature vector is fed to a linear SVM implemented with scikit-learn. To stabilize training and improve generalization, only the decoder is frozen and reused; the classifiers are trained from scratch in each experiment.

6.3 Experimental evaluation

All reported results are averaged over five random splits, and standard deviations are indicated. Baselines include ResNet50 (He et al., 2016), VGG16 (Simonyan and Zisserman, 2014), XceptionNet (Chollet, 2017), ViT-B-16 (Dosovitskiy et al.,

Table 6.1: Few-shot source attribution performance on eight generative models. Accuracy and macro-F1 are reported as mean ± standard deviation over five runs. Our method consistently outperforms the baseline detectors.

Training images per class		5	10	20
ResNet50	Accuracy	0.17 ± 0.01	0.20 ± 0.01	0.24 ± 0.01
	F1	0.17 ± 0.01	0.20 ± 0.01	0.24 ± 0.01
VGG16	Accuracy	0.17 ± 0.01	0.19 ± 0.01	0.23 ± 0.01
	F1	0.17 ± 0.01	0.19 ± 0.01	0.23 ± 0.01
XceptionNet	Accuracy	0.15 ± 0.01	0.16 ± 0.01	0.19 ± 0.01
	F1	0.12 ± 0.01	0.13 ± 0.01	0.17 ± 0.01
ViT-B-16	Accuracy	0.14 ± 0.01	0.16 ± 0.01	0.18 ± 0.00
	F1	0.14 ± 0.01	0.15 ± 0.01	0.18 ± 0.00
Swin-B	Accuracy	0.15 ± 0.01	0.20 ± 0.01	0.25 ± 0.01
	F1	0.15 ± 0.01	0.20 ± 0.01	0.25 ± 0.01
Swin-T	Accuracy	0.15 ± 0.00	0.17 ± 0.01	0.23 ± 0.01
	F1	0.14 ± 0.00	0.17 ± 0.01	0.23 ± 0.01
DE-FAKE	Accuracy	0.22 ± 0.01	0.30 ± 0.01	0.38 ± 0.01
	F1	0.20 ± 0.02	0.29 ± 0.01	0.38 ± 0.01
Ours	Accuracy	**0.29** ± 0.04	**0.38** ± 0.03	**0.44** ± 0.01
	F1	**0.29** ± 0.03	**0.36** ± 0.03	**0.42** ± 0.01

2020), Swin-B and Swin-T (Liu et al., 2021), and the CLIP-based DE-FAKE method (Sha et al., 2023). Across all settings, we emphasize class-balanced accuracy and macro-F1 to account for potential class imbalance and avoid dominance by easier classes.

Few-shot source attribution

Table 6.1 compares our method with the baseline detectors when trained on 5, 10, and 20 images per class. Our method consistently outperforms all competitors. When only 10 examples are available per model, it achieves 0.38 accuracy and 0.36 macro-F1, whereas DE-FAKE attains 0.30 accuracy and 0.29 macro-F1. The CNN and Transformer baselines remain below 0.25 accuracy even at 20 shots. These results confirm that, despite its small parameter count, the reconstruction-error representation is already sufficiently expressive to separate generators after minimal supervision. Qualitatively, the remaining misclassifications tend to occur among architecturally similar diffusion models, which is consistent with their overlapping artifact distributions.

The few-shot results should be read together with the parameter footprint reported in Table 6.2. While the strongest baseline, DE-FAKE, relies on a fixed 152M-parameter CLIP backbone, our method remains below one million trainable

Table 6.2: Trainable parameter count of DE-FAKE and of our method as the number of registered generators increases. For our method, the total grows linearly because each new class adds one tiny encoder while the shared decoder is reused.

Number of classes	2	3	4	5	6	7	8
DE-FAKE	152M	152M	152M	152M	152M	152M	152M
Ours	228k	321k	414k	508k	601k	694k	787k

Table 6.3: Ablation study on the decoder training strategy. Freezing the decoder after pretraining improves accuracy and F1 compared with finetuning or training from scratch.

Training images per class		5	10	20
From scratch	Acc.	0.26 ± 0.03	0.34 ± 0.01	0.41 ± 0.02
	F1	0.25 ± 0.03	0.31 ± 0.02	0.38 ± 0.02
Finetuned decoder	Acc.	0.26 ± 0.01	0.34 ± 0.03	0.40 ± 0.02
	F1	0.26 ± 0.01	0.32 ± 0.02	0.37 ± 0.02
Frozen decoder	Acc.	$\mathbf{0.29} \pm 0.04$	$\mathbf{0.38} \pm 0.03$	$\mathbf{0.44} \pm 0.01$
	F1	$\mathbf{0.29} \pm 0.03$	$\mathbf{0.36} \pm 0.03$	$\mathbf{0.42} \pm 0.01$

parameters even when eight candidate generators are registered. This compactness is not an accessory property: it is one of the main reasons why the approach remains practical in class-incremental scenarios, where new sources may have to be registered quickly and with very limited compute.

Frozen decoder vs finetuning. Table 6.3 analyses the effect of freezing the decoder after pretraining versus finetuning or training from scratch. Keeping the decoder fixed yields the best performance across all data budgets. This highlights the importance of the shared decoder as a prior: it anchors each encoder to a common reconstruction space, reducing overfitting when only a few samples are available. Intuitively, co-adapting the encoder and decoder weakens this anchor, allowing incompatible codes to emerge across classes; the effect is most evident at 5–10 shots, where data scarcity exacerbates variance.

Scalability over classes. The class-incremental nature of our method is assessed by increasing the number of generative models from 2 to 8. Table 6.4 shows that accuracy decreases gracefully as classes are added, from 0.77 with two models to 0.43 with eight. DE-FAKE likewise degrades but remains worse than our method at every point. Importantly, the parameter count of our method grows linearly, with each new encoder adding only about 93k parameters, whereas all baselines remain monolithic and cannot be extended without retraining. This scaling property is

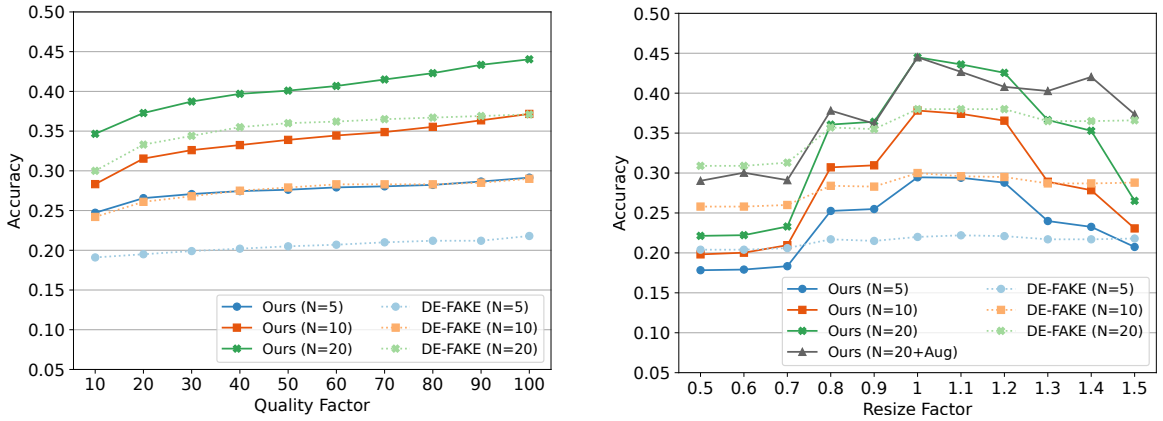
Table 6.4: Scalability of our method as the number of classes grows. Accuracy and F1 are reported for 20 training images per class.

Number of classes		2	3	4	5	6	7	8
ResNet50	Acc.	0.62	0.47	0.39	0.31	0.30	0.27	0.25
	F1	0.61	0.47	0.39	0.31	0.30	0.27	0.25
VGG16	Acc.	0.60	0.48	0.38	0.32	0.29	0.26	0.24
	F1	0.60	0.47	0.38	0.32	0.29	0.26	0.24
XceptionNet	Acc.	0.55	0.41	0.33	0.26	0.25	0.22	0.20
	F1	0.53	0.36	0.30	0.21	0.21	0.18	0.17
ViT-B-16	Acc.	0.57	0.39	0.32	0.26	0.23	0.21	0.19
	F1	0.57	0.39	0.32	0.26	0.23	0.21	0.19
Swin-B	Acc.	0.61	0.46	0.38	0.32	0.29	0.27	0.25
	F1	0.61	0.45	0.38	0.32	0.29	0.27	0.25
Swin-T	Acc.	0.56	0.43	0.34	0.28	0.26	0.25	0.22
	F1	0.56	0.42	0.34	0.28	0.26	0.25	0.22
DE-FAKE	Acc.	0.70	0.62	0.56	0.47	0.45	0.41	0.40
	F1	0.70	0.62	0.56	0.47	0.45	0.40	0.40
Ours	Acc.	**0.77**	**0.70**	**0.61**	**0.51**	**0.49**	**0.43**	**0.43**
	F1	**0.77**	**0.72**	**0.61**	**0.53**	**0.51**	**0.45**	**0.46**

pivotal for field deployment: forensic analysts can register a new model with a short specialization phase that does not affect existing capabilities.

Robustness to post-processing. Synthetic images encountered in practice may have undergone compression or resizing. The top panel of Figure 6.2 shows accuracy as a function of JPEG quality factor. Accuracy declines for all methods as compression increases, but our method maintains a consistent margin over DE-FAKE. The bottom panel plots accuracy versus resizing scale. DE-FAKE is initially more robust, yet when our method is trained with simple resizing augmentation ("20+Aug"), its performance surpasses DE-FAKE at almost all scales. These results indicate that our method remains competitive under common post-processing and that robustness can be improved via light augmentation without modifying the core architecture.

Joint detection and attribution. To jointly detect synthetic images and attribute them to their source, our method adds an extra encoder E_{real} trained on natural images, yielding a feature vector $(r_1(x), \ldots, r_k(x), r_{\text{real}}(x))$. Table 6.5 reports results when classifying nine classes (eight generators plus "real"). Our method achieves the best detection-and-attribution performance across all shot counts, reaching 0.43 accuracy and 0.41 macro-F1 with 20 examples per class. The model attributes ADM and GLIDE with high precision, while diffusion models with similar architectures,



JPEG compression Resizing

Figure 6.2: Performance of our method (solid lines) and DE-FAKE (dashed lines) under JPEG compression (left) and resizing (right). The variant "20+Aug" includes random resizing augmentation during training.

Table 6.5: Joint detection and attribution performance. Accuracy and F1 are reported as mean ± standard deviation over five runs.

Training images per class		5	10	20
ResNet50	Acc.	0.16 ± 0.01	0.19 ± 0.01	0.23 ± 0.02
	F1	0.15 ± 0.01	0.19 ± 0.01	0.23 ± 0.02
VGG16	Acc.	0.16 ± 0.01	0.19 ± 0.00	0.23 ± 0.01
	F1	0.16 ± 0.01	0.19 ± 0.00	0.22 ± 0.01
XceptionNet	Acc.	0.13 ± 0.01	0.14 ± 0.01	0.17 ± 0.01
	F1	0.10 ± 0.02	0.12 ± 0.02	0.15 ± 0.00
ViT-B-16	Acc.	0.13 ± 0.00	0.14 ± 0.01	0.17 ± 0.01
	F1	0.13 ± 0.00	0.14 ± 0.01	0.17 ± 0.01
Swin-B	Acc.	0.14 ± 0.01	0.17 ± 0.01	0.22 ± 0.01
	F1	0.14 ± 0.01	0.16 ± 0.01	0.21 ± 0.01
Swin-T	Acc.	0.13 ± 0.01	0.16 ± 0.01	0.21 ± 0.01
	F1	0.13 ± 0.00	0.16 ± 0.01	0.21 ± 0.01
DE-FAKE	Acc.	0.21 ± 0.01	0.27 ± 0.02	0.36 ± 0.01
	F1	0.20 ± 0.01	0.26 ± 0.01	0.36 ± 0.01
Ours	Acc.	**0.28** ± 0.02	**0.33** ± 0.01	**0.43** ± 0.03
	F1	**0.26** ± 0.02	**0.32** ± 0.02	**0.41** ± 0.04

such as SD-v4 and SD-v5, exhibit greater confusion, reflecting the inherent difficulty of distinguishing closely related generators.

Impact of GLCM entropy. Table 6.6 compares patch selection based on Shannon entropy and on GLCM entropy. While both yield comparable performance with 5 shots, the texture-aware criterion offers small but consistent improvements when

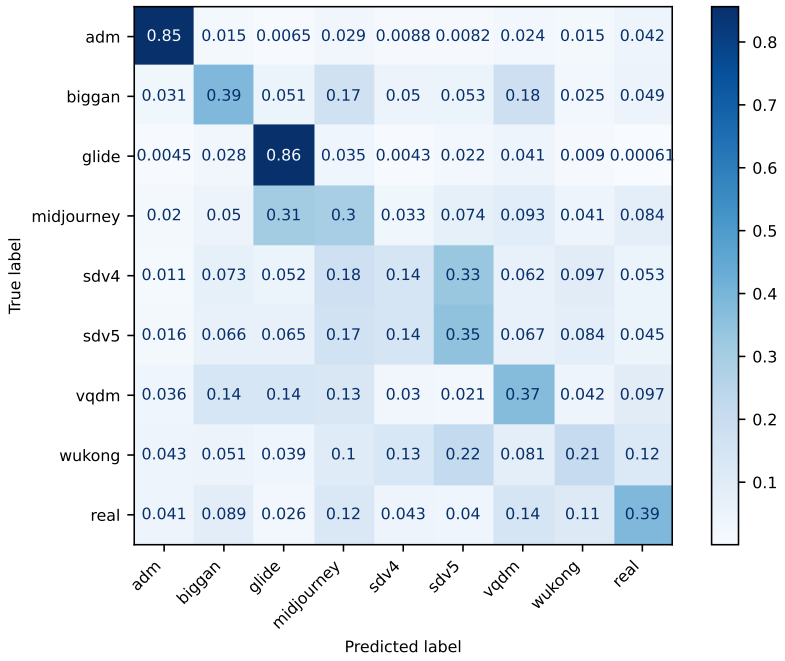


Figure 6.3: Confusion matrix of our method in the joint detection-and-attribution setting with 20 training images per class. The matrix shows that real images and several generators are identified reliably, while most residual confusion is concentrated among closely related diffusion-based models.

Table 6.6: Impact of the entropy criterion on joint detection-and-attribution. GLCM-based patch selection yields marginal gains at higher shot counts.

Entropy criterion	Metric	Training images per class		
		5	10	20
Shannon	Accuracy	0.28 ± 0.02	0.33 ± 0.01	0.43 ± 0.03
	F1	0.26 ± 0.02	0.32 ± 0.02	0.41 ± 0.04
GLCM	Accuracy	**0.28** ± 0.01	**0.37** ± 0.03	**0.44** ± 0.02
	F1	**0.28** ± 0.02	**0.36** ± 0.03	**0.42** ± 0.02

more training data are available, reaching 0.44 accuracy and 0.42 macro-F1 at 20 shots.

6.4 Discussion

The results show that compact autoencoders can provide an effective solution for few-shot source attribution when combined with a shared decoder and a lightweight discriminative layer. The main strength of the proposed approach lies not only in its competitive accuracy under severe data scarcity, but also in its modularity: each

new source can be incorporated by training only a dedicated encoder and updating the final classifier, without retraining the full system. This property is particularly relevant in realistic forensic scenarios, where the set of candidate generators evolves, and rapid adaptation is often more important than maximizing performance under large-scale supervised conditions. The confusion matrix further clarifies the behavior of the model: real images and several generators are identified reliably, while the remaining errors are concentrated among visually and architecturally similar diffusion-based models, where source-specific traces are naturally harder to disentangle. Robustness experiments under JPEG compression and resizing indicate that the method remains usable under common post-processing operations, and that simple task-aligned augmentation can further improve stability. More broadly, this chapter supports a methodological point that is consistent with the rest of the thesis: reconstruction-error-based pipelines can remain informative even outside standard anomaly detection, provided that the reconstruction space is structured to preserve class-dependent regularities. In this sense, the contribution of the chapter is not limited to multimedia forensics, but also reinforces the general utility of compact representation-and-score mechanisms in low-data and resource-constrained settings.

Chapter 7

Conclusions

This thesis addressed a common problem that appears in different forms across cardiology: how to assess whether an observation is compatible with normality when labels are scarce, patient populations are heterogeneous, and the consequences of overconfident decisions can be clinically relevant. The work approached this problem from two complementary directions. The first direction concerned electrogram signals, where normality cannot be described by a simple threshold and must instead be inferred from data through representation learning and anomaly scoring. The second direction concerned scalar clinical measurements whose interpretation depends on patient covariates, and where a useful decision-support tool should provide not only a score but also an indication of how reliable that score is in poorly covered regions of the reference population. Although the chapters adopt different modeling tools, they are connected by a common aspect: normality is treated as a data-driven concept, scores are preferred to rigid labels, and uncertainty is considered part of the output rather than an afterthought.

The thesis also included a complementary methodological chapter on tiny autoencoders for few-shot source attribution of AI-generated images. This chapter does not extend the clinical claims of the dissertation, yet it is not extraneous to its methodological core. It shows, in a different domain, that compact reconstruction-based representations can induce informative scores even when data are severely limited and when models must be updated incrementally.

7.1 Main findings of the thesis

The first main result of the thesis is that unsupervised deep anomaly detection can provide a meaningful and practically useful characterization of intracardiac electrogram morphology. In Chapter 3, reconstruction-based models and one-class objectives were used to derive anomaly scores directly from atrial EGMs, avoiding the need for manually assigned labels or manually tuned combinations of hand-

crafted indicators. The resulting scores were shown to correlate with established electrophysiological markers such as voltage, fractionation, and duration, while at the same time yielding spatial maps that are easier to interpret globally than a set of independent scalar biomarkers. From a clinical perspective, this matters because electroanatomical mapping is often interpreted through multiple indicators whose thresholds are partly conventional and whose joint reading can be cumbersome. The anomaly score does not eliminate the need for clinical interpretation, but it offers a coherent quantitative proxy for morphological abnormality and a potentially more reproducible substrate-oriented representation.

The second major result is methodological and concerns contextual anomaly detection. In Chapter 4, normalcy score was introduced as a probabilistic extension of classical Z-score reasoning. The main conceptual step was to treat the score itself as a random quantity induced by posterior uncertainty over the predictive mean and variance. This makes it possible to distinguish between two different sources of caution that are too often confounded in practice: intrinsic variability in the measurements and lack of knowledge due to limited or uneven reference data. By explicitly modeling both components, the method provides not only a central tendency of the score but also an interval-based description of reliability. This is important because many contextual anomaly detectors can rank examples, but do not communicate how much trust should be placed in a given ranking when the context is poorly supported by data. Normalcy score contributes both a modeling tool and a decision-support interface: it formalizes abnormality as deviation from context-dependent normality, and it makes visible when that assessment is itself uncertain.

The third major result is the translation of this uncertainty-aware viewpoint into a clinically established workflow. In Chapter 5, the classical Z-score for aortic diameters was reformulated in Bayesian terms and recast as a distribution rather than a deterministic number. This shift is modest in appearance but substantial in interpretation. Conventional Z-scores retain their value because they are simple and familiar to clinicians, yet they hide an important limitation: they implicitly assume that the reference equation used to standardize a diameter is itself known with sufficient precision. In practice, however, the available reference cohorts may be sparse or imbalanced in clinically relevant regions of the covariate space. The Bayesian Z-score addresses this limitation by providing both an expected score and a highest-density interval, thereby warning when the normalcy assessment is fragile. In the context of aortic surveillance, where decisions are often based on thresholds and repeated follow-up, this additional information can improve transparency and reduce the risk of false certainty.

A fourth result, complementary rather than central, concerns the role of reconstruction-based representations under severe data constraints. In Chapter 6,

compact autoencoders were used for few-shot source attribution of AI-generated images. The domain is different, but the methodological lesson is consistent with the rest of the thesis: reconstruction residuals can be turned into informative scores when the learned representation is sufficiently constrained and when the architecture is designed to preserve source-specific information. The chapter also highlights two practical themes that are relevant beyond multimedia forensics: data efficiency and modularity. These properties matter in clinical machine learning as well, where new acquisition settings, devices, or patient subgroups may appear over time and where retraining a large monolithic model can be impractical.

Taken together, these results support a broader claim. In label-scarce domains, useful decision-support tools do not necessarily arise from forcing a supervised labeling problem where labels are unstable or unavailable. They can instead be built by learning regularity from data, converting deviations from that regularity into scores, and making explicit when those scores should be trusted.

7.2 Methodological and clinical implications

A first methodological implication of the thesis is that score-based reasoning is often better suited than hard classification when the target concept is ambiguous. This is especially true in electrophysiology, where the notion of an abnormal EGM is not a universally agreed class label but a morphology-dependent judgment informed by multiple indicators and by spatial context. A score preserves ranking information, supports threshold selection downstream rather than upstream, and can be visualized in a way that is compatible with existing clinical workflows. The same argument applies to aortic normalcy assessment: what matters is not only whether a diameter crosses a threshold, but also how far it is from expected normality and how reliable that estimate is.

A second implication concerns uncertainty. In the thesis, uncertainty is not treated as a mere confidence decoration for a prediction. It has a more structural role: it tells the user whether the available data justify the interpretation of the score. This distinction is crucial in medical settings. An uncertain score does not necessarily indicate an abnormal patient; it may instead indicate that the model is extrapolating beyond the region where the reference data are informative. By surfacing this distinction, normalcy score and its application encourage a more conservative interaction between the model and the clinician. They are not only predictive tools but also instruments for communicating the limits of data-driven knowledge.

A third implication concerns deployability. Across the chapters, the methods were chosen not only for predictive performance but also for their ability to fit real constraints: limited data, moderate computational budgets, and the need for

interpretable outputs. This is visible in different ways. The EGM chapter compares models that produce direct anomaly scores from raw morphology rather than requiring a large manually engineered feature pipeline. The Bayesian chapters preserve the familiar language of Z-scores instead of replacing it with an entirely new abstraction. The tiny-autoencoder chapter shows that useful score-based systems can remain lightweight and modular even in class-incremental settings. These choices reflect a broader lesson: in translational machine learning, methodological elegance must be balanced with the practical conditions under which a system might actually be used.

From a clinical viewpoint, the thesis suggests that machine learning can contribute most effectively when it complements, rather than replaces, established reasoning. The EGM anomaly scores do not claim to supplant electrophysiological expertise; they provide a compact morphology-based lens that can be read alongside voltage and fractionation maps. Similarly, the Bayesian Z-score does not reject the clinical culture built around Z-score interpretation; it refines it by exposing when the underlying reference model is uncertain. In both cases, the goal is not automation for its own sake, but better-informed decision support.

7.3 Limitations and future perspectives

The thesis also has clear limitations, which should be acknowledged explicitly. The EGM study is based on a limited number of patients and, as is typical in this domain, it lacks a universally accepted ground truth for abnormality. Validation relies on indirect evidence, such as agreement with known biomarkers and spatial consistency on electroanatomical maps. These are meaningful signals, but they are not a substitute for large multicenter outcome-based validation. Future work should investigate whether anomaly-based substrate maps are associated with clinically relevant endpoints, such as arrhythmia recurrence after ablation or procedural success.

The contextual normalcy chapters inherit the assumptions of Gaussian-process modeling. These models are attractive because they are flexible, probabilistic, and well-suited to low- or moderate-dimensional covariate spaces, but they are not assumption-free. Their behavior depends on modeling choices and on the quality and representativeness of the reference data. The interval output of the score makes some of these limitations visible, but it does not remove them. In future work, it will be important to examine calibration under dataset shift, multicenter transferability, and the use of richer covariate sets without losing interpretability.

These limitations point to several promising directions. On the anomaly-detection side, one natural next step is to integrate morphology-based scores with spatial information and clinical metadata, so that local waveform anomalies

can be interpreted relative to patient-specific context. On the normalcy-assessment side, future work could extend scalar scores toward multivariate and longitudinal settings, where changes over time may be as important as deviation at a single visit. More broadly, there is room for decision-theoretic formulations in which thresholds are chosen not only from statistical criteria but also from explicit clinical costs. Such developments would move the thesis contributions from score estimation toward full risk-aware decision support.

Another promising direction concerns the interaction between anomaly detection and selective prediction. A score may be informative enough to support ranking, yet not reliable enough to justify a categorical decision on every patient or signal. Future models could combine abnormality estimation with explicit abstention policies, referral-to-expert mechanisms, or acquisition-aware confidence models. This is particularly relevant in medicine, where the most appropriate output of an algorithm is sometimes not a recommendation, but a warning that the available evidence is insufficient for a confident judgment.

The thesis deliberately focused on score construction and interpretation rather than on large-scale multicenter adaptation. Yet many clinically important sources of variability arise precisely at that level: electrode properties, filtering pipelines, scanner settings, institutional referral bias, or center-specific measurement conventions. Understanding how anomaly and normalcy scores behave under such shifts will be crucial for genuine clinical translation.

Appendix A

Publications

This chapter lists publications produced during the doctoral programme. The first group is directly related to the scientific content presented in this thesis. Additional publications not discussed in the main chapters are listed separately.

Publications directly related to this thesis

Journal papers

1. **Luca Bindini**, S. Pagani, A. Bernardini, B. Grossi, A. Giomi, A. Frontera, P. Frasconi, "All-in-one electrical atrial substrate indicators with deep anomaly detection", *Biomedical Signal Processing and Control*, vol. 98, Elsevier, 2024.

   `https://www.sciencedirect.com/science/article/pii/S174680942400795X`

   Candidate's contributions: developed deep anomaly detection algorithms, implemented experimental evaluation, and wrote the original draft.

2. **Luca Bindini**, L. Perini, S. Nistri, J. Davis, P. Frasconi, "Dealing with Uncertainty in Contextual Anomaly Detection", *Transactions on Machine Learning Research (TMLR)*, Journal of Machine Learning Research Inc., 2026.

   `https://openreview.net/forum?id=yLoXQDNwwa`

   Candidate's contributions: proposed novel uncertainty-aware anomaly detection methods, implemented experiments, and wrote the original draft.

3. **Luca Bindini**, L. Campens, J. Davis, L. Muino-Mosquera, S. D'hulst, J. De Backer, S. Nistri, P. Frasconi, "Towards a more reliable assessment of aortic diameters using a Bayesian Z-score", *Scientific Reports*, Nature Portfolio, 2026.

   `https://www.nature.com/articles/s41598-026-46006-x`

Candidate's contributions: developed the Bayesian Z-score model, implemented Gaussian process regression framework, designed and ran experiments, analyzed data, and wrote the original draft.

4. **Luca Bindini**, G. Bertazzini, D. Baracchi, D. Shullani, P. Frasconi, A. Piva, "Tiny Autoencoders for Few-Shot Source Attribution of AI-Generated Images", accepted for publication in *Journal on Information Security*, Springer, 2026.
   Candidate's contributions: conceived and implemented the framework, designed and conducted experiments, analyzed results, and wrote the original draft.

Peer reviewed conference and workshop papers

1. **Luca Bindini**, G. Bertazzini, D. Baracchi, D. Shullani, P. Frasconi, A. Piva, "Tiny autoencoders are effective few-shot generative model detectors", *IEEE Workshop on Information Forensics and Security (WIFS)*, IEEE, 2024.

   `https://ieeexplore.ieee.org/abstract/document/10810686`

   Candidate's contributions: developed and evaluated autoencoder architectures for deepfake attribution and wrote the original draft.

Other publications (not discussed in the thesis)

1. A. Bernardini, **Luca Bindini**, E. Antonucci, M. Berteotti, B. Giusti, S. Testa, G. Palareti, D. Poli, P. Frasconi, R. Marcucci, "Machine learning approach for prediction of outcomes in anticoagulated patients with atrial fibrillation", *International Journal of Cardiology*, vol. 407, Elsevier, 2024.

   `https://www.sciencedirect.com/science/article/abs/pii/S0167527324007101`

2. **Luca Bindini**, S. Giovannini, S. Marinai, V. Nardoni, K. Noor Ali, "Hierarchical structure understanding in complex tables with VLLMs: a benchmark and experiments", *GREC @ International Conference on Document Analysis (ICDAR) and Recognition*, Springer, 2025.

   `https://link.springer.com/chapter/10.1007/978-3-032-09371-4_1`

3. **Luca Bindini***, E. Ristori*, P. Frasconi, "Next-Scale Autoregression on Spectrograms for Sound Generation", *Multimodal Intelligence @ International Conference on Learning Representations (ICLR)*, OpenReview, 2026.

   `https://openreview.net/forum?id=nLr5craHWh`

4. **Luca Bindini***, E. Ristori*, P. Frasconi, "MARS: Sound Generation via Multi-Channel Autoregression on Spectrograms", *International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2026.

*Equal contribution.

Bibliography

Aggarwal, C. C. (2017). *Outlier ensembles*. Springer.

Albright, M., McCloskey, S., and Honeywell, A. (2019). Source generator attribution via inversion. In *CVPR workshops*, volume 8, page 3.

Altman, D. G. (1993). Construction of age-related reference centiles using absolute residuals. *Statistics in Medicine*, 12(10):917–924.

Ammer, T., Schützenmeister, A., Prokosch, H.-U., Rauh, M., Rank, C. M., and Zierk, J. (2023). A pipeline for the fully automated estimation of continuous reference intervals using real-world data. *Scientific Reports*, 13(1):13440.

An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special lecture on IE*, 2(1):1–18.

Anter, E. and Josephson, M. E. (2016). Bipolar voltage amplitude: what does it really mean? *Heart Rhythm*, 13(1):326–327.

Asch, F. M., Banchs, J., Price, R., Rigolin, V., Thomas, J. D., Weissman, N. J., and Lang, R. M. (2019). Need for a global definition of normative echo values—rationale and design of the world alliance of societies of echocardiography normal values study (wase). *Journal of the American Society of Echocardiography*, 32(1):157–162.e2.

Ballabio, D., Cassotti, M., Consonni, V., and Todeschini, R. (2015). QSAR fish toxicity. UCI Machine Learning Repository.

Bernardini, A., Bindini, L., Antonucci, E., Berteotti, M., Giusti, B., Testa, S., Palareti, G., Poli, D., Frasconi, P., and Marcucci, R. (2024). Machine learning approach for prediction of outcomes in anticoagulated patients with atrial fibrillation. *International Journal of Cardiology*, 407. Publisher: Elsevier.

Bindini, L., Bertazzini, G., Baracchi, D., Shullani, D., Frasconi, P., and Piva, A. (2024a). Tiny Autoencoders are Effective Few-Shot Generative Model Detectors. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. ISSN: 2157-4774.

Bindini, L., Campens, L., Davis, J., Muiño-Mosquera, L., D'hulst, S., De Backer, J., Nistri, S., and Frasconi, P. (2026a). Towards a more reliable assessment of aortic diameters using a bayesian z-score. *Scientific Reports*.

Bindini, L., Pagani, S., Bernardini, A., Grossi, B., Giomi, A., Frontera, A., and Frasconi, P. (2024b). All-in-one electrical atrial substrate indicators with deep anomaly detection. *Biomedical Signal Processing and Control*, 98:106737.

Bindini, L., Perini, L., Nistri, S., Davis, J., and Frasconi, P. (2026b). Dealing with uncertainty in contextual anomaly detection. *Transactions on Machine Learning Research*.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In Bach, F. R. and Blei, D. M., editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1613–1622. JMLR.org.

Bonilla, E. V., Chai, K., and Williams, C. (2007). Multi-task Gaussian process prediction. In *Adv. Neural Inf. Process. Syst.*, volume 20.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

Boyle, P. M., Zghaib, T., Zahid, S., Ali, R. L., Deng, D., Franceschi, W. H., Hakim, J. B., Murphy, M. J., Prakosa, A., Zimmerman, S. L., et al. (2019). Computationally guided personalized targeted ablation of persistent atrial fibrillation. *Nature biomedical engineering*, 3(11):870–879.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). LOF: identifying density-based local outliers. In *Proc. of the 2000 ACM SIGMOD Int. Conf. on Management of data*, pages 93–104.

Campens, L., Demulier, L., De Groote, K., Vandekerckhove, K., De Wolf, D., Roman, M. J., Devereux, R. B., De Paepe, A., and De Backer, J. (2014). Reference values for echocardiographic assessment of the diameter of the aortic root and ascending aorta spanning all age categories. *The American Journal of Cardiology*, 114(6):914–920.

Cantinotti, M., Giordano, R., Scalese, M., Murzi, B., Assanta, N., Spadoni, I., Maura, C., Marco, M., Molinaro, S., Kutty, S., and Iervasi, G. (2017). Nomograms for two-dimensional echocardiography derived valvular and arterial dimensions in caucasian children. *Journal of Cardiology*, 69(1):208–215.

Ceriotti, F. (2012). Establishing pediatric reference intervals: A challenging task. *Clinical Chemistry*, 58(5):808–810.

Chalapathy, R. and Chawla, S. (2019). Deep Learning for Anomaly Detection: A Survey.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):15:1–15:58.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

Chubb, H. and Simpson, J. M. (2012). The use of z-scores in paediatric cardiology. *Annals of Pediatric Cardiology*, 5(2):179.

Colan, S. D. (2013). The why and how of z scores. *Journal of the American Society of Echocardiography*, 26(1):38–40.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20:273–297.

Curtis, A., Smith, T., Ziganshin, B., and Elefteriades, J. (2016). The mystery of the z-score. *AORTA*, 04(04):124–130.

Dallaire, F., Bigras, J.-L., Prsa, M., and Dahdah, N. (2015). Bias related to body mass index in pediatric echocardiographic z scores. *Pediatric Cardiology*, 36(3):667–676.

Daubeney, P. E. F., Blackstone, E. H., Weintraub, R. G., Slavik, Z., Scanlon, J., and Webber, S. A. (1999). Relationship of the dimension of cardiac structures to body size: An echocardiographic study in normal infants and children. *Cardiology in the Young*, 9(4):402–410.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., and Udluft, S. (2018). Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pages 1184–1193. PMLR.

Devereux, R. B., De Simone, G., Arnett, D. K., Best, L. G., Boerwinkle, E., Howard, B. V., Kitzman, D., Lee, E. T., Mosley, T. H., Weder, A., and Roman, M. J. (2012). Normal limits in relation to age, body size and gender of two-dimensional echocardiographic aortic root dimensions in persons $\geq$15 years of age. *The American Journal of Cardiology*, 110(8):1189–1194.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29.

Feeny, A. K., Chung, M. K., Madabhushi, A., Attia, Z. I., Cikes, M., Firouznia, M., Friedman, P. A., Kalscheur, M. M., Kapa, S., Narayan, S. M., Noseworthy, P. A., Passman, R. S., Perez, M. V., Peters, N. S., Piccini, J. P., Tarakji, K. G., Thomas, S. A., Trayanova, N. A., Turakhia, M. P., and Wang, P. J. (2020). Artificial Intelligence and Machine Learning in Arrhythmias and Cardiac Electrophysiology. *Circulation: Arrhythmia and Electrophysiology*, 13(8):e007952.

Fernandes, S., Khairy, P., Graham, D. A., Colan, S. D., Galvin, T. C., Sanders, S. P., Singh, M. N., Bhatt, A., and Lacro, R. V. (2012). Bicuspid aortic valve and associated aortic dilation in the young. *Heart*, 98(13):1014–1019.

Frasconi, P., Baracchi, D., Giusti, B., Kura, A., Spaziani, G., Cherubini, A., Favilli, S., Di Lenarda, A., Pepe, G., and Nistri, S. (2021). Two-Dimensional Aortic Size Normalcy: A Novelty Detection Approach. *Diagnostics*, 11(2):220.

Frontera, A., Limite, L. R., Pagani, S., Cireddu, M., Vlachos, K., Martin, C., Takigawa, M., Kitamura, T., Bourier, F., Cheniti, G., Pambrun, T., Sacher, F., Derval, N., Hocini, M., Quarteroni, A., Della Bella, P., Haissaguerre, M., and Jaïs, P. (2022a). Electrogram fractionation during sinus rhythm occurs in normal voltage atrial tissue in patients with atrial fibrillation. *Pacing and Clinical Electrophysiology*, 45(2):219–228.

Frontera, A., Limite, L. R., Pagani, S., Hadjis, A., Cireddu, M., Sala, S., Tsitsinakis, G., Paglino, G., Peretto, G., Lipartiti, F., et al. (2021). Characterization of cardiac electrogram signals in atrial arrhythmias. *Minerva Cardiology and Angiology*, 69(1):70–80.

Frontera, A., Mahajan, R., Dallet, C., Vlachos, K., Kitamura, T., Takigawa, M., Cheniti, G., Martin, C., Duchateau, J., Lam, A., et al. (2019). Characterizing

localized reentry with high-resolution mapping: evidence for multiple slow conducting isthmuses within the circuit. *Heart Rhythm*, 16(5):679–685.

Frontera, A., Pagani, S., Limite, L. R., Peirone, A., Fioravanti, F., Enache, B., Cuellar Silva, J., Vlachos, K., Meyer, C., Montesano, G., Manzoni, A., Dedé, L., Quarteroni, A., Lațcu, D. G., Rossi, P., and Della Bella, P. (2022b). Slow Conduction Corridors and Pivot Sites Characterize the Electrical Remodeling in Atrial Fibrillation. *JACC: Clinical Electrophysiology*, 8(5):561–577.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Balcan, M. and Weinberger, K. Q., editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org.

Gao, C., Xu, Q., Qiao, P., Xu, K., Qian, X., and Dou, Y. (2024). Adapter-based incremental learning for face forgery detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4690–4694. IEEE.

Gautier, M., Detaint, D., Fermanian, C., Aegerter, P., Delorme, G., Arnoult, F., Milleron, O., Raoux, F., Stheneur, C., Boileau, C., Vahanian, A., and Jondeau, G. (2010). Nomograms for aortic root diameters in children using two-dimensional echocardiography. *The American Journal of Cardiology*, 105(6):888–894.

Gerritsma, J., Onnink, R., and Versluis, A. (1981). Yacht Hydrodynamics. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XG7R.

Gidaris, S., Singh, P., and Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Goldberg, P. W., Williams, C. K., and Bishop, C. M. (1997). Regression with input-dependent noise: A Gaussian process treatment. In *Adv. Neural Inf. Process. Syst.*, volume 10, pages 493–499.

Goldstein, H. (1972). The construction of standards for measurements subject to growth. *Human Biology*, 44(2):255–261.

Goldstein, M. and Dengel, A. (2012). Histogram-based outlier score (HBOS): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track*, 1:59–63.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., and Van Den Hengel, A. (2019). Memorizing Normality to Detect Anomaly: Memory-Augmented Deep Autoencoder for Unsupervised Anomaly Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1705–1714, Seoul, Korea (South). IEEE.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680.

Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. (2020). Bootstrap your own latent - A new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Haïssaguerre, M., Jaïs, P., Shah, D. C., Takahashi, A., Hocini, M., Quiniou, G., Garrigue, S., Le Mouroux, A., Le Métayer, P., and Clémenty, J. (1998). Spontaneous initiation of atrial fibrillation by ectopic beats originating in the pulmonary veins. *New England Journal of Medicine*, 339(10):659–666.

Haralick, R. M., Shanmugam, K., and Dinstein, I. H. (2007). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621.

Hauskrecht, M., Valko, M., Kveton, B., Visweswaran, S., and Cooper, G. F. (2007). Evidence-based anomaly detection in clinical domains. *AMIA Annual Symposium Proceedings*, 2007:319–323.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. B. (2022). Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. (2020). Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. Computer Vision Foundation / IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Healy, M. J. R. (1978). Statistics of growth standards. In *Principles and Prenatal Growth*, pages 169–181. Springer US, Boston, MA.

Heaukulani, C. and van der Wilk, M. (2019). Scalable bayesian dynamic covariance modeling with variational wishart and inverse wishart processes. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Hindricks, G., Potpara, T., Dagres, N., Arbelo, E., Bax, J. J., Blomström-Lundqvist, C., Boriani, G., Castella, M., Dan, G.-A., Dilaveris, P. E., Fauchier, L., Filippatos, G., Kalman, J. M., La Meir, M., Lane, D. A., Lebeau, J.-P., Lettino, M., Lip, G. Y. H., Pinto, F. J., Thomas, G. N., Valgimigli, M., Van Gelder, I. C., Van Putte, B. P., Watkins, C. L., ESC Scientific Document Group, Kirchhof, P., Kühne, M., Aboyans, V., Ahlsson, A., Balsam, P., Bauersachs, J., Benussi, S., Brandes, A., Braunschweig, F., Camm, A. J., Capodanno, D., Casadei, B., Conen, D., Crijns, H. J. G. M., Delgado, V., Dobrev, D., Drexel, H., Eckardt, L., Fitzsimons, D., Folliguet, T., Gale, C. P., Gorenek, B., Haeusler, K. G., Heidbuchel, H., Iung, B., Katus, H. A., Kotecha, D., Landmesser, U., Leclercq, C., Lewis, B. S., Mascherbauer, J., Merino, J. L., Merkely, B., Mont, L., Mueller, C., Nagy, K. V., Oldgren, J., Pavlović, N., Pedretti, R. F. E., Petersen, S. E., Piccini, J. P., Popescu, B. A., Pürerfellner, H., Richter, D. J., Roffi, M., Rubboli, A., Scherr, D., Schnabel, R. B., Simpson, I. A., Shlyakhto, E., Sinner, M. F., Steffel, J., Sousa-Uva, M., Suwalski, P., Svetlosak, M., Touyz, R. M., Dagres, N., Arbelo, E., Bax, J. J., Blomström-Lundqvist, C., Boriani, G., Castella, M., Dan, G.-A., Dilaveris, P. E., Fauchier, L., Filippatos, G., Kalman, J. M., La Meir, M., Lane, D. A., Lebeau, J.-P., Lettino, M., Lip, G. Y. H., Pinto, F. J., Neil Thomas, G., Valgimigli, M., Van Gelder, I. C., Watkins, C. L., Delassi, T., Sisakian, H. S., Scherr, D., Chasnoits, A., Pauw, M. D., Smajić, E., Shalganov, T., Avraamides, P., Kautzner, J., Gerdes, C., Alaziz, A. A., Kampus, P., Raatikainen, P., Boveda, S., Papiashvili, G., Eckardt, L., Vassilikos, V., Csanádi, Z., Arnar, D. O., Galvin, J., Barsheshet, A., Caldarola, P., Rakisheva, A., Bytyçi, I., Kerimkulova, A., Kalejs, O., Njeim, M., Puodziukynas, A., Groben, L., Sammut, M. A., Grosu, A., Boskovic, A., Moustaghfir, A., Groot, N. D., Poposka, L., Anfinsen, O.-G., Mitkowski, P. P., Cavaco, D. M., Siliste, C., Mikhaylov, E. N., Bertelli, L., Kojic, D., Hatala, R., Fras, Z., Arribas, F., Juhlin, T., Sticherling, C., Abid, L., Atar, I., Sychov, O., Bates, M. G. D., and Zakirov, N. U. (2021). 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration

with the European Association for Cardio-Thoracic Surgery (EACTS). *European Heart Journal*, 42(5):373–498.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Hirofumi, S., Fukuchi, K., Akimoto, Y., and Sakuma, J. (2022). Did you use my gan to generate fake? post-hoc attribution of gan generated images via latent recovery. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.*, 110(3):457–506.

Isselbacher, E. M., Preventza, O., others, and Woo, Y. J. (2022). 2022 ACC/AHA guideline for the diagnosis and management of aortic disease: A report of the american heart association/american college of cardiology joint committee on clinical practice guidelines. *Circulation*, 146(24):e334–e482.

Jadidi, A. S., Duncan, E., Miyazaki, S., Lellouche, N., Shah, A. J., Forclaz, A., Nault, I., Wright, M., Rivard, L., Liu, X., et al. (2012). Functional nature of electrogram fractionation demonstrated by left atrial high-density mapping. *Circulation: Arrhythmia and Electrophysiology*, 5(1):32–42.

Jeon, H., Bang, Y., Kim, J., and Woo, S. S. (2020). T-gd: Transferable gan-generated images detection framework. *arXiv preprint arXiv:2008.04115*.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):195.

Kendall, A. and Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, volume 30.

Kersting, K., Plagemann, C., Pfaff, P., and Burgard, W. (2007). Most likely heteroscedastic gaussian process regression. In *Proceedings of the 24th International Conference on Machine Learning*, pages 393–400. ACM.

Kim, J. B., Spotnitz, M., Lindsay, M. E., MacGillivray, T. E., Isselbacher, E. M., and Sundt, T. M. (2016). Risk of aortic dissection in the moderately dilated ascending aorta. *Journal of the American College of Cardiology*, 68(11):1209–1219.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y., editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kiureghian, A. D. and Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112.

Konings, K., Kirchhof, C., Smeets, J., Wellens, H., Penn, O. C., and Allessie, M. A. (1994). High-density mapping of electrically induced atrial fibrillation in humans. *Circulation*, 89(4):1665–1680.

Kruschke, J. K. (2015). *Doing Bayesian Data Analysis*. Academic Press, Boston, 2nd edition.

Kuo, Y.-H., Li, Z., and Kifer, D. (2018). Detecting outliers in data with correlated measures. In *Proc. of the 27th ACM Int. Conf. on information and knowledge management*, pages 287–296.

La Rosa, G., Quintanilla, J. G., Salgado, R., González-Ferrer, J. J., Cañadas-Godoy, V., Pérez-Villacastín, J., Jalife, J., Pérez-Castellano, N., and Filgueiras-Rama, D. (2021). Anatomical targets and expected outcomes of catheter-based ablation of atrial fibrillation in 2020. *Pacing and Clinical Electrophysiology*, 44(2):341–359.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413.

Lancellotti, P., Badano, L. P., others, and Maurer, G. (2013). Normal reference ranges for echocardiography: Rationale, study design, and methodology (norre study). *European Heart Journal - Cardiovascular Imaging*, 14(4):303–308.

Lau, D. H., Maesen, B., Zeemering, S., Kuklik, P., van Hunnik, A., Lankveld, T. A. R., Bidar, E., Verheule, S., Nijs, J., Maessen, J., Crijns, H., Sanders, P., and Schotten, U. (2015). Indices of bipolar complex fractionated atrial electrograms correlate poorly with each other and atrial fibrillation substrate complexity. *Heart Rhythm*, 12(7):1415–1423.

Le, Q. V., Smola, A. J., and Canu, S. (2005). Heteroscedastic Gaussian process regression. In *Proc. of the 22nd Int. Conf. on Machine Learning - ICML '05*, pages 489–496. ACM Press.

Lee, W., Riggs, T., Amula, V., Tsimis, M., Cutler, N., Bronsteen, R., and Comstock, C. H. (2010). Fetal echocardiography: Z-score reference ranges for a large patient population. *Ultrasound in Obstetrics & Gynecology*, 35(1):28–34.

Li, Z. and van Leeuwen, M. (2023). Explainable contextual anomaly detection using quantile regression forests. *Data Mining and Knowledge Discovery*, 37(6):2517–2563.

Liang, J. and Parthasarathy, S. (2016). Robust Contextual Outlier Detection: Where Context Meets Sparsity. In *Proc. of the 25th ACM Int. Conf. on Information and Knowledge Management*, pages 2167–2172. ACM.

Liao, S., Ragot, D., Nayyar, S., Suszko, A., Zhang, Z., Wang, B., and Chauhan, V. S. (2021). Deep Learning Classification of Unipolar Electrograms in Human Atrial Fibrillation: Application in Focal Source Mapping. *Frontiers in Physiology*, 12.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *Eighth IEEE Int. Conf. on Data Mining*, pages 413–422.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Lopez, L., Colan, S., Stylianou, M., Granger, S., Trachtenberg, F., Frommelt, P., Pearson, G., Camarda, J., Cnota, J., Cohen, M., Dragulescu, A., Frommelt, M., Garuba, O., Johnson, T., Lai, W., Mahgerefteh, J., Pignatelli, R., Prakash, A., Sachdeva, R., Soriano, B., Soslow, J., Spurney, C., Srivastava, S., Taylor, C., Thankavel, P., van der Velde, M., and Minich, L. (2017). Relationship of echocardiographic z scores adjusted for body surface area to age, sex, race, and ethnicity. *Circulation: Cardiovascular Imaging*, 10(11):e006979.

Magistri, S., Baracchi, D., Shullani, D., Bagdanov, A. D., and Piva, A. (2023). Towards continual social network identification. In *2023 11th International Workshop on Biometrics and Forensics*, pages 1–6. IEEE.

Magistri, S., Baracchi, D., Shullani, D., Bagdanov, A. D., and Piva, A. (2024). Continual learning for adaptive social network identification. *Pattern Recognition Letters*, 180:82–89.

Markou, M. and Singh, S. (2003). Novelty detection: A review—part 2: Neural network based approaches. *Signal Processing*, 83(12):2499–2521.

Marra, F., Saltori, C., Boato, G., and Verdoliva, L. (2019). Incremental learning for the detection and classification of gan-generated images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE.

Martinez-Millana, A., Hulst, J. M., Boon, M., Witters, P., Fernandez-Llatas, C., Asseiceira, I., Calvo-Lerma, J., Basagoiti, I., Traver, V., Boeck, K. D., and Ribes-Koninckx, C. (2018). Optimisation of children z-score calculation based on new statistical techniques. *PLOS ONE*, 13(12):e0208362.

Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J. (2017). GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 18(40):1–6.

Mawad, W., Drolet, C., Dahdah, N., and Dallaire, F. (2013). A review and critique of the statistical methods used to generate reference values in pediatric echocardiography. *Journal of the American Society of Echocardiography*, 26(1):29–37.

Meinshausen, N. and Ridgeway, G. (2006). Quantile regression forests. *Journal of Machine Learning Research*, 7(6).

Mendonca Costa, C., Anderson, G. C., Meijborg, V. M., O'Shea, C., Shattock, M. J., Kirchhof, P., Coronel, R., Niederer, S., Pavlovic, D., Dhanjal, T., et al. (2020). The amplitude-normalized area of a bipolar electrogram as a measure of local conduction delay in the heart. *Frontiers in Physiology*, 11:465.

Nademanee, K., McKenzie, J., Kosar, E., Schwab, M., Sunsaneewitayakul, B., Vasavakul, T., Khunnawat, C., and Ngarmukos, T. (2004). A new approach for catheter ablation of atrial fibrillation: Mapping of the electrophysiologic substrate. *Journal of the American College of Cardiology*, 43(11):2044–2053.

Nash, W., Sellers, T., Talbot, S., Cawthorn, A., and Ford, W. (1994). Abalone. UCI Machine Learning Repository.

Noroozi, M. and Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, volume 9910 of *Lecture Notes in Computer Science*, pages 69–84. Springer.

Pagani, S., Dede', L., Frontera, A., Salvador, M., Limite, L. R., Manzoni, A., Lipartiti, F., Tsitsinakis, G., Hadjis, A., Della Bella, P., and Quarteroni, A. (2021). A computational study of the electrophysiological substrate in patients suffering from atrial fibrillation. *Frontiers in Physiology*, 12.

Pan, K., Yin, Y., Wei, Y., Lin, F., Ba, Z., Liu, Z., Wang, Z., Cavallaro, L., and Ren, K. (2023). Dfil: Deepfake incremental learning by exploiting domain-invariant forgery clues. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8035–8046.

Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. (2021). Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 54(2):38:1–38:38. 00650.

Parameswaran, R., Al-Kaisey, A. M., and Kalman, J. M. (2021). Catheter ablation for atrial fibrillation: current indications and evolving technologies. *Nature Reviews Cardiology*, 18(3):210–225.

Patel, H. N., Miyoshi, T., Addetia, K., Citro, R., Daimon, M., Fajardo, P. G., Kasliwal, R. R., Kirkpatrick, J. N., Monaghan, M. J., Muraru, D., Ogunyankin, K. O., Park, S. W., Ronderos, R. E., Sadeghpour, A., Scalia, G. M., Takeuchi, M., Tsang, W., Tucay, E. S., Rodrigues, A. C. T., Amuthan, V., Zhang, Y., Schreckenberg, M., Blankenhagen, M., Degel, M., Hitschrich, N., Mor-Avi, V., Asch, F. M., and Lang, R. M. (2022a). Normal values of aortic root size according to age, sex, and race: Results of the world alliance of societies of echocardiography study. *Journal of the American Society of Echocardiography*, 35(3):267–274.

Patel, Z. B., Batra, N., and Murphy, K. (2022b). Uncertainty disentanglement with non-stationary heteroscedastic Gaussian processes for active learning. In *NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems*.

Perini, L. and Davis, J. (2023). Unsupervised anomaly detection with rejection. *Adv. Neural Inf. Process. Syst.*, 36:69673–69691.

Piccini, J. P., Lopes, R. D., Kong, M. H., Hasselblad, V., Jackson, K., and Al-Khatib, S. M. (2009). Pulmonary Vein Isolation for the Maintenance of Sinus Rhythm in Patients With Atrial Fibrillation: A Meta-Analysis of Randomized, Controlled Trials. *Circulation: Arrhythmia and Electrophysiology*, 2(6):626–633.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2016). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ramirez, F. D., Birnie, D. H., Nair, G. M., Szczotka, A., Redpath, C. J., Sadek, M. M., and Nery, P. B. (2017). Efficacy and safety of driver-guided catheter ablation for atrial fibrillation: A systematic review and meta-analysis. *Journal of Cardiovascular Electrophysiology*, 28(12):1371–1378.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Mass.

Ricci, F., Aung, N., others, and Petersen, S. E. (2021). Cardiovascular magnetic resonance reference values of mitral and tricuspid annular dimensions: The uk biobank cohort. *Journal of Cardiovascular Magnetic Resonance*, 23(1):5.

Ricker, J., Lukovnikov, D., and Fischer, A. (2024). Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9130–9140.

Rodrigo, M., Guillem, M. S., Climent, A. M., Pedrón-Torrecilla, J., Liberos, A., Millet, J., Fernández-Avilés, F., Atienza, F., and Berenfeld, O. (2014). Body surface localization of left and right atrial high-frequency rotors in atrial fibrillation patients: A clinical-computational study. *Heart Rhythm*, 11(9):1584–1591.

Roman, M. J., Devereux, R. B., Kramer-Fox, R., and O'Loughlin, J. (1989). Two-dimensional echocardiographic aortic root dimensions in normal children and adults. *The American journal of cardiology*, 64(8):507–512. 00846.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Rossi, P., Cauti, F. M., Niscola, M., Calore, F., Fanti, V., Polselli, M., Di Pastena, A., Iaia, L., and Bianchi, S. (2021). A novel ventricular map of electrograms duration as a method to identify areas of slow conduction for ventricular tachycardia ablation: The vedum pilot study. *Heart Rhythm*, 18(8):1253–1260.

Royston, P. (1991). Constructing time-specific reference ranges. *Statistics in Medicine*, 10(5):675–690.

Ruff, L., Görnitz, N., Deecke, L., Siddiqui, S. A., Vandermeulen, R. A., Binder, A., Müller, E., and Kloft, M. (2018). Deep one-class classification. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4390–4399. PMLR.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., and Müller, K.-R. (2021). A unifying review of deep and shallow anomaly detection. *Proc. IEEE*, 109(5):756–795.

Saul, A. D., Hensman, J., Vehtari, A., and Lawrence, N. D. (2016). Chained gaussian processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1431–1440. PMLR.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In Niethammer, M., Styner, M., Aylward, S. R., Zhu, H., Oguz, I., Yap, P., and Shen, D., editors, *Information Processing in Medical Imaging - 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings*, volume 10265 of *Lecture Notes in Computer Science*, pages 146–157. Springer.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471.

Senge, R., Bösner, S., Dembczyński, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., and Hüllermeier, E. (2014). Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29.

Sha, Z., Li, Z., Yu, N., and Zhang, Y. (2023). De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3418–3432.

Shieh, G. S. (1998). A weighted kendall's tau statistic. *Statistics & Probability Letters*, 39(1):17–24.

Sim, I., Bishop, M., O'Neill, M., and Williams, S. E. (2019). Left atrial voltage mapping: defining and targeting the atrial fibrillation substrate. *Journal of Interventional Cardiac Electrophysiology*, 56:213–227.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sluysmans, T. and Colan, S. D. (2016). Structural measurements and adjustments for growth. In *Echocardiography in Pediatric and Congenital Heart Disease*, chapter 5, pages 61–72. John Wiley & Sons, Ltd.

Snelson, E. and Ghahramani, Z. (2005). Sparse gaussian processes using pseudo-inputs. *Adv. Neural Inf. Process. Syst.*, 18.

Song, X., Wu, M., Jermaine, C., and Ranka, S. (2007). Conditional Anomaly Detection. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):631–645.

Tang, G., Bailey, J., Pei, J., and Dong, G. (2013). Mining multidimensional contextual outliers from categorical relational data. In *Proc. of the 25th Int. Conf. on Scientific and Statistical Database Management*, pages 1–4. ACM.

Tassone, F., Maiano, L., and Amerini, I. (2024). Continuous fake media detection: adapting deepfake detectors to new generative techniques. *arXiv preprint arXiv:2406.08171*.

Tax, D. and Duin, R. (2004). Support vector data description. *Machine Learning*, 54(1):45–66.

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *Artificial intelligence and statistics*, pages 567–574.

UCI (2021). Synchronous Machine. UCI Machine Learning Repository.

Valko, M., Kveton, B., Valizadegan, H., Cooper, G. F., and Hauskrecht, M. (2011). Conditional anomaly detection with soft harmonic functions. In *2011 IEEE 11th Int. Conf. on Data Mining*, pages 735–743.

van Kimmenade, R. R. J., Kempers, M., de Boer, M.-J., Loeys, B. L., and Timmermans, J. (2013). A clinical appraisal of different Z -score equations for aortic root assessment in the diagnostic evaluation of Marfan syndrome. *Genetics in Medicine*, 15(7):528–532.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103.

Vogler, J., Willems, S., Sultan, A., Schreiber, D., Lüker, J., Servatius, H., Schäffer, B., Moser, J., Hoffmann, B. A., and Steven, D. (2015). Pulmonary vein isolation versus defragmentation: the chase-af clinical trial. *Journal of the American College of Cardiology*, 66(24):2743–2752.

Vriz, O., Aboyans, V., D'Andrea, A., Ferrara, F., Acri, E., Limongelli, G., Della Corte, A., Driussi, C., Bettio, M., Pluchinotta, F. R., Citro, R., Russo, M. G., Isselbacher, E., and Bossone, E. (2014). Normal values of aortic root dimensions in healthy adults. *The American Journal of Cardiology*, 114(6):921–927.

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., and Li, H. (2023). Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455.

WHO Multicentre Growth Reference Study Group (2006). WHO child growth standards based on length/height, weight and age. *Acta Paediatrica*, 95(S450):76–85.

Williams, S. E., Roney, C. H., Connolly, A., Sim, I., Whitaker, J., O'Hare, D., Kotadia, I., O'Neill, L., Corrado, C., Bishop, M., Niederer, S. A., Wright, M., O'Neill, M., and Linton, N. W. F. (2021). Openep: A cross-platform electroanatomic mapping data format and analysis platform for electrophysiology research. *Frontiers in Physiology*, 12:646023.

Wolleb, J., Bieder, F., Sandkühler, R., and Cattin, P. C. (2022). Diffusion models for medical anomaly detection. In Wang, L., Dou, Q., Fletcher, P. T., Speidel, S., and Li, S., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Lecture Notes in Computer Science, pages 35–45, Cham. Springer Nature Switzerland.

Wong, G. R., Nalliah, C. J., Lee, G., Voskoboinik, A., Prabhu, S., Parameswaran, R., Sugumar, H., Anderson, R. D., McLellan, A., Ling, L.-H., et al. (2019). Dynamic atrial substrate during high-density mapping of paroxysmal and persistent af: implications for substrate ablation. *JACC: Clinical Electrophysiology*, 5(11):1265–1277.

Yeh, I.-C. (1998). Concrete Compressive Strength. UCI Machine Learning Repository.

Zahid, S., Cochet, H., Boyle, P. M., Schwarz, E. L., Whyte, K. N., Vigmond, E. J., Dubois, R., Hocini, M., Haïssaguerre, M., Jaïs, P., and Trayanova, N. A. (2016). Patient-derived models link re-entrant driver localization in atrial fibrillation to fibrosis spatial pattern. *Cardiovascular Research*, 110(3):443–454.

Zhao, Y., Nasrullah, Z., and Li, Z. (2019). Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7.

Zhou, C. and Paffenroth, R. C. (2017). Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 665–674.

Zhou, X., Liu, H., Pourpanah, F., Zeng, T., and Wang, X. (2022). A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. *Neurocomputing*, 489:449–465.

Zierk, J., Metzler, M., and Rauh, M. (2021). Data mining of pediatric reference intervals. *Journal of Laboratory Medicine*, 45(6):311–317.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. (2018). Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.