THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Adaptive Representation Learning for Prediction and Semantic Segmentation

David Hagerman

Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden, 2026

Adaptive Representation Learning for Prediction and Semantic Segmentation

David Hagerman
ISBN 978-91-8103-448-6

Acknowledgements, dedications, and similar personal statements in this thesis, reflect the author's own views.

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5905
ISSN 0346-718X

Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
Phone: +46 (0)31 772 1000

Cover:
A statue at the summit of Mount Fuji. The statue boundaries contain fine geometric detail and require high-resolution representation, whereas the surrounding sky and clouds are semantically homogeneous and can be represented more coarsely. Photograph by the author.

Printed by Chalmers Digital Printing
Gothenburg, Sweden, August 2026

Adaptive Representation Learning for Prediction and Semantic Segmentation

David Hagerman
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Deep learning models for prediction and segmentation are often designed under idealised assumptions: that inputs are complete, relevant measurements are available, and visual data can be processed at a fixed spatial resolution. Clinical and biomedical applications often violate these assumptions. Observations may be incomplete, heterogeneous, or costly to acquire, and the information needed for prediction or segmentation may be concentrated in a small part of the input.

This thesis develops deep learning methods that use information and computation selectively. For prediction under observation constraints, GenoARM (Paper A) formulates antimicrobial resistance prediction as a joint problem of gene-test selection and resistance classification, using reinforcement learning to construct compact genomic measurement panels. A multi-stream LVEF prediction method (Paper E) estimates cardiac function from echocardiographic examinations where views may be missing, duplicated, or misclassified, combining image sequences and optical flow to handle variable input configurations.

For segmentation, ARTA (Paper B) introduces adaptive mixed-resolution token allocation, concentrating high-resolution tokens in semantically complex regions while keeping simpler regions coarse. BATS (Paper C) extends this principle to volumetric medical segmentation using dense multi-scale allocation and Parent Attention for cross-scale context. SwInception (Paper D) complements these adaptive methods by strengthening local vision Transformers with a multi-scale convolutional structure. Together, the papers show how adaptive representation learning can improve robustness and efficiency under incomplete observations, costly measurements, and spatially uneven visual information.

Keywords: Deep learning, representation learning, medical data, computer vision, semantic segmentation, adaptive computation.

To Adam.

List of Publications

This thesis is based on the following publications:

[A] **David Hagerman**, Anna Johnning, Roman Naeem, Fredrik Kahl, Erik Kristiansson, Lennart Svensson, "Optimizing gene-based testing for antibiotic resistance prediction". AAAI 2025.

[B] **David Hagerman**[*], Roman Naeem[*], Erik Brorsson, Fredrik Kahl, Lennart Svensson, "ARTA: Adaptive Mixed-Resolution Token Allocation for Efficient Dense Feature Extraction". arXiv Preprint 2026.

[C] **David Hagerman**, Roman Naeem, Fredrik Kahl, "BATS: Resource-Efficient Volumetric Segmentation with Boundary-Aware Mixed-Resolution Tokens". arXiv Preprint 2026.

[D] **David Hagerman**, Roman Naeem, Jakob Lindqvist, Carl Lindström, Fredrik Kahl, Lennart Svensson, "SwInception - Local Attention Meets Convolutions". ICPRAI 2024.

[E] Jennifer Alvén, Eva Hagberg, **David Hagerman**, Richard Petersen, Ola Hjelmgren, "A deep multi-stream model for robust prediction of left ventricular ejection fraction in 2D echocardiography". Scientific Reports 2024.

Other publications by the author, not included in this thesis, are:

[F] Eva Hagberg, **David Hagerman**, Richard Johansson, Nasser Hosseini, Jan Liu, Elin Björnsson, Jennifer Alvén, Ola Hjelmgren, "Semi-supervised learning with natural language processing for right ventricle classification in echocardiography—a scalable approach". *Computers in Biology and Medicine*, Vol 143, Page 105282, 2022.

[G] Georg Hess, Johan Jaxing, Elias Svensson, **David Hagerman**, Christoffer Petersson, Lennart Svensson, "Masked autoencoder for self-supervised pre-training on lidar point clouds". *WACV*, Pages 350-359, 2023.

[H] Jennifer Alvén, Richard Petersen, **David Hagerman**, Mårten Sandstedt, Pieter Kitslaar, Göran Bergström, Erika Fagman, Ola Hjelmgren, "PlaqueViT:

a vision transformer model for fully automatic vessel and plaque segmentation in coronary computed tomography angiography". *European Radiology*, Vol 35, Issue 8, Pages 4461-4471, 2025.

[I] Roman Naeem, **David Hagerman**, Lennart Svensson, Fredrik Kahl, "Trexplorer: recurrent detr for topologically correct tree centerline tracking". *MICCAI*, Pages 744-754, 2024.

[J] Roman Naeem, **David Hagerman**, Jennifer Alvén, Lennart Svensson, Fredrik Kahl, "Trexplorer super: Topologically correct centerline tree tracking of tubular objects in ct volumes". *MICCAI*, Pages 595-605, 2025.

[K] Roman Naeem, **David Hagerman**, Jennifer Alvén, Fredrik Kahl, "RefTr: Recurrent Refinement of Confluent Trajectories for 3D Vascular Tree Centerline Graphs". *MICCAI*, 2026.

Contents

Acknowledgments

When I was 21, I dropped out of university and thought to myself, "Higher education is probably not for me." When I later started thinking about university again, my plan was, at most, to complete a bachelor's degree. Now, roughly ten years later, as I write my PhD thesis, I realize that what you want and where you are meant to go are things you discover along the way.

There are many people who made this journey possible.

First, I want to thank my original supervisor, the late Lennart Svensson, who sadly did not live to see this thesis finished. He was brilliant, curious, and encouraging, and always kind and caring toward the people around him. When he fell ill and had to step back from supervision, I lost not just a supervisor, but a mentor I deeply admired. I am grateful for the time I had with him.

I am deeply grateful to Fredrik Kahl, who started as my co-supervisor and later took over as my main supervisor, guiding and supporting me throughout this work. I also want to thank Roman Naeem for the close collaboration over all these years. I hope we can continue that even after this. I am grateful to Ida Häggström, who stepped in as co-supervisor during a difficult period, and to Jennifer Alven, who gave me the first push toward pursuing a PhD in the first place.

I would also like to thank the entire Signal Processing group for making our corridor a pleasant and fun place to work.

I am grateful to Erik Kristiansson and Anna Johnning from the Fraunhofer-Chalmers Centre for a fun and inspiring collaboration, and for sharing their expertise on antibiotic resistance, which made this work possible. I also want to thank Eva Hagberg and Ola Hjelmgren from Cardiobotics at Sahlgrenska for their clinical expertise, annotations, and the many discussions that shaped the development of the models for predicting cardiovascular health markers.

Finally, I want to thank my family. My wife Anastasia, for her continuous support and love throughout; without her, this would not have been possible. My son Adam, who helped by making me resilient to sleepless nights. And my mother Lisbeth and my sister Sara, for their encouragement and belief in me along the way.

This project was funded by MedTech West in a joint collaboration between Chalmers University of Technology and Sahlgrenska University Hospital.

Acronyms

AI:	Artificial Intelligence
AFA:	Active Feature Acquisition
AP:	Average Precision
AR:	Average Recall
ARTA:	Adaptive Mixed-Resolution Token Allocation
BATS:	Boundary-Aware Token Selection
CNN:	Convolutional Neural Network
CT:	Computed Tomography
FFN:	Feed-Forward Network
Dice:	Dice-Sørensen coefficient
FLOPs:	Floating Point Operations per second
MAE:	Mean Absolute Error
MDP:	Markov Decision Process
mIoU:	Mean Intersection over Union
MLP:	Multi-Layer Perceptron
MRI:	Magnetic Resonance Imaging
PCR:	Polymerase ChainReaction
ReLU:	Rectified Linear Unit
RL:	Reinforcement Learning
SOTA:	State of the Art
ViT:	Vision Transformer

Part I

Overview

CHAPTER 1

Introduction

1.1 Introduction

Many practical deep learning problems do not provide complete, reliable, or uniformly informative inputs. Echocardiographic examinations may contain missing, duplicated, or incorrectly labelled video sequences. Relevant signals, such as cardiac motion, may remain implicit in the raw image sequence. Diagnostic settings may require selecting a small set of measurements rather than acquiring all available information. In semantic segmentation, fine spatial detail may be essential near boundaries and small structures but unnecessary across large homogeneous regions.

This thesis studies how deep learning models can adapt to such settings by using information and computation selectively. The central question is which observations should be acquired, which structures should be made explicit in the representation, and which spatial regions should receive fine-resolution processing.

The thesis develops this question in two connected directions. The first concerns the observations available to a model. GenoARM (Paper A) studies how to select a compact set of genomic measurements for antimicrobial

resistance prediction when complete measurement is impractical. The LVEF paper (Paper E) studies prediction from variable sets of echocardiographic video sequences and uses motion-derived input to make temporal information explicit.

The second direction concerns visual representation and computation. SwInception (Paper D) studies how local and multi-scale architectural priors can strengthen transformer-based volumetric segmentation. ARTA (Paper B) and BATS (Paper C) allocate spatial resolution according to the predicted difficulty of different image or volume regions.

Echocardiographic assessment of cardiac function illustrates the first observation problem. An examination typically contains several video sequences acquired from standard imaging planes, commonly referred to as views. Echocardiography is widely used because it is non-invasive, cost-effective, and clinically informative [1]. Estimation of left ventricular ejection fraction is a central part of assessing systolic function [2], but it is affected by image quality, operator dependence, and interobserver variability [3]. Deep learning methods can support more consistent assessment [4], but routine examinations do not always contain a complete and correctly identified set of standard views. Individual videos may be missing, duplicated, or assigned to the wrong view, and acquisition conditions vary between patients, operators, and ultrasound systems.

A model designed for a complete and well-labelled set of views may perform well on curated data but be less reliable in routine clinical use. Robust prediction therefore requires aggregating information from a variable set of echocardiographic videos and tolerating uncertainty in their availability and labels.

LVEF prediction also presents a representational challenge. Ejection fraction describes the change in ventricular volume between diastole and systole, so appearance-based features alone leave part of the relevant temporal signal implicit. Motion-derived representations such as optical flow make displacement between frames explicit and provide complementary information about ventricular contraction. Paper E combines multiple echocardiographic views with optical flow to address both variable view availability and the temporal nature of LVEF estimation.

Antimicrobial resistance prediction raises a related but distinct observation problem. Here, the relevant measurements may exist in principle, but acquiring all of them can be too slow or expensive for rapid clinical decision-making.

Antimicrobial resistance is a major threat to global health and creates an urgent need for diagnostic tools that can support effective antibiotic prescription [5], [6], [7]. Deep learning models trained on complete pathogen genomic information have shown strong potential for predicting resistance [8], [9], [10], [11]. However, complete genome sequencing is not always compatible with the time and cost constraints of routine diagnostic use. Faster alternatives, such as PCR-based tests, measure only a targeted subset of known resistance genes or mutations.

This changes the task from prediction alone to prediction under measurement constraints. Thousands of genes and mutations may be associated with resistance [12], but measuring all of them defeats the purpose of a fast and cost-effective test. Selecting only the most frequent genes is also insufficient, since prevalence does not necessarily coincide with diagnostic value, and co-located resistance genes may provide overlapping information. A useful model should therefore support accurate prediction while identifying compact and informative subsets of measurements. GenoARM (Paper A) addresses this problem by learning to construct gene panels for resistance prediction rather than assuming that all genomic measurements are available.

Semantic segmentation raises the same selectivity problem in spatial form. A model must predict a label at every pixel or voxel, but not every location requires the same amount of computation. Large homogeneous regions may be represented coarsely, whereas boundaries, lesions, small structures, and ambiguous regions require finer spatial detail. This trade-off becomes especially demanding in high-resolution images and volumetric medical data, where memory and computation increase rapidly with spatial resolution.

Standard segmentation architectures therefore face a resolution trade-off. Processing all regions at high resolution preserves detail, but spends substantial capacity on regions that may not need it. Processing at lower resolution improves efficiency, but risks losing the small structures and boundaries that often determine segmentation quality. Multi-scale encoder-decoder architectures [13] partly address this tension by combining coarse semantic context with finer spatial information, and hybrid convolutional and transformer-based designs further combine local spatial bias with more flexible contextual interaction. SwInception (Paper D) explores this direction by strengthening volumetric transformer-based segmentation with local and multi-scale convolutional structure. However, using multiple scales does not by itself make the allocation of resolution adaptive. In most architectures, the resolution hierarchy is fixed

in advance, so all images and regions are processed according to the same predetermined pattern, regardless of where the difficult or semantically important structures actually occur. ARTA (Paper B) addresses this limitation in two-dimensional semantic segmentation by making token resolution depend on predicted semantic complexity.

Volumetric medical segmentation is a particularly demanding case of this resolution trade-off. Three-dimensional scans are commonly processed as overlapping crops because dense processing of an entire high-resolution volume may exceed available memory. Within each crop, boundaries, thin structures, lesions, and small targets may require fine spatial detail, while large homogeneous regions can be represented more coarsely. This makes adaptive token allocation particularly attractive in volumetric data, where selectively concentrating fine-resolution processing can provide substantial memory and computational benefits. BATS (Paper C) extends adaptive mixed-resolution allocation to this setting.

These problems differ in form, but each requires the model to work with information that is unevenly available or useful. The relevant view may be missing, the useful gene may not have been measured, the clinically meaningful motion may be implicit, or the important segmentation region may occupy only a small part of the image or volume. Selective use of information and computation is therefore the unifying concern of this thesis: adapting deep learning models to settings where observations, structures, and spatial detail differ in availability, reliability, cost, and importance.

1.2 Research Objectives

The overall aim of this thesis is to develop deep learning methods that use information and computation more selectively. This aim is studied in settings where observations may be incomplete or heterogeneous, where informative measurements may be costly to acquire, and where visual prediction requires efficient use of structure, spatial resolution, and computational resources.

This aim is addressed through the following research objectives:

1. **Prediction from partial and heterogeneous observations.** Develop models that remain reliable when the available observations are incomplete or imperfectly specified, as in echocardiographic examinations containing missing, duplicated, or incorrectly labelled views.

2. **Structure-aware visual representation learning.** Investigate how incorporating structural priors into visual representations, including motion, locality, and scale, can improve learning in image-based prediction and segmentation.

3. **Adaptive selection and refinement of measurements and spatial regions.** Develop methods that identify which observations, measurements, or regions are most informative to acquire or refine, including adaptive selection of genomic measurements and adaptive mixed-resolution refinement of image regions.

4. **Efficient semantic segmentation in two-dimensional and volumetric data.** Design and evaluate segmentation models that improve the trade-off between accuracy and computational cost, with particular emphasis on high-resolution images and volumetric medical data where segmentation difficulty is spatially concentrated.

1.3 Contributions

The thesis includes five papers that address these objectives from complementary perspectives.

GenoARM (Paper A) addresses prediction when complete measurement is impractical. Rather than assuming that full genomic information is available, it formulates gene selection as a sequential decision-making problem. The method asks a practical diagnostic question: if only a small panel of genes can be measured, which genes should be included so that resistance can still be predicted accurately? The contribution is to couple resistance prediction with the design of a compact and informative measurement set.

ARTA (Paper B) develops adaptive mixed-resolution token allocation for semantic segmentation. ARTA concentrates high-resolution tokens in semantically complex regions while keeping simpler regions coarse. The efficiency question becomes not how to process a dense token set more cheaply, but where fine tokens should be introduced.

BATS (Paper C) extends this adaptive mixed-resolution principle to volumetric medical segmentation. It addresses the stronger memory and computational constraints of three-dimensional data, where segmentation difficulty is typically concentrated around small structures, lesions, and boundaries. The paper

adapts the allocation principle to three-dimensional input crops using a dense boundary predictor to identify where finer resolution is needed, parallel scoring of all refinement levels to prevent erroneous coarse decisions from suppressing fine-scale evidence, and parent cluster attention to provide hierarchical cross-scale context.

SwInception (Paper D) combines transformer-based representation learning with convolutional branches and multi-scale feature extraction for volumetric medical image segmentation. The architecture combines broader contextual modelling with local and hierarchical spatial priors, which are important for representing anatomical structures at different scales. It contributes to the thesis by showing how explicit architectural structure can support structure-aware representation learning in volumetric data.

The multi-stream LVEF prediction paper (Paper E) proposes a deep model for predicting left ventricular ejection fraction from echocardiographic examinations containing up to four automatically identified standard views. Each view is represented by both its image sequence and corresponding optical flow, allowing the model to combine appearance and motion information. The model is designed to tolerate missing, misclassified, and duplicated views through pre-training, sampling strategies, and parameter sharing. The paper therefore treats variable view availability and motion representation as connected problems: the model must predict from the echocardiographic videos available in a particular examination while capturing the temporal change that defines LVEF.

Together, Papers A and E address constraints on the observations available to a model, while Papers B–D address how spatial structure, resolution, and computation should be represented and allocated in visual models.

1.4 Thesis Outline

The remainder of the thesis is organized as follows.

Chapter 2 introduces the methodological background for the thesis. It describes the tasks and data modalities considered in the included papers: echocardiographic image sequences, genomic measurements, natural images, and volumetric medical images, and reviews classical and deep learning approaches, including representation learning, reinforcement learning, convolutional networks, transformers, and semantic segmentation.

Chapter 3 focuses on learning from partial, irregular, and costly observations. It examines two settings where the standard fixed-input assumption breaks down: prediction from variable multi-view clinical echocardiographic examinations, and antimicrobial resistance prediction from selectively acquired genomic measurements. The chapter positions these problems relative to missing-data learning, multi-view learning, feature selection, and active feature acquisition.

Chapter 4 discusses efficient, structure-aware, and adaptive visual representation learning. It considers why semantic segmentation requires both local spatial detail and broader context, and why uniform dense processing is often inefficient. The chapter introduces multi-scale representations, visual inductive biases, hybrid CNN–transformer architectures, efficient attention mechanisms, and adaptive mixed-resolution processing, and discusses the additional challenges that arise in volumetric medical segmentation.

Chapter 5 summarizes the included papers and relates them to the research objectives. Each paper is described in terms of its motivation, main methodological contribution, experimental setting, and role in the overall thesis.

Chapter 6 concludes the thesis by summarizing the main findings, discussing limitations, and outlining directions for future work.

1.5 Chapter Summary

The papers in this thesis address settings where observations may be incomplete or costly, relevant structure may remain implicit, and spatial difficulty may be unevenly distributed. The following chapter introduces the methodological background needed to understand the proposed approaches to observation selection, structure-aware representation learning, and adaptive spatial computation.

CHAPTER 2

Background

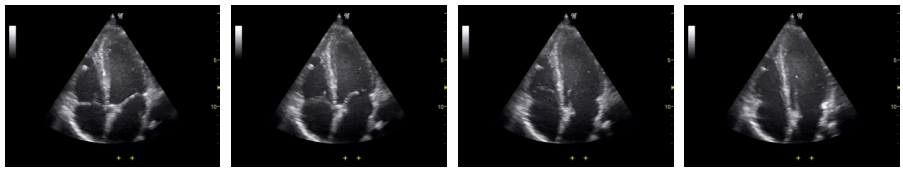
The included papers apply deep learning to genomic data and associated metadata, natural images, volumetric medical images, and echocardiographic image sequences. These settings differ in both input modality and prediction target, ranging from scalar and multi-label prediction to two- and three-dimensional semantic segmentation.

This chapter introduces the tasks, data modalities, and methodological concepts needed to understand the included work. It begins with the four application settings, then reviews the transition from handcrafted to learned representations, reinforcement learning for sequential selection, convolutional networks, transformers, and semantic segmentation.

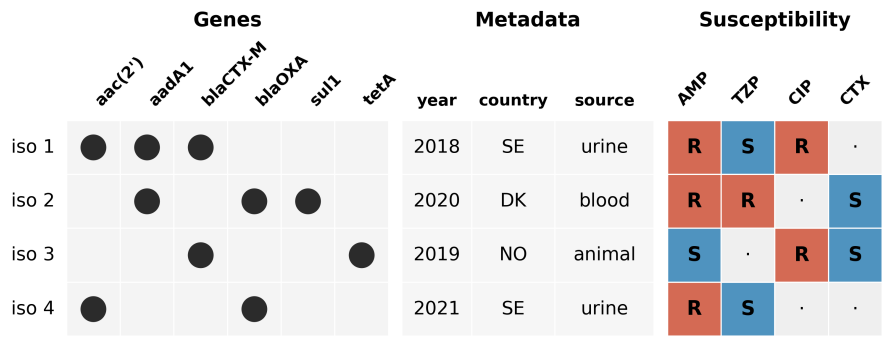
2.1 Tasks and Data Modalities

The included papers differ in both input modality and prediction target. Figures 2.1 and 2.2 illustrate the four main task settings considered in the thesis.

Echocardiographic assessment of left ventricular ejection fraction is a scalar prediction task based on spatiotemporal medical image data [1], [2], [3]. The



(a)

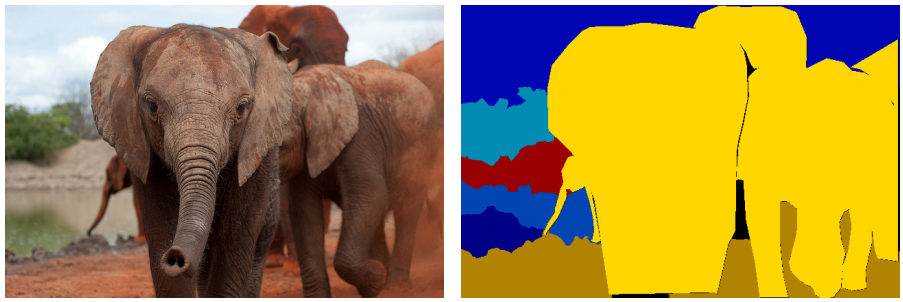


(b)

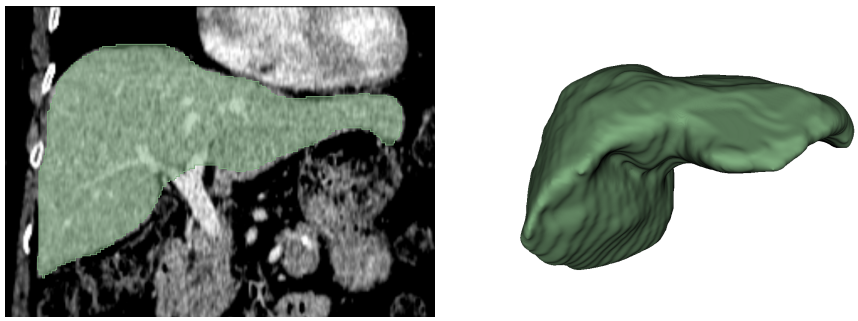
Figure 2.1: Examples of prediction tasks and data modalities considered in the thesis. Echocardiographic prediction (a) uses a temporal medical image sequence as input, while AMR prediction (b) uses sparse genomic markers and metadata to predict resistance phenotypes across antibiotics. Image credits and license information are provided in Image Credits and Licenses.

input consists of transthoracic two-dimensional echocardiographic image sequences, each showing the heart from a particular imaging plane, or view, over the cardiac cycle. Although the output is a single continuous value, the relevant signal is not contained in any individual frame. Ejection fraction describes the change in ventricular volume between diastole and systole, so the model must represent both cardiac appearance and motion.

Echocardiography also has acquisition-specific variability. Ultrasound images are affected by speckle noise, acoustic shadowing, probe positioning, patient anatomy, and the available acoustic window [4], [14]. The same anatomical view may therefore appear differently across patients, operators, ultrasound systems, and acquisition protocols. The temporal dimension adds further variation



(a)



(b)

Figure 2.2: Examples of semantic segmentation tasks and data modalities considered in the thesis. Natural-image segmentation (a) assigns labels to pixels in two-dimensional images, while volumetric medical segmentation (b) assigns labels to voxels in three-dimensional CT or MR data. Image credits and license information are provided in Image Credits and Licenses.

in frame rate, cardiac phase coverage, and number of frames. In routine clinical examinations, the available views may also be missing, duplicated, or imperfectly labelled. These properties make LVEF prediction a problem of spatiotemporal representation learning under variable and incomplete clinical input. Paper E addresses this setting using multiple echocardiographic views and motion-derived representations.

Antimicrobial resistance prediction estimates whether a bacterial isolate is resistant or susceptible to one or more antibiotics from genomic measure-

ments [15]. These measurements may be supplemented with isolate metadata [16]. The genomic input considered in this thesis consists of genes, mutations, and other markers associated with antimicrobial resistance. The metadata describe the country of origin, the source from which the isolate was obtained, and the year of collection. These variables provide epidemiological context, as resistance prevalence and pathogen populations may vary geographically, across isolation sources, and over time [16], [17], [18].

The relevant biological signal may arise from acquired resistance genes, resistance-conferring mutations, or interactions between multiple genetic variants [12], [19]. The effect of an individual variant may also depend on the surrounding genetic background [19]. In practice, the genomic feature representation depends on how resistance determinants are defined and annotated, and therefore on the database and annotation pipeline used [12]. Consequently, a frequently occurring marker is not necessarily the most discriminative for separating resistant from susceptible isolates, and informative predictions may depend on combinations of genomic markers and epidemiological variables.

GenoARM (Paper A) couples resistance prediction with genomic measurement selection. Complete genomic profiles may be informative, but obtaining them can be too slow or costly for a targeted diagnostic test. Such tests require a compact set of genes or mutations to be chosen in advance, while contextual metadata such as country, source, and year may remain available to the predictor.

The task therefore combines two problems: predicting resistance and selecting which genomic markers should be measured. This selection is non-trivial because resistance markers frequently co-occur and may provide redundant information. The value of adding a marker consequently depends on the markers already included in the panel. GenoARM addresses this dependency by constructing compact panels through sequential decision-making.

Semantic segmentation assigns a class label to each spatial location in an image or volume. Unlike scalar prediction, segmentation produces a spatially organized output. Segmentation quality depends on correct semantic recognition and spatial precision, including boundary accuracy, preservation of small structures, and consistency across neighboring regions.

Natural-image benchmarks for semantic segmentation include ADE20K, COCO-Stuff, and Cityscapes [20], [21], [22]. These datasets contain substantial variation in object size, texture, scene composition, and boundary complexity.

They therefore provide a demanding two-dimensional test bed for evaluating whether adaptive mixed-resolution representations can preserve fine structures and object boundaries while representing simpler regions more coarsely. ARTA (Paper B) develops and evaluates this principle in natural images before it is extended to volumetric medical data in BATS (Paper C).

Volumetric medical images introduce a different set of constraints. The medical segmentation papers in this thesis use computed tomography and magnetic resonance images. CT intensities are measured in Hounsfield units and are commonly interpreted through task-specific windowing. MR images do not have a standardized intensity scale across scanners and protocols, and are often acquired in different sequences or contrasts that may be treated as separate input channels depending on the task [23]. MR data therefore commonly require intensity normalization before being used as model input [24].

Both CT and MR volumes are memory-intensive, especially when high spatial resolution is needed. Target structures may occupy only a small fraction of the scan, while surrounding anatomy remains necessary for localization and context. Medical segmentation also involves class imbalance, small structures, fuzzy boundaries, and variation across scanners, protocols, and patient populations [25]. The diagnostically important information is often spatially localized, concentrated near boundaries, lesions, or small anatomical targets. This makes uniform high-resolution processing inefficient: much of the volume may need to be inspected, but not all of it needs the same level of detail. This setting is central to both BATS (Paper C), which adapts mixed-resolution allocation to 3D segmentation, and SwInception (Paper D), which studies fixed local and multi-scale architectural priors for volumetric segmentation.

A further shared challenge is the cost and variability of annotation. Medical annotations require trained experts and can exhibit substantial interobserver variability depending on image quality and target definition [26], [27]. Annotated medical imaging datasets are often small by deep learning standards; for many specialized tasks, datasets with tens to hundreds of annotated cases are common. This makes architectural priors and efficient use of supervision particularly important.

2.2 From Handcrafted to Learned Representations

Before deep learning became dominant, prediction and image analysis commonly relied on handcrafted features and task-specific pipelines. In image analysis, these included edge detection [28], deformable models [29], atlas registration [30], vesselness filtering [31], and graph-based optimization [32]. Biomedical prediction from non-image data similarly relied on predefined feature sets, statistical models, and explicit feature-selection procedures [33], [34].

These approaches encode domain knowledge directly. A vesselness filter, for example, assumes that the target has a tubular appearance [31], while an atlas-based method assumes correspondence to a reference anatomy [30]. Such methods can be interpretable and computationally efficient when their assumptions match the data. Their performance may, however, degrade when acquisition conditions, anatomy, or biological context differ from those anticipated during design.

Deep learning methods learn intermediate representations jointly with the prediction or segmentation objective, rather than defining all relevant features in advance [35]. Convolutional neural networks learn hierarchical visual features directly from images [36], [37], while transformer-based models provide flexible mechanisms for modeling interactions between input elements [38], [39]. In genomic and other high-dimensional biomedical data, deep learning provides a way to model complex relationships among many potentially interacting features [8], [40].

In supervised deep learning, a model f_θ with parameters θ is trained from examples $\{(x_i, y_i)\}_{i=1}^{N}$ by minimizing an empirical loss,

$$\theta^* = \arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i), y_i)\,. \tag{2.1}$$

The choice of loss $\mathcal{L}$ depends on the task: regression losses are used for scalar prediction, classification losses for resistance or class prediction, and dense segmentation losses for pixel- or voxel-level outputs. This formulation also makes clear where the fixed-input assumption enters: the model f_θ is usually designed for inputs with a predefined structure.

This shift does not remove the need for structure or prior assumptions. It changes where such structure enters the model. Structure may be introduced through the input representation, as with motion-derived features in image

sequences; through the architecture, as with locality and weight sharing in convolutional networks; through initialization or biological prior information; or through adaptive mechanisms that decide which measurements or spatial regions should receive more capacity. These choices matter especially when annotated data are limited, inputs are heterogeneous, or uniform processing is computationally expensive.

The following sections describe the main learning mechanisms and architectural structures used in the included papers.

2.3 Representation Learning

Representation learning transforms raw input data into features that are useful for a downstream task [41]. In deep learning, these representations are learned jointly with the prediction objective. Depending on the task, a representation may be a vector summarizing an image sequence, a set of features derived from genomic measurements, a spatial feature map, or a sequence of tokens.

A common way to express this is to decompose a predictor into a representation map and a task head,

$$z = h_\theta(x), \qquad \hat{y} = g_\phi(z), \tag{2.2}$$

where h_θ transforms the input into a learned representation z, and g_ϕ maps that representation to the prediction. In segmentation, z is usually spatially organized rather than a single vector, since the output must remain aligned with the image or volume.

What makes a representation useful depends on the task. In some cases, it should be invariant to irrelevant variation, such as changes in illumination, scanner settings, or acquisition conditions. In others, it must preserve variation that is essential. Motion over time is central to estimating cardiac function; spatial detail at boundaries is central to segmentation quality. A useful representation must therefore suppress some sources of variation while preserving others, depending on the prediction target.

Learned representations are shaped by design choices: preprocessing, input channels, architecture, initialization, tokenization, and adaptive selection mechanisms. These choices determine not only what information is extracted, but also at what spatial, temporal, or semantic granularity it is represented. These choices determine what information remains available to later layers and

at what spatial, temporal, or semantic granularity it is represented. A model cannot later select or refine information that has already been discarded by its representation.

Some tasks require global representations that summarize the input as a whole; others require spatially organized representations that retain local detail. Segmentation cannot rely on a single global vector, since the output must be aligned with the spatial structure of the image. Modern segmentation models therefore use hierarchical or multi-scale representations, where coarse features capture broader context and finer features preserve local detail. Token-based models introduce another form of granularity, where images or feature maps are represented as sets of local elements. Mixed-resolution representations extend this further by allowing different parts of the input to be represented at different levels of detail. This idea is central to ARTA (Paper B) and BATS (Paper C).

2.4 Reinforcement Learning

GenoARM (Paper A) uses reinforcement learning because it does not only predict resistance from a fixed input; it constructs a gene panel through a sequence of choices. The value of each choice depends on which genes have already been selected. A gene that is informative on its own may be redundant once a related marker has been included, while a less frequent gene may be valuable if it contributes complementary information. This context-dependence makes the problem different from ranking genes independently.

In reinforcement learning, an agent learns to choose actions based on feedback from an environment [42]. Instead of receiving a direct target for every action, the agent receives rewards indicating how useful its actions were. At time step t, the agent observes a state s_t, chooses an action a_t, and receives a reward r_t. The policy $\pi(a_t \mid s_t)$ defines the probability of choosing an action given the current state. The objective is to maximize expected cumulative reward over an episode.

For a policy with parameters θ, this objective can be written as

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r_t\right], \tag{2.3}$$

where $\gamma \in [0, 1]$ is a discount factor and T is the episode length. In the gene-

panel setting of GenoARM (Paper A), an episode corresponds to constructing one panel, the actions correspond to selecting genes or mutations, and the reward is based on the predictive usefulness of the resulting panel.

This formulation is useful when the value of an action depends on context and cannot be evaluated in isolation. It also introduces the credit assignment problem: the model must learn which earlier actions contributed to a successful or unsuccessful outcome. Reinforcement learning further involves a trade-off between exploration and exploitation, since the agent must try actions whose value is uncertain while also exploiting actions already believed to produce high reward.

Many reinforcement learning methods learn a parameterized policy directly. Policy-gradient methods update policy parameters in a direction that increases expected reward [43], but simple updates can be unstable when rewards are noisy or delayed. Proximal Policy Optimization (PPO) addresses this by limiting how much the policy is allowed to change during each update [44]. Its clipped objective discourages excessively large steps relative to the previous policy, making training more stable while still allowing improvement.

In GenoARM, the policy selects one additional genomic marker at each step based on the current panel. The reward measures the predictive usefulness of the resulting panel across antibiotics, allowing the agent to account for interactions and redundancy between markers.

2.5 Convolutional Networks

Convolutional networks encode two assumptions that are particularly useful for visual data: nearby pixels or voxels tend to be related, and similar local patterns may occur at different spatial locations.

A convolution applies a small learnable filter to local neighborhoods and produces a feature map recording the filter's response at each spatial location [36], [37]. For a two-dimensional input feature map x, a convolutional layer can be written as

$$y_{c_{\text{out}}}(u, v) = b_{c_{\text{out}}} + \sum_{c_{\text{in}}} \sum_{\Delta u, \Delta v} W_{c_{\text{out}}, c_{\text{in}}, \Delta u, \Delta v}\, x_{c_{\text{in}}}(u + \Delta u, v + \Delta v), \qquad (2.4)$$

where W is a learned kernel, b is a bias term, and the same weights are reused at every spatial position. This weight sharing makes convolutional layers

parameter-efficient and gives them a form of translation equivariance: if an input pattern shifts spatially, the corresponding feature response shifts in the same way. In volumetric data, the same operation extends by adding a depth coordinate and a third spatial kernel dimension.

CNNs build hierarchical representations. Early layers capture local low-level patterns such as edges, textures, and intensity transitions; deeper layers combine these into more abstract features with progressively larger receptive fields. Downsampling through pooling or strided convolution increases spatial context while reducing feature map resolution. In volumetric data, maintaining high-resolution feature maps throughout the network is often infeasible due to memory cost, making this hierarchy especially important.

For segmentation, this hierarchy is commonly organized as an encoder–decoder. The encoder progressively reduces spatial resolution while increasing the receptive field, and the decoder restores spatial resolution to produce the output map. U-Net [13] became the dominant design in biomedical segmentation through skip connections that pass encoder feature maps to the decoder, combining fine spatial detail with higher-level semantic features. Residual connections further improved optimization in deeper networks [45]. Volumetric extensions such as 3D U-Net and V-Net applied the same principle to three-dimensional data [46], [47], and nnU-Net systematized these designs into a self-configuring framework [24].

The properties that make CNNs effective also define their limitations. Since convolutional layers operate locally, long-range interactions are modeled only indirectly through depth and downsampling. Dense convolutional processing can also be expensive when high resolution must be maintained, especially in volumetric data. These limitations motivate attention-based and hybrid architectures, but do not make convolutional structure obsolete. SwInception (Paper D), for example, combines convolutional locality and multi-scale processing with transformer-based contextual modelling.

2.6 Transformers

Transformer models provide a flexible mechanism for combining information across spatial locations, time points, and input streams. They represent the input as a sequence of tokens and use self-attention to determine how information should be exchanged between them [38]. This is useful when relevant

relationships extend beyond a small local neighborhood, but full self-attention becomes expensive as the number of tokens increases. The computational cost is therefore a central concern in high-resolution and volumetric vision models.

Tokens and Self-Attention

A transformer operates on a sequence of tokens, each represented by an embedding vector [38]. In language models, tokens correspond to words or subwords. In vision models, tokens may correspond to image patches, spatial feature vectors, voxel patches, or features extracted by a preceding convolutional network. They can also represent temporal elements, such as frame-level features in an echocardiographic sequence.

The central operation is self-attention. For each token, the model computes three vectors by learned linear projections: a query, a key, and a value. Given matrices of queries Q, keys K, and values V, scaled dot-product attention is defined as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{2.5}$$

where d_k is the key dimension. Attention weights are determined by similarities between queries and keys, and the output is a weighted combination of the value vectors. In practice, attention is computed in multiple heads, allowing the model to learn several interaction patterns in parallel. A standard transformer block combines multi-head self-attention with a feed-forward network, normalization layers, and residual connections.

In multi-head attention, the same operation is applied in several learned subspaces:

$$\begin{aligned} \text{MHA}(X) &= \text{Concat}(H_1, \ldots, H_h)W^O, \\ H_j &= \text{Attention}(XW_j^Q, XW_j^K, XW_j^V), \end{aligned} \tag{2.6}$$

where h is the number of heads. Multiple heads allow the model to represent different interaction patterns in parallel, such as local similarity, long-range context, or cross-scale relationships.

The key difference from convolution is that self-attention defines interactions through learned content similarity rather than a fixed local filter. This makes it attractive when relevant interactions are not strictly local. The cost is that full self-attention scales quadratically with the number of tokens, which becomes expensive for high-resolution images and particularly for volumetric

data. Vision transformers therefore usually modify the plain attention pattern through hierarchy, locality, sparsity, or windowing.

Vision Transformers

Vision transformers adapt the transformer to image data by converting images or feature maps into tokens [39]. In the original formulation, an image is divided into non-overlapping patches, each flattened and projected into an embedding vector with positional information added. The resulting sequence is processed by transformer blocks. Output tokens can be used for image-level prediction or reshaped into a spatial feature map for dense prediction tasks.

This plain formulation is a useful conceptual baseline, but the papers in this thesis build on more structured transformer variants. For segmentation and volumetric analysis, the number of tokens is large and the spatial organization of the input is important. Standard self-attention can model interactions between distant tokens directly, but it does not by itself encode locality or translation equivariance. Vision transformers therefore commonly introduce hierarchy, local or windowed attention, positional information, or convolutional components. These modifications are not only computational shortcuts; they also shape the inductive bias of the model.

Figure 2.3 summarizes this process, from patch extraction and embedding to attention-based processing and spatial decoding.

Efficient Attention for Vision

Most efficient vision transformers reduce cost by restricting the attention pattern, so each token interacts only with a subset of other tokens.

A common strategy is window-based attention. The Swin Transformer [48] partitions the feature map into non-overlapping windows and applies self-attention independently inside each, reducing attention cost for a fixed window size. To avoid isolating windows completely, Swin alternates between regular and shifted window partitions, allowing information to pass across window boundaries in successive layers. It also builds hierarchical feature maps by progressively merging patches, making it suitable for dense prediction tasks.

Windowed attention assumes a regular grid of tokens. AutoFocusFormer [49] introduces cluster attention for token sets that are sparse or non-uniform. Tokens are grouped into local clusters, and attention is computed over neigh-

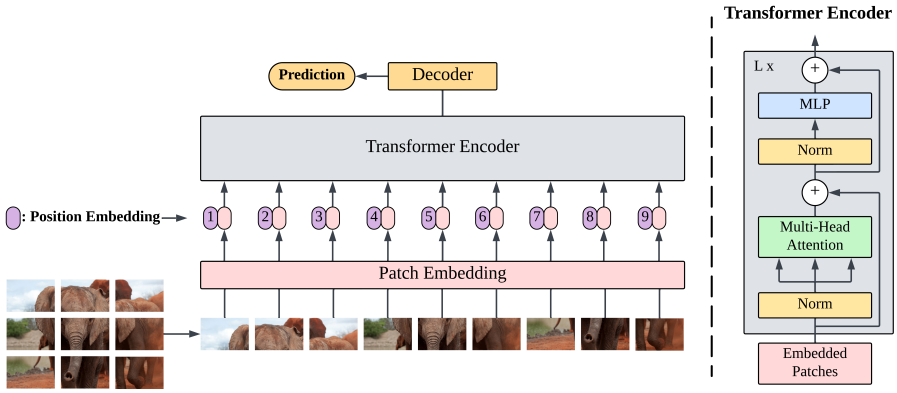


Figure 2.3: Conceptual overview of a vision transformer for image data. The input image or feature map is divided into patches, each patch is projected to a patch embedding, positional information is added, and the resulting token sequence is processed by attention blocks. For dense prediction tasks such as semantic segmentation, the output tokens must be converted back into a spatial feature map or decoded into segmentation masks.

borhoods formed from nearby clusters. A balanced clustering procedure orders token positions using a space-filling curve and partitions the ordered tokens into equally sized clusters, avoiding padding overhead from uneven cluster sizes. Mixed-resolution token sets need this kind of mechanism. In ARTA (Paper B) and BATS (Paper C), tokens correspond to regions of different sizes and do not form a regular dense grid, so a fixed window is no longer the natural unit of attention.

Figure 2.4 illustrates different ways of restricting the spatial interaction pattern of self-attention.

Hybrid CNN–Transformer Architectures

Many vision models combine convolutional and transformer-based components rather than using either in isolation. Convolutional layers provide efficient local feature extraction, weight sharing, and strong spatial inductive bias; transformer layers provide flexible interactions between tokens and can model longer-range dependencies more directly. This combination is useful in seg-

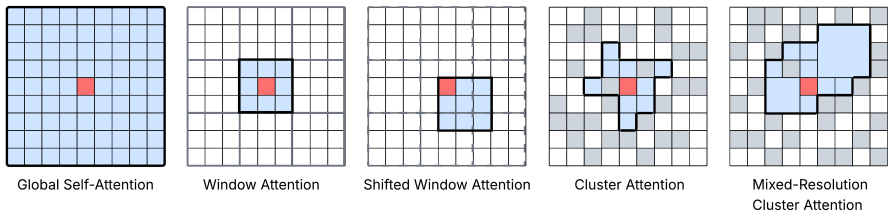


Figure 2.4: Conceptual comparison of spatial attention patterns in vision transformers. The red cell denotes the query token, while blue cells or regions denote tokens that can contribute to its output. Solid outlines mark the attended neighborhood, dashed lines indicate the previous window partition, and gray cells indicate spatial locations without an explicit token.

mentation because the task requires both local spatial precision and broader contextual reasoning.

Some models use a convolutional stem before transformer blocks, so that the tokens processed by attention are already derived from local image features. Others combine convolutional encoders with transformer bottlenecks, or use convolutional decoders to restore spatial detail from transformer-encoded features. In segmentation, convolutional layers are often used to preserve spatial detail while attention layers improve contextual reasoning.

For image sequences and echocardiography, convolutional layers may extract spatial or spatiotemporal features while transformer layers model relationships across time. This separates local visual feature extraction from temporal interaction modeling, which is useful when the target depends on motion across frames rather than single-frame appearance.

Hybrid architectures are especially common in medical image segmentation, where datasets are often small and purely attention-based models may require more data or regularization to learn useful spatial structure. UNETR and SwinUNETR are influential examples [50], [51]: both combine convolutional and attention-based components to support local precision and broader contextual reasoning. SwInception (Paper D) follows this motivation, but introduces additional local and multi-scale convolutional processing inside a Swin-based volumetric architecture.

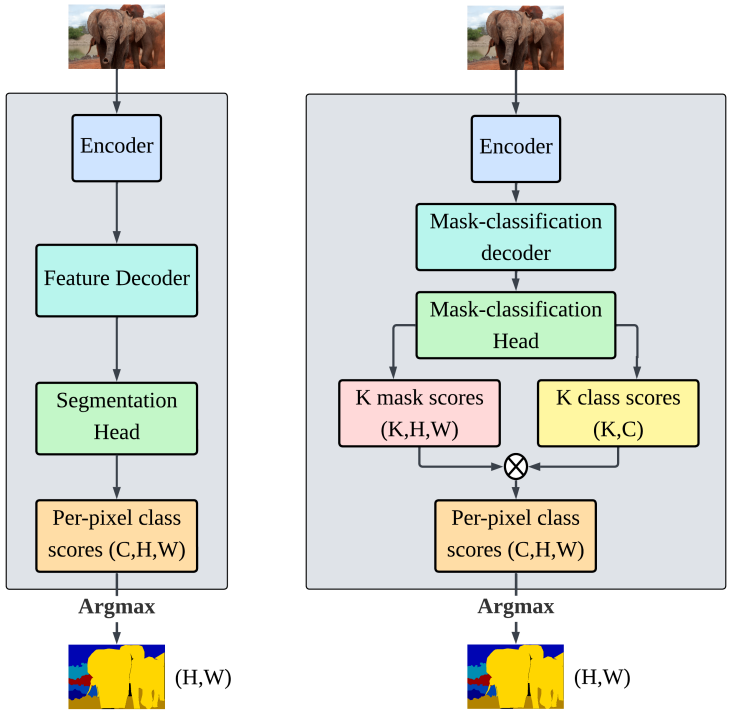
2.7 Semantic Segmentation and Structured Outputs

Semantic segmentation assigns a semantic class label to each location in the input: pixels in two-dimensional images and voxels in volumes. Unlike image-level prediction, it produces an output map spatially aligned with the input. The same class may appear in several disconnected regions, and the goal is to identify all locations belonging to each semantic category.

Although implemented as classification at each pixel or voxel, the output is structured. Neighboring predictions are not independent: boundaries should align with image evidence, small structures should be preserved, and predictions should remain spatially consistent. A model must recognize what is present while also retaining enough spatial detail to assign accurate labels near class boundaries, thin structures, and small targets.

Fully convolutional networks established end-to-end dense prediction from image features [52]. Encoder-decoder models such as U-Net [13] became central in biomedical segmentation: the encoder extracts increasingly abstract features while reducing resolution, and the decoder restores resolution to produce a segmentation map. Skip connections pass encoder feature maps to the decoder, combining fine spatial information with higher-level semantic features. This design reflects the core tension in segmentation: semantic context benefits from depth and downsampling, while spatial precision requires preserving fine detail.

Modern segmentation models also use mask-based formulations. In direct per-location classification, decoded features are mapped to class scores at each pixel or voxel. Mask-classification models instead predict a set of mask scores together with corresponding class scores, which are combined to obtain per-location class scores. During training, these continuous score outputs are used for the loss; discrete segmentation maps are produced only when hard predictions are needed, such as during inference, visualization, or evaluation. Query-based transformer decoders, as in DETR [53], MaskFormer, and Mask2Former [54], [55], implement this by having learned queries attend to image features and produce mask predictions. The decoder design is architecture-specific; for example, Mask2Former combines a pixel decoder with a masked-attention Transformer decoder. ARTA (Paper B) builds on this formulation for adaptive mixed-resolution segmentation.



(a) Per-location classification (b) Mask classification

Figure 2.5: Two output formulations for semantic segmentation. Per-location classification predicts spatial class scores directly, while mask classification combines mask and class scores. Bottom maps show hard predictions; dimensions are shown in 2D.

Figure 2.5 schematically contrasts direct per-location classification with mask classification.

Volumetric segmentation extends semantic segmentation to three-dimensional data. Architectures such as 3D U-Net, V-Net, nnU-Net, UNETR, SwinUNETR, and nnFormer have been developed for this setting [24], [46], [47], [50], [51], [56]. Full-resolution volumes are usually too large to process directly during training, so models are trained on cropped subvolumes, often with foreground oversampling to reduce class imbalance. Sliding-window inference combines

local predictions into a complete segmentation map at test time. BATS (Paper C) and SwInception (Paper D) both operate in this volumetric setting, but address different parts of the modelling problem: BATS adapts where fine-resolution tokens are used, while SwInception strengthens the fixed volumetric backbone with local and multi-scale convolutional structure.

Patch size determines the balance between spatial context and memory use. Small crops permit higher local resolution but contain less anatomical context, whereas larger crops provide more context at greater computational cost. Adaptive resolution offers another option: preserving enough coarse context for localization while concentrating fine-resolution processing around spatially demanding structures.

Segmentation models are commonly trained using cross-entropy loss for per-location classification. For a voxel or pixel u with ground-truth class y_u and predicted class probability $p_{u,c}$, the cross-entropy loss is

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{u} \log p_{u,y_u}. \tag{2.7}$$

Dice loss and the Dice score are widely used in medical segmentation because they measure region overlap and are less dominated by background than plain accuracy [47], [57]. For a binary foreground class, the soft Dice score can be written as

$$\mathrm{Dice} = \frac{2\sum_u p_u y_u + \epsilon}{\sum_u p_u + \sum_u y_u + \epsilon}, \tag{2.8}$$

where p_u is the predicted foreground probability and y_u is the binary ground-truth label.

2.8 Chapter Summary

This chapter introduced the task settings and methodological foundations of the included papers: genomic and echocardiographic prediction, two- and three-dimensional semantic segmentation, learned representations, reinforcement learning, convolutional networks, and transformers.

Chapter 3 examines prediction under partial, variable, and costly observations. Chapter 4 then focuses on structure-aware visual representation learning and the allocation of spatial resolution and computation.

Image Credits and Licenses

The echocardiographic sequence in Figure 2.1a is adapted from *Ultrasound of human heart apical 4-cahmber view.gif* by Fruehaufsteher2, licensed under CC BY-SA 3.0. The panel was created by extracting frames from the original animation, cropping and resizing them, and arranging them into a temporal sequence.

Figure 2.1b is an illustrative AMR example constructed for this thesis. It visualizes example gene-presence indicators, metadata, and susceptibility phenotypes.

The natural-image segmentation example in Figure 2.2a uses MS COCO image `000000007108.jpg`, which is marked in the COCO metadata as having no known copyright restrictions, together with the corresponding COCO-Stuff semantic annotation, licensed under CC BY 4.0. The panel was created by resizing and arranging the image and semantic label map side by side.

The volumetric medical segmentation example in Figure 2.2b uses adapted image and annotation data from the Medical Segmentation Decathlon dataset, licensed under CC BY-SA 4.0. The panel was created by extracting and windowing a CT slice, overlaying the segmentation, rendering the segmentation as a 3D surface, and arranging the views side by side.

CHAPTER 3

Learning from Partial, Irregular, and Costly Observations

Supervised deep learning is commonly formulated for samples with a fixed input structure: each sample contains the same types of observations, and each input position has a consistent meaning across the dataset. Clinical and biomedical data often depart from this formulation. An echocardiographic examination may contain a variable set of videos, while a targeted diagnostic test may deliberately measure only a subset of the available genomic markers. Metadata may also be missing, and the target labels required for training may be only partially observed.

This chapter examines how these conditions affect model design and learning. The multi-stream LVEF prediction paper (Paper E) addresses prediction from observations produced by a variable clinical workflow. GenoARM (Paper A) addresses measurement selection, where genomic markers are chosen before prediction according to available context. GenoARM also operates with missing metadata and partially observed antibiotic susceptibility labels.

3.1 The Fixed-Input Assumption

Standard supervised learning represents each sample as an input $x_i \in \mathcal{X}$ and learns a function $f : \mathcal{X} \to \mathcal{Y}$ from examples (x_i, y_i). In many practical models, $\mathcal{X}$ is implemented as a fixed-size vector or tensor whose positions have consistent meanings across samples. In image classification, this may mean representing all images using the same dimensions and channels. In tabular prediction, it may mean that every sample contains the same measured variables. This common structure simplifies architecture design, batching, and optimization.

In many applications, however, the input is produced by an acquisition process. Let

$$\mathcal{O} = \{1, \ldots, M\}$$

denote the possible observation types. In a fixed-input setting, every sample contains one observation of each expected type. In a variable-input setting, the acquired collection A_i may omit expected observations, contain repeated observations, or include uncertain type assignments.

A different situation arises when acquisition is deliberately restricted. Rather than receiving all possible measurements, the predictor is given measurements from a selected subset

$$S_i \subseteq \mathcal{O}.$$

The subset may be shared across all samples or chosen according to context associated with the sample. In either case, measurement selection determines what information will be available before prediction is performed.

Input availability should be distinguished from supervision availability. A model may produce a fixed set of outputs even though only some target labels are observed for each training sample, in which case the loss is computed only from the available labels. GenoARM combines deliberate measurement restriction with missing metadata and partial supervision.

Figure 3.1 summarizes the distinction between a complete fixed input, a variable acquired input, and a deliberately selected measurement set.

A related fixed-structure assumption appears inside visual models. Convolutional layers and many vision-transformer operations are designed for features arranged on dense, regular grids. ARTA (Paper B) and BATS (Paper C) instead construct sparse, mixed-resolution token sets whose spatial structure varies with the input. This is not an acquisition problem—the original image

$$\mathcal{O} = \{1, 2, 3, 4, 5\}$$

Observation regime	Example input relative to $\mathcal{O}$
Fixed input	$A_i = [1, 2, 3, 4, 5]$
Variable clinical input	$A_i = [1, 2, 2, 5^?]$
Selected measurements	$S_i = \{2, 4\} \subset \mathcal{O}, \quad A_i = [2, 4]$

Figure 3.1: Input regimes discussed in this chapter. Fixed-input models receive all expected observation types. Variable clinical inputs may omit, repeat, or uncertainly specify observations, illustrated by the repeated entry 2 and uncertain entry $5^?$. Selected-measurement models deliberately restrict acquisition to a subset S_i, which may be fixed globally or chosen according to contextual information.

or volume is still available—but it creates a related architectural problem: subsequent layers must operate on an irregular representation. Chapter 4 examines this setting.

3.2 Variable Multi-View Prediction in Clinical Imaging

Echocardiography provides a concrete example of prediction from irregular clinical observations. It is widely used for cardiovascular evaluation because it is accessible, non-invasive, and cost-effective [1]. One important quantity derived from echocardiography is left ventricular ejection fraction (LVEF), which describes the percentage of the left ventricular end-diastolic volume ejected during contraction. LVEF is commonly estimated using biplane Simpson's method [2], but in clinical practice it may also rely on visual assessment. Interobserver variability remains a practical limitation, particularly when image quality is poor or the reader is less experienced [3]. Deep learning methods have therefore been investigated as a way to support more automated and consistent assessment [4].

Existing approaches to LVEF prediction differ in how they use echocardiographic views. Some methods segment cardiac chambers and compute LVEF from the resulting measurements [58], [59], [60], [61], while others predict LVEF

directly without an intermediate segmentation step [62], [63], [64], [65], [66], [67]. Many works use curated datasets with known view labels, and several focus on a single view, especially the apical four-chamber view [62], [64], [66], [67], [68]. Others use two or more views from predefined sets [58], [59], [60], [61], [63], [65]. These studies establish the usefulness of deep learning for LVEF estimation, but many of them assume that the relevant views are known, fixed, and available.

Routine clinical echocardiographic data often violate this assumption. A full examination may contain multiple videos from several views, but their number and quality vary. Expected views may be absent, several videos may correspond to the same view, and an automatic view-classification system may assign an incorrect label. Recordings also differ in ultrasound system, examining physician, frame rate, and image quality. The problem therefore combines elements of multi-view learning, missing-modality learning, and set-based aggregation, while adding uncertainty from the clinical acquisition and preprocessing pipeline.

The multi-stream LVEF prediction paper (Paper E) uses four automatically selected standard transthoracic 2DE views: apical two-chamber, apical three-chamber, apical four-chamber, and parasternal long axis. Each view is represented by both its image sequence and corresponding optical flow, resulting in eight input streams. Shared weights reduce dependence on separate view-specific parameters, while the training procedure exposes the model to alternative view combinations. When an expected view is unavailable, another available video is substituted according to a predefined priority order, and duplicate views are used when constructing training inputs. The resulting model can exploit complementary information from multiple views without requiring every examination to contain the same complete and correctly identified view set.

3.3 Context-Dependent Measurement Selection

Antimicrobial resistance prediction raises a different observation problem. Here, the relevant measurements may exist in principle, but acquiring all of them can be too slow or expensive for routine diagnostic use. Antimicrobial resistance is a major global health challenge, and diagnostic tools that support timely, targeted antibiotic treatment are therefore clinically valuable [5], [17]. Because

resistance phenotypes are often associated with acquired resistance genes and chromosomal mutations, genomic information can be highly informative for predicting antimicrobial susceptibility [15], [69].

Machine-learning approaches, including deep models, have shown that whole-genome sequence data can support accurate antimicrobial resistance prediction [8], [9], [10], [11], [40]. However, whole-genome sequencing may not always be suitable for routine, time-sensitive diagnostics because the complete workflow requires sequencing infrastructure, bioinformatic processing, validation, and quality assurance [15], [70]. Targeted molecular tests such as PCR can provide a faster and more focused alternative, but they require the relevant resistance genes or mutations to be specified in advance [71]. This changes what has to be learned. Rather than assuming that whole-genome information is available, the model must support prediction from a deliberately limited measurement set.

Let $\mathcal{G}$ denote the set of candidate genes or mutations, and let $S \subseteq \mathcal{G}$ be a selected panel with at most K measurements. The idealized selection problem can be written as

$$S^* = \arg \max_{S \subseteq \mathcal{G},\, |S| \leq K} U(S), \tag{3.1}$$

where $U(S)$ measures the predictive usefulness of a model that only has access to the measurements in S. This objective is combinatorial: even for a moderate number of candidate markers, the number of possible panels is too large to evaluate exhaustively.

Existing work approaches parts of this problem from several directions. One is AMR prediction from genomic data, where most approaches use full gene sequences or gene-level representations [8], [9], [10], [11]. Another is the identification of important resistance genes through genetic algorithms or full-genome analyses [72], [73], [74]. More broadly, the problem connects to feature selection and black-box optimization [75], [76], [77], [78].

The problem also relates to active feature acquisition, where a model decides which features to observe before predicting. Many such methods formulate information acquisition as a sequential decision problem, deciding whether and what to observe before making a prediction [79], [80], [81]. Other methods select features after observing the full vector [82], or assume that each acquired feature guides the next acquisition decision [83], [84]. These assumptions do not map cleanly onto targeted gene testing. At deployment, the measurements in a panel are acquired together rather than sequentially, so each observed

gene cannot guide the acquisition of the next. A greedy or wrapper-based selection procedure could also build a gene set iteratively, but evaluating each candidate subset through model training is expensive, and such approaches do not learn a reusable selection policy conditioned on the current subset.

The core difficulty is constructing a compact gene panel that works well across multiple resistance phenotypes. With thousands of genes and mutations associated with antibiotic resistance [12], many markers are rare, phenotype-specific, or redundant with other markers. Two genes may co-occur frequently or indicate the same resistance mechanism, so measuring both adds little. A rare gene may be important if it captures resistance cases that more common markers miss. A useful panel should therefore balance informativeness and complementarity: genes that together cover resistance phenotypes, rather than simply the strongest individual predictors.

The general formulation above does not specify whether the same panel is used in every context. In GenoARM, panel selection is conditioned on metadata. Let m_i denote the country of origin, sample source, and collection year associated with isolate i. The selection policy defines

$$S_{m_i} = \pi(m_i) \subseteq \mathcal{G}, \tag{3.2}$$

where $\mathcal{G}$ is the set of candidate genomic markers. Isolates with the same metadata configuration receive the same panel, while different configurations may lead to different panels. The policy may nevertheless produce identical panels when it learns that a difference in metadata does not affect which markers are useful.

Three types of variables must be distinguished in this setting. First, the genes and mutations are the measurements selected by the policy. Second, country, sample source, and year are contextual metadata used both for panel selection and resistance prediction; these variables may themselves be missing, in which case a dedicated missing category is used. Third, antibiotic susceptibility outcomes are the target labels. These labels are also incomplete because each isolate is typically tested against only a subset of the antibiotics considered by the model.

The resistance predictor outputs probabilities for the complete antibiotic set, even though only some susceptibility labels are observed for a given isolate. Let $\mathcal{C}_i$ denote the antibiotics with observed labels for isolate i. Its training loss

is computed as

$$\mathcal{L}_i = \frac{1}{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} \ell(\hat{y}_{ic}, y_{ic}) \,, \tag{3.3}$$

where $\hat{y}_{ic}$ is the predicted resistance probability and ℓ is the per-antibiotic loss. Predictions for antibiotics without observed susceptibility outcomes do not contribute to the loss.

GenoARM (Paper A) constructs each metadata-conditioned panel through a sequence of conditional choices. Although the final panel is an unordered set, it is built iteratively: at each step, the value of adding a genomic marker depends on the markers already selected. The predictive performance obtained from the selected markers and available metadata provides the reward for the selection policy. The predictor therefore evaluates the usefulness of a panel, while the policy learns which genomic measurements are complementary under a given metadata context.

3.4 Open Challenges

Realistic evaluation remains difficult in both settings. For echocardiography, artificially dropping or duplicating views supports controlled experiments but may not reproduce the joint distribution of missingness, image quality, acquisition variation, and view-classification errors found in routine examinations. Evaluation should therefore include naturally variable clinical data and examine performance across view combinations, acquisition conditions, and patient populations.

For AMR prediction, metadata-conditioned panels may not transfer unchanged across hospitals, countries, time periods, pathogen populations, or resistance databases. Evaluation must also account for incomplete susceptibility testing and determine whether the observed labels adequately represent the antibiotics and isolates on which the system would be used. Clinical adoption further requires that selected markers correspond to feasible tests and can be related to known resistance mechanisms.

3.5 Chapter Summary

This chapter distinguished between prediction from variably acquired observations and prediction using deliberately selected measurements. Paper E

aggregates an irregular set of echocardiographic views, while GenoARM selects metadata-conditioned genomic panels, accommodates missing metadata, and learns from partially observed susceptibility labels. The next chapter turns to a related departure from fixed structure within visual models, focusing on dense, hybrid, and adaptive mixed-resolution representations.

CHAPTER 4

Adaptive Representation Learning for Efficient Segmentation

Semantic segmentation requires spatially resolved predictions. Unlike image-level prediction, where a model may compress the input into a single global representation, segmentation models must preserve spatial information throughout the network. This creates a tension between local precision and broader context: boundaries, thin structures, small objects, and lesions require fine detail, while correct interpretation often depends on surrounding anatomy or scene structure. In high-resolution images and volumetric medical data, maintaining both becomes computationally expensive.

This chapter examines three approaches to balancing spatial detail, context, and computational cost. SwInception (Paper D) introduces fixed local and multi-scale architectural priors for volumetric segmentation. ARTA (Paper B) makes token resolution input-dependent in two-dimensional semantic segmentation, and BATS (Paper C) extends this principle to volumetric medical data.

4.1 The Cost of Spatial Detail in Semantic Segmentation

In a segmentation map, the difficult regions are rarely spread uniformly across the input. Class boundaries, thin structures, small objects, and lesions may occupy only a small part of an image or volume, but they often determine the quality of the final prediction. Large homogeneous regions are usually less demanding. In natural images, sky, road, walls, and object interiors may be locally homogeneous. In medical volumes, background, air, and healthy tissue can often be labelled without maintaining full-resolution processing everywhere.

A fixed resolution schedule is therefore a compromise. Processing the whole input finely preserves detail, but spends computation and representational capacity on regions that may not require it. Processing the input coarsely is more efficient, but risks losing precisely the boundaries and small structures that make segmentation difficult.

Most segmentation architectures still process visual data according to a predetermined spatial schedule. A convolutional network applies filters over a regular grid. A vision transformer partitions the image into fixed-size patches. A volumetric model processes a dense voxel grid or dense crop. These choices are practical: they preserve spatial regularity, make batching straightforward, and are well supported by current hardware. However, they allocate computation according to the grid rather than according to the semantic complexity of the image.

This creates both a computational and a representational cost. The computational cost is direct: high-resolution feature maps are expensive, especially in 3D, where increasing resolution affects three spatial dimensions. The representational cost is subtler. A token spanning a class boundary must represent several things at once: the classes present, their relative proportions, their spatial arrangement, and the geometry of the boundary. A token covering a mostly homogeneous region can instead represent the appearance and context of a single dominant class. With fixed tokenization, both regions receive the same representational budget even though they place very different demands on the model.

A simple way to describe the allocation problem is to view an image or volume as a set of spatial regions represented by tokens. Let a token t_i cover a

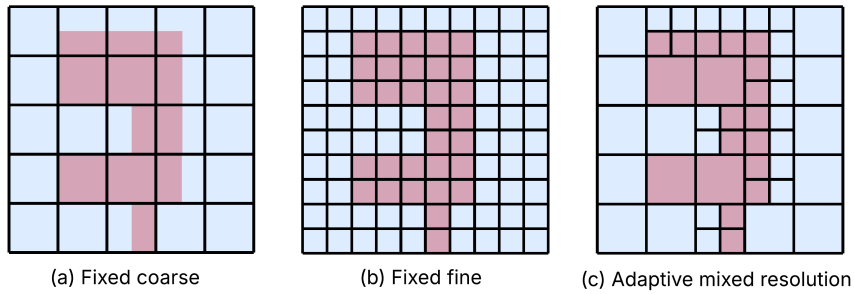


Figure 4.1: Spatial resolution as an allocation problem in segmentation. A fixed coarse representation is efficient but under-resolves boundaries and thin structures, whereas a fixed fine representation preserves detail at higher cost. Adaptive hierarchical representations add finer tokens in spatially complex regions while retaining coarser tokens for multi-scale context.

spatial region R_i at resolution level s_i. In a fixed-resolution model, all regions at the same stage share the same resolution level, regardless of their content. In a mixed-resolution model, the resolution level becomes input-dependent:

$$\mathcal{T}(x) = \{(R_i, s_i) : i = 1, \ldots, N(x)\}, \tag{4.1}$$

where both the number of tokens $N(x)$ and their resolutions may depend on the input x. The token set need not form a non-overlapping partition at a single scale. In the hierarchical representations used by ARTA and BATS, spatial regions may be represented simultaneously at several resolutions: refining a region introduces finer tokens while retaining its coarser ancestors.

Figure 4.1 illustrates the finest resolution allocated to each region. In ARTA and BATS, the corresponding coarser tokens are retained to provide multi-scale context.

4.2 Multi-Scale Representations for Context and Detail

Most segmentation architectures address this tension through fixed multi-scale hierarchies. Early layers in a visual model preserve relatively high spatial resolution and capture local patterns, while deeper layers operate

at lower resolution and capture broader semantic context. Encoder-decoder architectures use this hierarchy by reducing resolution to extract abstract features, then restoring resolution to produce spatially aligned predictions.

This design is central to semantic segmentation because neither local detail nor global context is sufficient on its own. High-resolution features support precise localization, but may lack the context needed to distinguish visually similar structures. Low-resolution features capture broader context, but can blur boundaries and lose small objects. Architectures such as U-Net [13] and feature-pyramid networks [85] formalize the combination through skip connections, lateral connections, and top-down feature fusion.

Multi-scale processing is useful across segmentation settings because the model must combine local spatial detail with broader context. In natural-image segmentation, this helps represent objects, parts, boundaries, and scene layout at different spatial scales, as in ARTA (Paper B). In medical segmentation, the scale differences can be sharper: large organs require anatomical context, while small lesions, vessels, and boundaries require fine spatial detail. This is the setting of SwInception (Paper D) and BATS (Paper C), which address volumetric segmentation through fixed multi-scale architectural structure and adaptive mixed-resolution allocation, respectively.

The limitation of fixed multi-scale processing is that the hierarchy is applied uniformly. Every input passes through the same resolution schedule, regardless of where boundaries, small structures, or homogeneous regions occur. ARTA (Paper B) and BATS (Paper C) retain the hierarchical representation but make its depth spatially adaptive. If a region is refined to level s, it is represented by tokens at levels $1, \ldots, s$, where level 1 is the coarsest. Adaptive allocation therefore determines where finer levels are added, while coarser tokens remain available for contextual reasoning.

4.3 Inductive Biases and Hybrid Visual Architectures

Representation learning is shaped by data and objectives, but also by inductive biases: structural assumptions that make some solutions easier for the model to learn than others. In visual learning, these assumptions matter because images and volumes have strong spatial structure. They matter even more in medical imaging, where annotated datasets are often small and variation

across scanners, protocols, and patient populations can be substantial.

Convolutional networks encode locality and translation equivariance. Local filters assume that nearby pixels or voxels are related; weight sharing assumes that useful local patterns may occur at many spatial locations. These properties make CNNs effective and data-efficient for local visual representation learning.

Transformer-based models make a different trade-off. Self-attention allows tokens to interact according to learned similarity rather than fixed local neighborhoods, which gives the model more flexibility in how it combines information. Plain attention has weaker built-in spatial structure than convolution, so vision transformers usually introduce additional structure through patch embeddings, positional encodings, hierarchy, local or windowed attention, or convolutional components. These choices affect both computation and representation: they determine which interactions are cheap, which are direct, and which must be learned from data.

Hybrid CNN–transformer architectures combine these two sources of structure. Convolutional components provide local feature extraction, spatial regularity, and data-efficient inductive bias. Transformer components provide more flexible interaction between features. This combination is useful for segmentation because the task requires both local spatial precision and broader contextual reasoning.

SwInception (Paper D) applies this hybrid approach to volumetric medical image segmentation. It builds on the Swin Transformer, which introduces locality through windowed self-attention, and adds Inception-style convolutional branches inside the feed-forward layers. The branches process features using several receptive fields, introducing local and multi-scale spatial priors directly within each transformer block.

Adaptive mixed-resolution models require a different architectural capability: attention must operate efficiently on token sets that are sparse, irregular, and hierarchical.

4.4 Efficient and Sparse Attention for Segmentation

Transformer-based segmentation models need mechanisms for feature interaction, but full self-attention is expensive. For N tokens with feature dimension d, full self-attention forms an $N \times N$ attention matrix and has cost on the

order of

$$\mathcal{O}(N^2 d). \tag{4.2}$$

This becomes costly for dense high-resolution segmentation features, especially in volumetric data where the number of spatial elements grows cubically with resolution.

A common response is windowed attention. The Swin Transformer [48] computes attention within local windows rather than across the full image, reducing the cost of attention for a fixed window size. Shifted windows allow information to pass across window boundaries in successive layers. This makes Swin useful for segmentation because it combines some of the locality of convolution with the content-dependent interaction of attention.

Cluster attention, introduced in AutoFocusFormer [49], supports attention over sparse and non-uniform token sets. Token centres are ordered spatially using a space-filling curve and partitioned into balanced clusters containing the same number of tokens. This produces compact groups without the padding overhead associated with clusters of unequal size.

For each query token, a fixed number of spatially nearby clusters is selected based on their centroids. The union of the tokens in these clusters forms the query's key–value neighbourhood. Different queries may therefore use overlapping neighbourhoods, allowing information to pass across cluster boundaries while keeping the attention operation local. Unlike windowed attention, this construction does not require tokens to occupy every position of a dense regular grid. Tokens representing regions of different sizes can therefore participate in the same attention operation, which is essential for the irregular mixed-resolution token sets used by ARTA (Paper B) and BATS (Paper C).

Other efficiency methods reduce the number of tokens that reach later layers. Token pruning removes tokens estimated to be uninformative, while token merging combines tokens that appear redundant. These approaches can reduce later computation, but many of them start from a dense representation. For segmentation, and especially for volumetric segmentation, that means part of the high-resolution cost has already been paid before pruning or merging occurs.

These methods reduce cost at different points in the pipeline. Windowed and cluster attention make token interaction cheaper. Token pruning and merging reduce the token set after features have been computed. Adaptive mixed-resolution allocation, as used in ARTA (Paper B) and BATS (Paper C),

changes where fine tokens are introduced in the first place.

4.5 Adaptive Computation and Mixed-Resolution Processing

In segmentation, not all spatial regions need the same resolution. Boundaries, thin structures, and object-dense areas benefit from fine detail; homogeneous regions can often be represented more coarsely. Adaptive computation methods use this imbalance to concentrate representational capacity where it is more likely to affect the prediction.

In a fixed tokenization, the number of tokens at a given resolution is determined by the input size and patch size. For a two-dimensional image of size $H \times W$ and patch size p, this gives approximately

$$N_{\mathrm{2D}} \approx \frac{HW}{p^2} \tag{4.3}$$

tokens. Reducing the patch size increases spatial detail, but also increases the number of tokens that must be processed. Adaptive mixed-resolution processing changes this relationship by allowing only selected regions to move to finer token scales, while other regions remain coarse.

Fixed multi-scale architectures process every input through the same hierarchy, while pruning and merging methods typically begin with a dense token set and reduce it after features have already been computed. ARTA (Paper B) instead begins with a coarse representation and selectively extends the hierarchy to finer scales.

ARTA uses an allocator that predicts class-boundary evidence and identifies which spatial regions require additional resolution. Regions likely to contain semantic transitions are subdivided, while homogeneous regions remain coarse. Importantly, refinement adds tokens rather than replacing the existing representation. A region refined to level s retains its ancestor tokens at levels $1, \ldots, s-1$, allowing fine local features to coexist with progressively broader context.

The resulting tokens differ in both size and spatial arrangement and no longer form a dense regular grid. ARTA processes them using cluster attention, which groups nearby mixed-resolution tokens into balanced local neighborhoods. The allocator therefore determines where additional spatial detail is introduced,

while cluster attention allows information to be exchanged within the resulting irregular hierarchy.

4.6 From Two-Dimensional to Volumetric Segmentation

Volumetric segmentation makes the allocation problem harder. Increasing resolution in 3D affects height, width, and depth, so the number of fine-scale elements grows much faster than in 2D. Full-volume processing at high resolution is usually infeasible, and practical systems rely on crop-based training and sliding-window inference. These choices control memory use, but they also limit how much context the model sees at one time.

The scaling problem is stronger in volumetric data. For a three-dimensional volume of size $H \times W \times D$ and cubic patch size p, the number of tokens is approximately

$$N_{\mathrm{3D}} \approx \frac{HWD}{p^3}. \tag{4.4}$$

Reducing the patch size therefore increases the number of tokens along three spatial dimensions. This cubic growth makes uniform fine-resolution processing much more costly than in two-dimensional images, and makes missed allocation decisions more consequential: a coarse 3D token may cover a large spatial region that contains a small lesion or thin structure.

The target structure often occupies a small fraction of the scan. A lesion, vessel, or small organ may be surrounded by large regions of background or healthy tissue. The model still needs enough coverage to find the target, but most of its capacity should be spent near the target or its boundary. This is the same motivation as in 2D segmentation, but the consequences of a poor allocation are more severe.

Extending adaptive allocation from 2D to 3D exposes three concrete problems. First, sequential coarse-to-fine allocation can make early errors irreversible. If a region is not selected at a coarse level, it is not evaluated at any of the finer levels that descend from it. This vulnerability is present in ARTA and becomes more consequential in three dimensions, where a coarse token covers a larger spatial region and may contain an entire small lesion or thin structure.

Second, fine-token growth is cubic. As resolution increases, local neighborhoods can become dominated by fine tokens, weakening the intended interaction between coarse context and fine detail. Third, positional encoding and relative position bias become more cumbersome because the number of possible spatial offsets grows substantially in three dimensions.

BATS (Paper C) changes both allocation and cross-scale interaction to address these issues. A dense boundary predictor scores all refinement levels in parallel, so fine-scale evidence is evaluated even when the corresponding coarse region receives a low score. A fine-first cascade then converts these predictions into a valid hierarchical mixed-resolution token set.

BATS also introduces parent cluster attention. Standard cluster attention provides local interaction between spatially nearby mixed-resolution tokens, but a neighborhood dominated by fine tokens may contain little coarse context. Parent cluster attention therefore injects the coarser hierarchical ancestors of each query token directly into its attention neighborhood. Fine tokens consequently receive deterministic cross-scale context rather than relying on their ancestor tokens to occur in the local cluster neighborhood.

BATS also adapts positional encoding to the volumetric setting using three-dimensional rotary positional encoding. Rotary transformations are applied independently along the three spatial axes, replacing the additive relative-position bias used in the original cluster-attention formulation.

BATS retains ARTA's principle of concentrating fine-resolution processing in spatially demanding regions, but modifies the allocation and attention mechanisms for three-dimensional data. Parallel scale scoring reduces dependence on early coarse decisions, while parent cluster attention preserves reliable cross-scale context as the number of fine tokens grows.

4.7 Open Challenges

The allocation criterion itself remains an open design choice. ARTA and BATS use predicted class-boundary evidence as a proxy for where additional resolution is valuable: regions likely to contain more than one class are refined, while regions predicted to be homogeneous remain coarse. This is intuitive for segmentation, but it does not directly optimize the final accuracy–efficiency trade-off. Fine resolution may also be useful because of uncertainty, small isolated targets, ambiguous texture, or task-specific clinical importance, while

some class boundaries may not require the finest representation. Alternative allocators could therefore estimate the expected benefit of refinement more directly rather than relying only on boundary evidence.

The efficiency of adaptive allocation also depends on how well irregular token sets can be processed. Cluster construction, neighbourhood search, routing, and irregular memory access introduce overhead that can offset the reduction in token count. Depending on the dataset and resulting token density, adaptive processing may therefore be faster or slower than a dense baseline. Improving sparse attention and mixed-resolution implementations is consequently an important direction, especially for volumetric data where the potential savings are large but the cost of inefficient token interaction is also greater.

Allocation errors remain another important failure mode. Sequential coarse-to-fine allocation, as used in ARTA, is particularly vulnerable because a region missed at an early stage cannot be recovered at a finer level. BATS reduces this dependence by scoring all levels in parallel, but can still under-refine regions when the boundary predictor fails to identify relevant fine-scale evidence. Allocation maps and segmentation failures should therefore be examined together, especially for small structures and lesions.

Finally, allocation behavior may not transfer unchanged between datasets, modalities, or annotation protocols. A model may require different refinement patterns when object scales, boundary complexity, or target prevalence change. Evaluating adaptive representations therefore requires studying both final segmentation performance and how the learned allocation changes across domains.

4.8 Chapter Summary

This chapter described fixed and adaptive approaches to combining spatial detail with broader context in semantic segmentation. SwInception strengthens a fixed volumetric hierarchy using local and multi-scale convolutional priors. ARTA constructs hierarchical mixed-resolution token sets through input-dependent refinement, while BATS adapts this principle to three-dimensional data through parallel scale scoring and parent cluster attention.

CHAPTER 5

Summary of included papers

This chapter provides a summary of the included papers.

5.1 Paper A

David Hagerman, Anna Johnning, Roman Naeem, Fredrik Kahl, Erik Kristiansson, Lennart Svensson
Optimizing gene-based testing for antibiotic resistance prediction
Proceedings of the AAAI Conference on Artificial Intelligence
Vol. 39, No. 27, pp. 28033-28041, 2025

Summary: This paper presents GenoARM, a method for antimicrobial resistance prediction when only a limited set of genomic measurements can be acquired. Instead of assuming that complete genomic information is available, the method treats gene-test selection and resistance prediction as a joint problem. A reinforcement learning policy constructs compact gene panels by selecting genes or mutations sequentially, where each choice is evaluated in the context of the genes already selected. This allows the model to account for

redundancy between resistance markers and to prioritize measurements that are jointly informative across multiple antibiotics. The selected panel is then used by a prediction model to estimate resistance probabilities. Experiments show that GenoARM can identify small subsets of genomic features that retain strong predictive performance, supporting the use of learned gene-panel design for faster and more cost-effective antimicrobial resistance testing.

Contributions: The project was initiated in collaboration between David, Lennart, Anna, and Erik. David developed the main method, led the implementation and experimental evaluation, analyzed the results, and wrote the main manuscript draft. Lennart, Fredrik, and Roman contributed to method development, result interpretation, additional experiment planning, and manuscript revision. Anna and Erik contributed expert knowledge on genomics and antimicrobial resistance, supported interpretation of the findings, and contributed to manuscript revision. All authors contributed to and approved the final version of the paper.

5.2 Paper B

David Hagerman[*], Roman Naeem[*], Erik Brorsson, Fredrik Kahl, Lennart Svensson
ARTA: Adaptive Mixed-Resolution Token Allocation for Efficient Dense Feature Extraction
arXiv Preprint
arXiv:2603.26258, 2026.

[*]Shared first authorship.

Summary: This paper introduces ARTA, an adaptive vision Transformer for semantic segmentation that represents images with tokens at different spatial resolutions. Instead of processing the image with a fixed dense token grid, the model begins from a coarse representation and progressively refines regions predicted to contain semantic complexity, such as object boundaries and fine structures. This produces a sparse mixed-resolution token set in which simple regions remain compact while more difficult regions are represented in greater detail. The tokens are then refined jointly using attention over local mixed-resolution neighborhoods, allowing coarse contextual information and fine spatial detail to interact. Experiments on standard semantic segmentation

benchmarks show that ARTA maintains competitive segmentation accuracy while reducing computational cost, demonstrating the value of allocating resolution according to image content rather than using a uniform processing schedule.

Contributions: David and Roman contributed to idea generation, model implementation, result analysis, preparation of figures, and paper writing. David conducted the experiments. Erik, Fredrik, and Lennart contributed to idea generation, result analysis, and paper writing.

5.3 Paper C

David Hagerman, Roman Naeem, Fredrik Kahl
BATS: Resource-Efficient Volumetric Segmentation with Boundary-Aware Mixed-Resolution Tokens
arXiv Preprint, 2026

.

Summary: This paper presents BATS, a volumetric medical image segmentation architecture that concentrates fine-resolution processing near predicted class boundaries, thin structures, and small targets while representing homogeneous regions more coarsely. A dense boundary predictor scores all refinement levels in parallel, and a fine-first context cascade converts these predictions into an input-dependent hierarchy of mixed-resolution tokens containing each selected token and its coarser ancestors. This avoids hard gating by coarse decisions, allowing valid fine-scale evidence to be retained even when the corresponding coarse region is not selected. The resulting sparse hierarchy is processed using parent cluster attention, which injects hierarchical ancestor tokens into local attention neighbourhoods to provide deterministic cross-scale context without dense multi-scale refiner features or token-level nearest-neighbour search. BATS is evaluated on five public CT and MRI datasets under the standardised nnU-Net Revisited protocol. It achieves the highest LiTS Dice among the compared methods and averages 0.37 Dice points below MedNeXt-L k3 across all five datasets. Relative to this dense baseline, BATS reduces peak allocated GPU memory by more than 53% on KiTS, LiTS, and BraTS, and reduces inference time by 30.3% on KiTS and 27.2% on LiTS, although it is slower on the more token-dense BraTS.

Contributions: David initiated the project, developed the main method, led the implementation and experimental evaluation, analyzed the results, and wrote the main manuscript draft. Fredrik and Roman contributed through regular discussion of ideas, interpretation of results, suggestions for additional experiments, and manuscript revision. All authors contributed to and approved the final version of the paper.

5.4 Paper D

David Hagerman, Roman Naeem, Jakob Lindqvist, Carl Lindström, Fredrik Kahl, Lennart Svensson
SwInception - Local Attention Meets Convolutions
Published in International Conference on Pattern Recognition and Artificial Intelligence
pp. 3–17, Nature Singapore, 2024

Summary: This paper presents SwInception, a hybrid CNN–Transformer architecture for volumetric medical image segmentation. The method builds on the Swin Transformer framework, which uses local windowed self-attention to model contextual interactions efficiently, and strengthens it with Inception-style convolutional branches inside the network blocks. These branches introduce additional local and multi-scale spatial processing, giving the model a stronger inductive bias for three-dimensional medical images where annotated data are often limited. The architecture follows an encoder-decoder design for dense voxel-level prediction, combining attention-based feature interaction with convolutional feature extraction at multiple scales. Experiments on volumetric medical segmentation tasks show that SwInception improves segmentation performance compared with several established CNN- and Transformer-based baselines.

Contributions: David initiated the project, developed the main method, led the implementation and experimental evaluation, analyzed the results, and wrote the main manuscript draft. Lennart, Fredrik, Roman, Jakob and Carl contributed through regular discussion of ideas, interpretation of results, suggestions for additional experiments, and manuscript revision. All authors contributed to and approved the final version of the paper.

5.5 Paper E

Jennifer Alvén, Eva Hagberg, **David Hagerman**, Richard Petersen, Ola Hjelmgren
A deep multi-stream model for robust prediction of left ventricular ejection fraction in 2D echocardiography
Scientific Reports
Vol. 14, No. 1, 2024

Summary: This paper presents a multi-stream deep learning method for predicting left ventricular ejection fraction from clinical 2D echocardiographic examinations. The method uses four standard transthoracic views and processes both the original image sequences and optical flow, allowing the model to use information about cardiac appearance as well as motion. Since routine echocardiographic examinations may contain missing, duplicated, or misclassified views, the model is designed to be less dependent on receiving a complete and perfectly labelled view set. It uses shared weights across views, pre-training, and sampling strategies that expose the model to different view configurations during training. The paper shows that combining multiple views with motion-derived input can support robust LVEF prediction from clinically variable data, and contributes to the thesis by addressing prediction from incomplete and heterogeneous observations while making cardiac motion explicit in the learned representation.

Contributions: Jennifer and Ola initiated the project and contributed to the original idea and study direction. David contributed to early method exploration together with Jennifer and led the main model implementation, experimental setup, and core experimental evaluation. Jennifer, David, Eva, and Ola regularly discussed results, prioritized follow-up experiments, refined the method, interpreted the findings, and contributed to writing the manuscript. Eva and Ola also contributed medical expertise and dataset annotation/generation. Jennifer and Richard contributed to extending the implementation to additional open datasets, supporting additional experiments, and strengthening the revised manuscript. All authors contributed to manuscript revision and approved the final version.

CHAPTER 6

Concluding Remarks and Future Work

This thesis studied how deep learning models can adapt when information is incomplete, costly to acquire, implicit in the input, or unevenly distributed across space. The included papers addressed this through context-dependent genomic measurement selection, prediction from variable echocardiographic examinations, fixed structure-aware volumetric architectures, and adaptive allocation of spatial resolution. The results show that representation and acquisition choices can be learned rather than treated as fixed, but also that these choices introduce new failure modes and deployment constraints.

6.1 From Gene Panels to Diagnostic Tests

GenoARM selects genomic measurements according to contextual metadata, but a clinical laboratory cannot maintain a separate PCR panel for every possible combination. A more realistic formulation would constrain the system to a limited number of deployable panels. Given a budget of N panels, the problem becomes deciding which contexts justify specialized panels and which should share a general panel.

Panel selection could also be conditioned on test results that are already

available. An initial test may narrow the plausible resistance mechanisms and make a more targeted follow-up panel useful. However, every additional conditioning variable increases the number of panels that may need to be designed, validated, and maintained. Future work should therefore optimize predictive performance jointly with panel count, assay cost, test availability, and stability across pathogen populations, locations, and time periods.

6.2 Robust Prediction from Echocardiographic Examinations

The multi-stream LVEF model learns robustness mainly from the variation naturally present in clinical data. Rare input configurations may therefore receive little training signal; for example, some standard views may be absent only infrequently. Robustness could be trained more explicitly by augmenting the available view set through view dropping, duplication, substitution, or alternative combinations sampled from the same examination.

Future models could also incorporate view-quality estimates and view-classification confidence during aggregation. Low-quality or uncertain videos could then contribute less to the final estimate, and the model could indicate when the available examination does not support a reliable prediction. Evaluation should ultimately cover the complete pipeline from acquisition and automatic view selection to prediction and clinical interpretation.

6.3 Learning Where to Allocate Spatial Detail

ARTA and BATS use predicted class boundaries as a proxy for where fine spatial resolution is useful. This is intuitive for segmentation, but it may not produce the best accuracy–efficiency trade-off in every setting. Fine resolution may also be valuable in uncertain regions, around small isolated targets, or where weak contrast makes classification difficult. Future allocators could therefore estimate the expected benefit of refinement more directly rather than relying only on whether a patch contains multiple classes.

A related direction is to couple allocation to the final segmentation objective. The current allocators are trained using an explicit boundary-prediction loss, while the segmentation loss supervises the representation produced after allo-

cation. Ideally, the final task loss would guide both segmentation and token placement, together with a penalty on computational cost. The main difficulty is that allocation decisions are discrete, making credit assignment from the final prediction to individual refinement choices difficult.

Adaptive allocation also depends on efficient sparse processing. Cluster construction, gathering and scattering, and irregular memory access can offset the savings from reducing the token count. This explains why BATS improves runtime on boundary-sparse datasets but can be slower when the retained hierarchy becomes dense. Improving sparse attention and mixed-resolution operators is therefore an important direction alongside improved allocation criteria. These adaptive mechanisms could also be combined with stronger local and multi-scale architectural priors, as explored in SwInception.

6.4 Final Remarks

The central conclusion of this thesis is that the information presented to a model, and the resolution at which it is represented, should be treated as part of the learning problem. Making these choices adaptive can reduce unnecessary measurements and computation, but it also gives the model responsibility for deciding what information is preserved. Future progress therefore depends not only on better selection and allocation mechanisms, but on ensuring that their decisions remain reliable, efficient, and practical in the settings where the models are intended to operate.

References

[1] A. Papolos, J. Narula, C. Bavishi, F. A. Chaudhry, and P. P. Sengupta, "US hospital use of echocardiography: Insights from the nationwide inpatient sample," *J. Am. Coll. Cardiol.*, vol. 67, no. 5, pp. 502–511, 2016.

[2] N. B. Schiller et al., "Left ventricular volume from paired biplane two-dimensional echocardiography," *Circulation*, vol. 60, no. 3, pp. 547–555, 1979.

[3] N. T. Kouris, V. S. Kostopoulos, G. A. Psarrou, P. M. Kostakou, C. Tzavara, and C. D. Olympios, "Left ventricular ejection fraction and global longitudinal strain variability between methodology and experience," *Echocardiography*, vol. 38, no. 4, pp. 582–589, 2021.

[4] G. Litjens et al., "State-of-the-art deep learning in cardiovascular image analysis," *JACC Cardiovasc. Imaging*, vol. 12, no. 8 Part 1, pp. 1549–1565, 2019.

[5] R. Laxminarayan et al., "Antibiotic resistance—the need for global solutions," *The Lancet infectious diseases*, vol. 13, no. 12, pp. 1057–1098, 2013.

[6] J. O'Neill, "Antimicrobial resistance: Tackling a crisis for the health and wealth of nations," *Rev. Antimicrob. Resist.*, 2014.

[7] J. O'Neill, "Tackling drug-resistant infections globally: Final report and recommendations," 2016.

[8] M. Tharmakulasingam, B. Gardner, R. L. Ragione, and A. Fernando, "Explainable Deep Learning Approach for Multilabel Classification of Antimicrobial Resistance With Missing Labels," *IEEE Access*, vol. 10, pp. 113 073–113 085, 2022, Conference Name: IEEE Access, ISSN: 2169-3536.

[9] N. Macesic, O. J. Bear Don't Walk, I. Pe'er, N. P. Tatonetti, A. Y. Peleg, and A.-C. Uhlemann, "Predicting Phenotypic Polymyxin Resistance in Klebsiella pneumoniae through Machine Learning Analysis of Genomic Data," *mSystems*, vol. 5, no. 3, 10.1128/msystems.00656–19, May 2020, Publisher: American Society for Microbiology.

[10] C. Jin, C. Jia, W. Hu, H. Xu, Y. Shen, and M. Yue, "Predicting antimicrobial resistance in *E. coli* with discriminative position fused deep learning classifier," *Computational and Structural Biotechnology Journal*, vol. 23, pp. 559–565, Dec. 2024, ISSN: 2001-0370.

[11] X. Kuang, F. Wang, K. M. Hernandez, Z. Zhang, and R. L. Grossman, "Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN," en, *Scientific Reports*, vol. 12, no. 1, p. 2427, Feb. 2022, Number: 1 Publisher: Nature Publishing Group, ISSN: 2045-2322.

[12] B. P. Alcock et al., "Card 2023: Expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database," *Nucleic Acids Research*, vol. 51, pp. D690–D699, 2023.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," en, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24574-4.

[14] J. A. Noble and D. Boukerroui, "Ultrasound image segmentation: A survey," *IEEE Transactions on medical imaging*, vol. 25, no. 8, pp. 987–1010, 2006.

[15] M. Su, S. W. Satola, and T. D. Read, "Genome-based prediction of bacterial antibiotic resistance," *Journal of clinical microbiology*, vol. 57, no. 3, pp. 10–1128, 2019.

[16] D. Moradigaravand, M. Palm, A. Farewell, V. Mustonen, J. Warringer, and L. Parts, "Prediction of antibiotic resistance in escherichia coli from large-scale pan-genome data," *PLoS computational biology*, vol. 14, no. 12, e1006258, 2018.

[17] C. J. Murray et al., "Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis," *The lancet*, vol. 399, no. 10325, pp. 629–655, 2022.

[18] D. J. Larsson and C.-F. Flach, "Antibiotic resistance in the environment," *Nature Reviews Microbiology*, vol. 20, no. 5, pp. 257–269, 2022.

[19] A. Wong, "Epistasis and the evolution of antimicrobial resistance," *Frontiers in Microbiology*, vol. 8, p. 246, 2017.

[20] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene Parsing through ADE20K Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, ISSN: 1063-6919, Jul. 2017, pp. 5122–5130.

[21] H. Caesar, J. Uijlings, and V. Ferrari, "COCO-Stuff: Thing and Stuff Classes in Context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1209–1218.

[22] M. Cordts et al., "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3213–3223.

[23] B. Menze et al., "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[24] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," en, *Nat Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021, Number: 2 Publisher: Nature Publishing Group, ISSN: 1548-7105.

[25] M. Antonelli et al., "The Medical Segmentation Decathlon," en, *Nat Commun*, vol. 13, no. 1, p. 4128, Jul. 2022, Number: 1 Publisher: Nature Publishing Group, ISSN: 2041-1723.

[26] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[27] L. Maier-Hein et al., "Metrics reloaded: Recommendations for image analysis validation," *Nature methods*, vol. 21, no. 2, pp. 195–212, 2024.

[28] J. Canny, "A computational approach to edge detection," *IEEE Transactions on pattern analysis and machine intelligence*, no. 6, pp. 679–698, 2009.

[29] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[30] D. Rueckert, M. J. Clarkson, D. L. Hill, and D. J. Hawkes, "Non-rigid registration using higher-order mutual information," in *Medical Imaging 2000: Image Processing*, SPIE, vol. 3979, 2000, pp. 438–447.

[31] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *International conference on medical image computing and computer-assisted intervention*, Springer, 1998, pp. 130–137.

[32] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, IEEE, vol. 1, 2001, pp. 105–112.

[33] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[34] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[35] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[36] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 2002.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[38] A. Vaswani et al., "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[39] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[40] M. Tharmakulasingam, B. Gardner, R. La Ragione, and A. Fernando, "Rectified Classifier Chains for Prediction of Antibiotic Resistance From Multi-Labelled Data With Missing Labels," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 1, pp. 625–636, Jan. 2023, Conference Name: IEEE/ACM Transactions on Computational Biology and Bioinformatics, ISSN: 1557-9964.

[41] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[42] R. S. Sutton, A. G. Barto, et al., *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1.

[43] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3, pp. 229–256, 1992.

[44] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[46] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*, Springer, 2016, pp. 424–432.

[47] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, Ieee, 2016, pp. 565–571.

[48] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 10 012–10 022.

[49] C. Ziwen et al., "AutoFocusFormer: Image Segmentation off the Grid," en, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 227–18 236.

[50] A. Hatamizadeh et al., "Unetr: Transformers for 3d medical image segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2022, pp. 574–584.

[51] Y. Tang et al., "Self-supervised pre-training of swin transformers for 3d medical image analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 20 730–20 740.

[52] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[53] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, Springer, 2020, pp. 213–229.

[54] B. Cheng, A. Schwing, and A. Kirillov, "Per-Pixel Classification is Not All You Need for Semantic Segmentation," in *Advances in Neural Information Processing Systems*, vol. 34, Curran Associates, Inc., 2021, pp. 17 864–17 875.

[55] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-Attention Mask Transformer for Universal Image Segmentation," en, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1290–1299.

[56] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, and Y. Yu, *nnFormer: Interleaved Transformer for Volumetric Segmentation*, arXiv:2109.03201 [cs], Feb. 2022.

[57] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[58] S. Leclerc et al., "Deep learning for segmentation using an open large-scale dataset in 2D echocardiography," *IEEE Trans. Med. Imaging*, vol. 38, no. 9, pp. 2198–2210, 2019.

[59] X. Liu et al., "Deep learning-based automated left ventricular ejection fraction assessment using 2-D echocardiography," *Am. J. Physiol. - Heart Circ. Physiol.*, vol. 321, no. 2, H390–H399, 2021.

[60] E. Smistad et al., "Real-time automatic ejection fraction and foreshortening detection using deep learning," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 67, no. 12, pp. 2595–2604, 2020.

[61] J. Zhang et al., "Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy," *Circulation*, vol. 138, no. 16, pp. 1623–1635, 2018.

[62] D. Ouyang et al., "Video-based AI for beat-to-beat assessment of cardiac function," *Nature*, vol. 580, no. 7802, pp. 252–256, 2020.

[63] D. Behnami et al., "Dual-view joint estimation of left ventricular ejection fraction with uncertainty modelling in echocardiograms," in *Med. Image Comput. Comput. Assist. Interv.*, Springer, 2019, pp. 696–704.

[64] A. Ghorbani et al., "Deep learning interpretation of echocardiograms," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–10, 2020.

[65] K. Kusunose et al., "Deep learning for assessment of left ventricular ejection fraction from echocardiographic images," *J. Am. Soc. Echocardiogr.*, 2020.

[66] H. Reynaud, A. Vlontzos, B. Hou, A. Beqiri, P. Leeson, and B. Kainz, "Ultrasound video transformers for cardiac ejection fraction estimation," in *Med. Image Comput. Comput. Assist. Interv.*, Springer, 2021, pp. 495–505.

[67] J. F. Silva, J. M. Silva, A. Guerra, S. Matos, and C. Costa, "Ejection fraction classification in transthoracic echocardiography using a deep learning approach," in *Proc. IEEE Symp. Comput.-Based Med. Syst.*, IEEE, 2018, pp. 123–128.

[68] M. M. K. Esfeh, C. Luong, D. Behnami, T. Tsang, and P. Abolmaesumi, "A deep bayesian video analysis framework: Towards a more robust estimation of ejection fraction," in *Med. Image Comput. Comput. Assist. Interv.*, Springer, 2020, pp. 582–590.

[69] J. M. Blair, M. A. Webber, A. J. Baylay, D. O. Ogbolu, and L. J. Piddock, "Molecular mechanisms of antibiotic resistance," *Nature reviews microbiology*, vol. 13, no. 1, pp. 42–51, 2015.

[70] E. M. Ransom, R. F. Potter, G. Dantas, and C.-A. D. Burnham, "Genomic prediction of antimicrobial resistance: Ready or not, here it comes!" *Clinical chemistry*, vol. 66, no. 10, pp. 1278–1289, 2020.

[71] M. F. Anjum, E. Zankari, and H. Hasman, "Molecular methods for detection of antimicrobial resistance," *Antimicrobial resistance in bacteria from livestock and companion animals*, pp. 33–50, 2018.

[72] H.-L. Her and Y.-W. Wu, "A pan-genome-based machine learning approach for predicting antimicrobial resistance activities of the Escherichia coli strains," *Bioinformatics*, vol. 34, no. 13, pp. i89–i95, Jul. 2018, ISSN: 1367-4803.

[73] J. C. Hyun, E. S. Kavvas, J. M. Monk, and B. O. Palsson, "Machine learning with random subspace ensembles identifies antimicrobial resistance determinants from pan-genomes of three pathogens," en, *PLOS Computational Biology*, vol. 16, no. 3, e1007608, Mar. 2020, Publisher: Public Library of Science, ISSN: 1553-7358.

[74] E. S. Kavvas et al., "Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance," *Nature Communications*, vol. 9, p. 4306, Oct. 2018, ISSN: 2041-1723.

[75] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, 2016.

[76] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, *Evolution strategies as a scalable alternative to reinforcement learning*, 2017.

[77] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in bioinformatics*, vol. 2, 2022.

[78] M. Alirezanejad, R. Enayatifar, H. Motameni, and H. Nematzadeh, "Heuristic filter feature selection methods for medical datasets," *Genomics*, vol. 112, no. 2, pp. 1173–1181, 2020, ISSN: 0888-7543.

[79] H. A. Nam, S. Fleming, and E. Brunskill, "Reinforcement Learning with State Observation Costs in Action-Contingent Noiselessly Observable Markov Decision Processes," in *Advances in Neural Information Processing Systems*, vol. 34, Curran Associates, Inc., 2021, pp. 15 650–15 666.

[80] Y. Li and J. Oliva, "Active Feature Acquisition with Generative Surrogate Models," en, in *Proceedings of the 38th International Conference on Machine Learning*, ISSN: 2640-3498, PMLR, Jul. 2021, pp. 6450–6459.

[81] H. Yin, Y. Li, S. J. Pan, C. Zhang, and S. Tschiatschek, *Reinforcement Learning with Efficient Active Feature Acquisition*, arXiv:2011.00825 [cs], Nov. 2020.

[82] J. Yoon, J. Jordon, and M. v. d. Schaar, "INVASE: Instance-wise Variable Selection using Neural Networks," en, in *International Conference on Learning Representations*, Sep. 2018.

[83] S. Zannone, J. M. H. Lobato, C. Zhang, and K. Palla, "ODIN: Optimal Discovery of High-value INformation Using Model-based Deep Reinforcement Learning," en-US, in *Real-world Sequential Decision Making Workshop, ICML*, Jun. 2019.

[84] Z. Yu, Y. Li, J. Kim, K. Huang, Y. Luo, and M. Wang, *Deep Reinforcement Learning for Cost-Effective Medical Diagnosis*, arXiv:2302.10261 [cs], Feb. 2023.

[85] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.