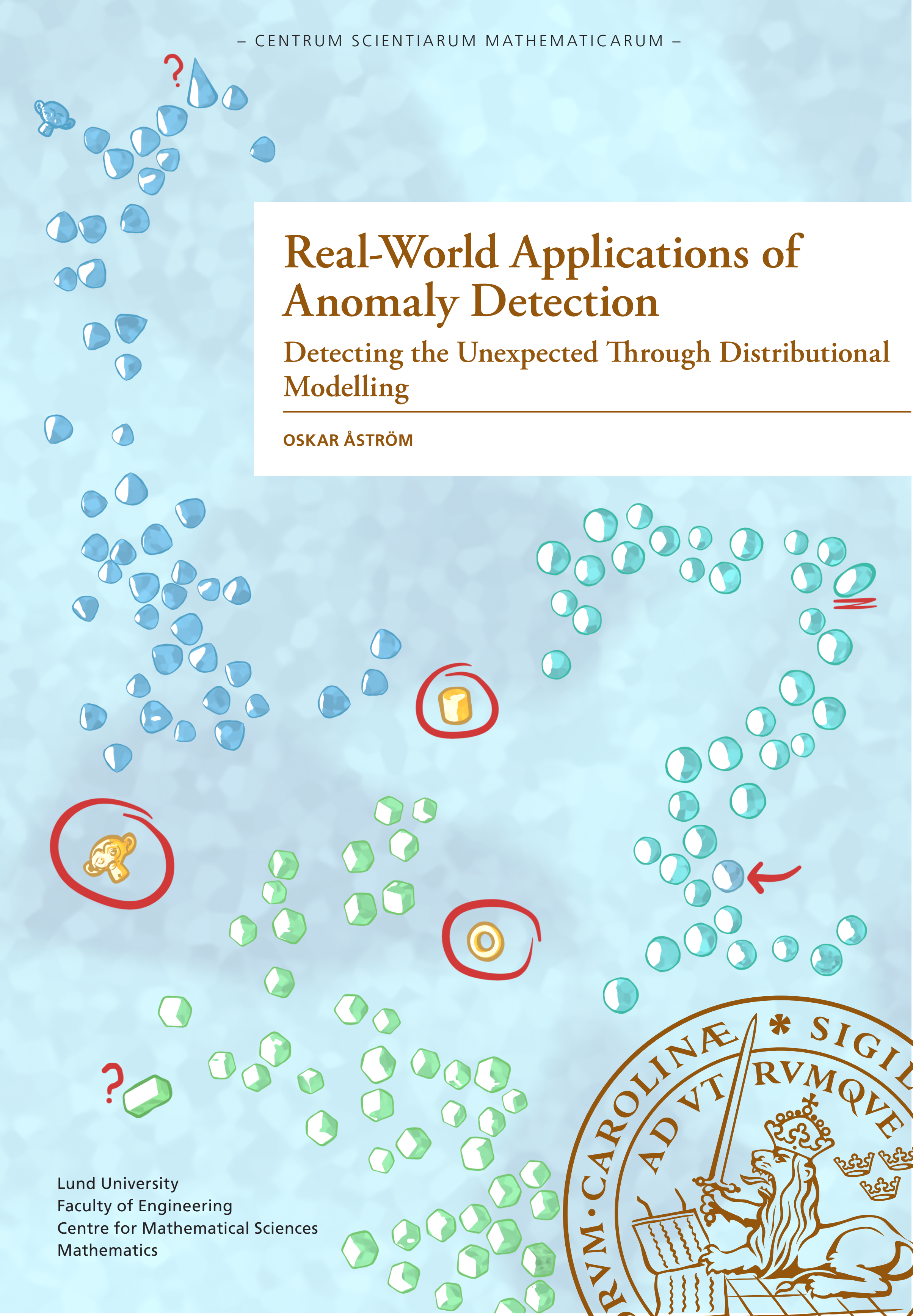
Real-World Applications of Anomaly Detection

Detecting the Unexpected Through Distributional Modelling

Åström, Oskar

2026

– CENTRUM SCIENTIARUM MATHEMATICARUM –



Real-World Applications of Anomaly Detection

Detecting the Unexpected Through Distributional Modelling

OSKAR ÅSTRÖM

Lund University
Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Real-World Applications of Anomaly Detection

Detecting the Unexpected Through Distributional Modelling

by Oskar Åström

LUND
UNIVERSITY

LICENTIATE THESIS

which, with due permission of the Faculty of Engineering at Lund University, will be publicly defended on Tuesday 10th of February, 2026, at 13:15 in lecture hall MH:309A at the Centre for Mathematical Sciences.

Thesis advisors:
Alexandros Sopasakis, Tony Stillfjord, Ola Hall

Faculty opponent:
Dr. Aleksis Pirinen, RISE Research Institutes of Sweden

Organization
LUND UNIVERSITY
Centre for Mathematical Sciences
Box 118
SE–221 00 LUND
Sweden

Document name
Licentiate thesis

Date of presentation
2026-02-10

Sponsoring organization
Rymdstyrelsen, eSSENCE, FORMAS, Swedish Research Council

Author(s)
Oskar Åström

Title and subtitle
Real-World Applications of Anomaly Detection – Detecting the Unexpected Through Distributional Modelling

Abstract

In many machine learning tasks, the premise is designed around predetermined targets and clear expectations of model behaviour. In such cases, there is a direct definition of the optimal mappings between inputs and outputs, which can be learned given sufficiently sized datasets and models. However, in many real-world scenarios, tasks are often not as well-posed and instead defined around detecting the unexpected, the anomalies.

There are many ways of modelling distributions of data points, but in cases of complex high-dimensional data, like images, traditional parametric distributions often fall short. The large non-linear dependencies between pixel values and the cluster-like properties of natural categories make image distributions difficult to model. Instead, recent years have seen advances by using neural networks recontextualized as parametric distributions to construct probabilistic models of natural images.

This thesis investigates how such methods hold up in real-world applications. Modelling data in the wild results in several challenges compared to the controlled conditions of many benchmarks. Instead, by applying these methods in real-world settings, they can be evaluated on their impact and usefulness on downstream tasks. By moving research and method development closer to the intended applications, this thesis aims to highlight some of the benefits that can be gained from bridging the gap between theory and practice.

This thesis contains three main research contributions. The first is a theoretical method development paper that delves into the statistics and machine learning techniques used in the field of anomaly detection. This paper investigates how conditional distributions can be modelled better in variational autoencoder (VAE) models. Commonly, such methods use conditional class clusters which are fully learned by the model. This paper finds that VAE-style models can generalize better with small amounts of rigidity in cluster positions.

The second paper applies these techniques to the field of breast cancer diagnosis. Traditional mammography is a reliable way of diagnosing breast cancer, but is not available globally due to economic constraints. Point-of-care Ultrasound (POCUS) is a promising alternative. However, such images are harder to capture and can contain artifacts that make diagnosis difficult. By modelling the distribution of properly captured POCUS images, we are able to filter out images with artifacts that make them unsuitable for diagnosis.

Paper three applies distributional modelling to the agricultural sector to model how crop yield is distributed over fields using graph neural networks. Using publicly available remote sensing data from the Sentinel-1 and Sentinel-2 satellites, the model is able to estimate how harvest levels were distributed in the past and how the yield will vary in future years. The goal of this study is to provide farmers with more information on how yield is distributed, thereby decreasing cost and mitigating eutrophication caused by over-fertilization.

Key words
Anomaly Detection; Machine Learning, Remote Sensing, Environmental Monitoring, Out-of-Distribution Detection, Medical Image Analysis

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title
1404-028X

ISBN
978-91-8104-838-4 (print)
978-91-8104-839-1 (electronic)

Recipient's notes

Number of pages
xii+103

Price

Security classification

Signature

Date 2026-01-13

Real-World Applications of Anomaly Detection

Detecting the Unexpected Through Distributional Modelling

by Oskar Åström

LUND UNIVERSITY

Copyright

Published by

Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden
www.maths.lu.se

Licentiate Theses in Mathematical Sciences 2026:1
ISSN: 1404-028X
ISBN: 978-91-8104-838-4 (print)
ISBN: 978-91-8104-839-1 (electronic)
LUTFMA-LUTFTM-2002-2026-2026

Printed in Sweden by Media-Tryck, Lund University, Lund 2026

Printed matter
3041 0903

Media-Tryck is a Nordic Swan Ecolabel
certified provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se

MADE IN SWEDEN

Alice laughed.
'There's no use trying,' she said.
'One can't believe impossible things.'

'I daresay you haven't had much practice,'
said the Queen.

– Lewis Carroll (1871)
Through the Looking Glass

Abstract

In many machine learning tasks, the premise is designed around predetermined targets and clear expectations of model behaviour. In such cases, there is a direct definition of the optimal mappings between inputs and outputs, which can be learned given sufficiently sized datasets and models. However, in many real-world scenarios, tasks are often not as well-posed and instead defined around detecting the unexpected, the anomalies.

There are many ways of modelling distributions of data points, but in cases of complex high-dimensional data, like images, traditional parametric distributions often fall short. The large non-linear dependencies between pixel values and the cluster-like properties of natural categories make image distributions difficult to model. Instead, recent years have seen advances by using neural networks recontextualized as parametric distributions to construct probabilistic models of natural images.

This thesis investigates how such methods hold up in real-world applications. Modelling data in the wild results in several challenges compared to the controlled conditions of many benchmarks. Instead, by applying these methods in real-world settings, they can be evaluated on their impact and usefulness on downstream tasks. By moving research and method development closer to the intended applications, this thesis aims to highlight some of the benefits that can be gained from bridging the gap between theory and practice.

This thesis contains three main research contributions. The first is a theoretical method development paper that delves into the statistics and machine learning techniques used in the field of anomaly detection. This paper investigates how conditional distributions can be modelled better in variational autoencoder (VAE) models. Commonly, such methods use conditional class clusters which are fully learned by the model. This paper finds that VAE-style models can generalize better with small amounts of rigidity in cluster positions.

The second paper applies these techniques to the field of breast cancer diagnosis. Traditional mammography is a reliable way of diagnosing breast cancer, but is not available globally due to economic constraints. Point-of-care Ultrasound (POCUS) is a promising alternative. However, such images are harder to capture and can contain artifacts that make diagnosis difficult. By modelling the distribution of properly captured POCUS images, we are able to filter out images with artifacts that make them unsuitable for diagnosis.

Paper three applies distributional modelling to the agricultural sector to model how crop yield is distributed over fields using graph neural networks. Using publicly available remote sensing data from the Sentinel-1 and Sentinel-2 satellites, the model is able to estimate how harvest levels were distributed in the past and how the yield will vary in future years. The goal of this study is to provide farmers with more information on how yield is distributed, thereby decreasing cost and mitigating eutrophication caused by over-fertilization.

List of Publications

This thesis is based on the following publications, referred to by their Roman numerals. They are reproduced and included in this thesis with the permission of their respective publishers. The author's contributions to each paper are listed below.

Main papers

I **Latent Rigidity Regularization for Conditional VAEs in Anomaly Detection**
O. Åström, A. Sopasakis
Working Paper. In Submission.

Author's contributions: The original idea was formulated by me based on the CLVAE architecture formulated in a 2019 paper by Erik Norlander and Alexandros Sopasakis [25]. The implementation of code was done by me, and the paper is predominantly written by me with input from AS.

II **Out-of-Distribution Detection in Point-of-Care Ultrasound Breast Imaging using Variational Autoencoders**
O. Åström, J. Karlsson
Proceedings of the Scandinavian Conference on Image Analysis (SCIA), 2025[3].

Author's contributions: The idea of implementing a conditional VAE framework for anomaly detection in breast cancer was formulated by me and JK, based on previous research by JK. The implementation of code and writing was done by me and JK.

III **Predicting Intra-Field Yield Variations for Winter Wheat using Remote Sensing and Graph Attention Networks**
O. Åström, S. Månsson, I. Lazar, M. Nilsson, J. Ekelöf, A. Oxenstierna, A. Sopasakis
Computers and Electronics in Agriculture 237 (2025): 110499 [4].

Author's contributions: The core idea was formulated as a group. The code was implemented by me, SM, and IL, with SM and IL focusing on data collection and preprocessing, and me focusing on method development and model training. The paper was written jointly by all authors, with me predominantly writing the methodology, results, discussion, and conclusions sections.

Additional publications by the author

VI **Enhancing carbon emission reduction strategies using OCO and ICOS data**
O. Åström, C. Geldhauser, M. Grillitsch, O. Hall, A. Sopasakis
Scientific Reports 15.1 (2025): 36297.

V **Graph-enhanced diffusion models for spatiotemporal imputation**
D. Liu, O. Åström, A. Sopasakis
In submission: *Integrated Computer-Aided Engineering (ICAE)*

Acknowledgements

I would like to express my deepest gratitude to my main supervisor, Alexandros Sopasakis, for your support and ever-optimistic mood. Your guidance and words of encouragement have made the embarkment of my academic journey not only filled with scientific discoveries but also joy. I am also grateful to my co-supervisors, Tony Stillfjord and Ola Hall, for your collaboration and discussions over the past two years. I would also like to thank all my project colleagues at Niftitech, Hushållningssällskapet Skåne, and Nordic Beet Research. Our endless brainstorming sessions and lofty visions are truly the hallmarks of science in progress.

To all my incredible doctoral colleagues, both within and outside the Centre for Mathematical Sciences, I want to thank you all for the wonderful times we've had together so far. The nonsensical fika-break discussions and the emotional support from our knitting sessions are among the highlights of the workplace.

Lastly, I want to thank my whole family, especially my mom and dad, for their unconditional support and encouragement throughout my life. I would also like to thank my grandpa for his endless optimism and inspiration. Finally, to my partner, Hannah: I could not have done this without your love and support. Thank you for the joy and happiness you bring to me. You are the sun in my life.

Funding

My work has been partially supported by grants from eSSENCE no. 138227, FORMAS no. 2022-151862, Rymdstyrelsen no. 2022-00282, and AgTech Sweden. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

Contents

Chapter 1

Introduction

As society increasingly adopts data-driven machine learning systems, reliability, robustness, and safety become ever more critical. From industrial processes and medical imaging to environmental monitoring and ecosystem inventorying, machine learning models are often expected not only to make accurate predictions under normal conditions, but also to recognize when the underlying assumptions no longer hold. In such cases, the ability to detect unexpected or abnormal inputs becomes as important as the primary task itself.

Anomaly detection addresses this challenge by identifying data points that deviate from what is considered normal and expected [33]. In traditional supervised machine learning problems, data used during training is assumed to reflect the conditions faced in the downstream application. This assumption rarely holds true in the real world, as all possible failure modes, artifacts, edge cases, unexpected corruptions, and novel conditions are infeasible to express in advance. The real world is noisy and unpredictable, and as such, accurately detecting anomalies becomes crucial to prevent systems that are unreliable or unstable to distributional shifts [27].

A common strategy for anomaly detection is to construct probabilistic models over a data distribution representing normal behaviour, and to flag observations that are unexpected under the model as anomalous [24]. While conceptually simple, this approach can be exceedingly challenging for high-dimensional complex data, such as images. The problem of anomaly detection is further exacerbated by the fact that the definition of *normal* is subjective and ill-posed. As many real-world tasks are deeply context-dependent, there is often no single true distribution of what is to be ”expected”. This, in conjunction with the fact that natural images are high-dimensional objects that exhibit non-linear dependencies and often belong to multi-modal distributions, makes anomaly detection a truly non-trivial task [18].

Recent advances in deep learning have provided powerful methods for representing complex, high-dimensional data distributions. In particular, representational learning methods such as variational autoencoders (VAEs) enable the transformation of high-dimensional observations into lower-dimensional latent spaces that more clearly express semantically meaningful structures [2, 16, 24]. Still, while much of prior work shows strong performance on benchmark datasets in controlled settings, the true test lies in how well these models perform in the real world. Methods that fail under real-world conditions, while scientifically interesting, offer limited practical value and risk undermining trust in deployed machine learning systems.

The purpose of this thesis is to investigate how distributional modelling is best applied for anomaly detection tasks in real-world settings, and where systems can fail in the wild. The focus is on unsupervised methods, primarily using VAEs, to explore how latent space structures influence robustness, stability, and anomaly detection performance. Through a combination of theoretical and practical work, this thesis aims to bridge the gap between abstract anomaly detection methods and their deployment in the wild.

The thesis consists of three main research contributions. The first investigates how the latent space structure in class-conditional VAEs is affected by cluster rigidity, and whether rigidity can be used to improve anomaly detection and prevent issues such as posterior collapse. The second paper applies these class-conditional VAEs in the real-world application of breast cancer diagnosis. This paper investigates how anomaly detection can be applied to point-of-care ultrasound imaging, where the identification and removal of improperly captured images with artifacts such as blur or acoustic shadows is crucial for ensuring reliable medical diagnoses. The third paper explores how graph neural networks can be used to model and predict the distribution of crop yield using remote sensing data, where accurate knowledge of yield distribution can help reduce overfertilization and environmental impacts in the agricultural sector.

Before the presentation of the three main research contributions, important background information is presented in the following sections. Chapter 2 will highlight core machine learning concepts and common techniques. This is followed by Chapter 3, giving an overview of what anomaly detection is and the challenges faced when modelling real-world data. Chapter 4 focuses on variational autoencoders, the main model architecture used in this thesis, deriving the statistical methods used for modelling distributions of complex high-dimensional data. This is followed by Chapters 5 and 6, each giving an introduction to the two application-based papers. Chapter 5 presents background on the medical application in Paper II and the data used. Finally, Chapter 6 presents the context for the remote sensing, weather, and soil data used in Paper III, along with background on the project's agricultural applications.

Chapter 2

Machine Learning Techniques

This chapter will outline some of the basic machine learning techniques used in this thesis. Starting with traditional neural networks in Section 2.1, the core ideas behind machine learning will be explained. Following this, the most common practices regarding evaluation, dataset usage, and performance metrics will be described. Finally, the more complex methods, such as Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs), will be described. Exposition of the Variational Autoencoder network (VAE) is found separately in Chapter 4 as its function is key to this thesis and benefits from additional focus.

2.1 Neural Networks

Neural networks have seen an upswing in the last few decades and are the bedrock of modern machine learning. While the focus on neural networks has predominantly increased in recent years, the techniques used date back to the 1940s, with the work of Warren McCulloch and Walter Pitts [23]. This was taken into practice with the realization of the perceptron by Frank Rosenblatt in 1958 [22], albeit in a physical and mechanical form. The difference between modern-day neural networks and those developed in the late 20th century is primarily in scale rather than kind, and many of the same methods and techniques remain.

The core idea behind neural networks is built on nodes and weights. Networks consist of layers of nodes, each containing a value, where the value of a node depends on the values of nodes in the previous layer. Figure 2.1 shows an example network, consisting of an input node x, an output node y, and several hidden nodes h_i^l. These are connected using weights w_{ij}^l that are used as a scaling factor when updating the state of a node. The value of a node is defined using the update function

$$h_i^{l+1} = \sigma(b_i^l + \sum_j w_{ji}^l h_j^l), \tag{2.1}$$

where b_i^l is an additional parameter of the model referred to as the bias. The function σ is called an activation function, which will be discussed in Section 2.1.1. The goal is to find parameters w and b such that the final output y matches a desired target when inputting a corresponding x. There is no limit to how many layers can be used, and of what size. Furthermore, the input and output layers often also consist of many nodes.

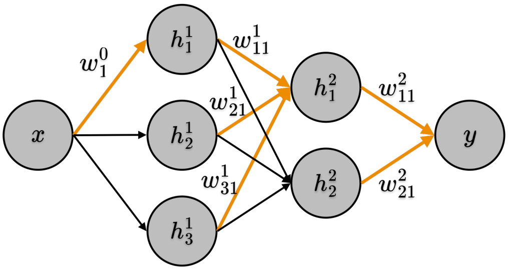


Figure 2.1: An example of a neural network with input x, output y and nodes h_i^l connected using weights w_{ij}^l.

In practice, these computations are performed in matrix form. The general formulation is then expressed as

$$h^{l+1} = \sigma(b^l + W^l h^l), \tag{2.2}$$

where h^l and b^l are the vectorized collection of nodes and biases, respectively, and W^l is a matrix consisting of the weights w_{ij}^l.

2.1.1 Activation Functions

Since the interactions between weights and nodes are linear, the final output would have an affine function of the input if it weren't for the activation function σ. The original idea is to mimic the behaviour of biological neurons that fire or *activate* if they receive a strong enough signal. In practice, the main benefit is that it introduces non-linearity into the model, allowing for the mimicking of a wider and more complex family of functions. There are several choices of activation function, with one of the earliest being the Sigmoid function

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}}. \tag{2.3}$$

This function always outputs a value between 0 and 1, which fits with the biological interpretation that was prevalent in the early days of machine learning. However, in modern times, the most common activation function is the Rectified Linear Unit (ReLU),

$$\text{ReLU}(x) = \begin{cases} 0, & \text{if}\, x < 0 \\ x, & \text{if}\, x \geq 0. \end{cases} \tag{2.4}$$

This is often used as it is very quick to compute and has non-vanishing gradients for positive values. A variant of this, referred to as LeakyReLU, is defined as

$$\text{LeakyReLU}(x) = \begin{cases} \alpha \cdot x, & \text{if}\, x < 0 \\ x, & \text{if}\, x \geq 0, \end{cases} \tag{2.5}$$

where α is a positive number smaller than 1.

2.1.2 Datasets in Machine Learning

The other core component when training a neural network is the dataset. This can take many forms depending on the task. The two types of tasks that will be outlined here are supervised learning and unsupervised learning.

In supervised learning, datasets consist of input-output pairs. The goal of the model is to return the target output given the corresponding input. The input-output pairs can consist of anything, from image-label pairs in classification to sentence-next-word pairs in text generation. In this thesis, supervised learning is used in Paper III, where inputs are represented as a graph structure containing satellite, weather, and soil measurements across a field, and outputs are represented as a similar graph structure containing the crop yield.

In contrast, unsupervised learning does not use target outputs. Instead, the goal is to find efficient representations and patterns in the input data. This is often used as a preprocessing or feature extraction step for some other downstream task. Unsupervised learning can be more difficult to train because the objective of the downstream task is not typically built into the learning process. However, this task ambiguity can be a benefit as biases introduced in the supervision process are avoided. This approach is common in anomaly detection as it is desirable to have a model that is ambiguous to the type of anomaly. In the real world, outcomes can deviate from expectations in many ways, so training this detection model in a supervised manner often biases it towards finding a specific kind of anomaly. This poses a risk that anomalies you haven't considered may go undetected due to their absence in the training data. Unsupervised learning is therefore used in Papers I and II.

When training a model, it is common practice to split the dataset into three subsets: the training set, the validation set, and the test set. The training set is the bulk of the data and is used to train the weights of the neural network. The validation set is then used to check that the model can still perform well on data it hasn't seen before. Due to the large number of parameters in modern neural networks, one has to be sure that the model learns patterns that are true in general instead of just memorizing the inputs in the training set. By checking the performance on the validation set during training, one can ensure that the model's capability will likely transfer to unseen data.

After training, the test set is used as the final metric for how well the model performs on completely unseen data. Ideally, the test set is from a slightly different source to ensure minimal overlap between the training and test sets that could taint the test set integrity.

2.1.3 Learning the Parameters

While neural network training is often referred to as *learning*, the process is actually fairly mechanical in nature. Neural networks are developed using a loss function $\mathcal{L}$. This describes how poorly the model performs some task. Common losses are the Mean Squared Error (MSE) in regression tasks or Binary Cross-Entropy (BCE) in classification tasks

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_i (y_i - \hat{y}_i)^2 \quad \mathcal{L}_{BCE} = -\frac{1}{N}\sum_i y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i), \tag{2.6}$$

where $\hat{y}_i$ are the model outputs and y_i are the corresponding target outputs. Since the loss function is dependent on the model output, it is also a function of the model weights. During training, the loss function is recontextualized such that the weights are considered variables, and the input-output values are considered fixed values. Importantly, all composite functions in the network are differentiable. Hence, the partial derivatives of the loss with respect to each weight in the model can be computed, $\delta\mathcal{L}/\delta w_i$. The objective is to find parameters w_i such that the loss function is minimized. This is done in practice through gradient descent. This method computes the gradient of the loss function with respect to the weights and shifts the weights a small step in the opposite direction, essentially following the downward slope of the loss function. Several modifications of this exist, such as adding momentum terms, introducing amounts of random noise, or adaptively updating the size of the gradient steps throughout the training [29].

2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a subgenre of NNs and are the predominant layer architecture used for image inputs [6]. Layers then consist of image-like structures instead of vectors. The next layer is constructed by convolving the previous layer with a kernel. This means that nodes in the network are only connected with a small subset of the previous layer, vastly reducing the number of connections in the network. Additionally, since the same kernel is used for the entire image, only a very small number of weights have to be used, further reducing the computational cost.

Figure 2.2 shows an example of such a convolution operation in a CNN. The 3x3 kernel maps the 25 pixels in the input image to 9 output pixels. In a fully connected layer, this would require $(25+1)\cdot 9 = 234$ weights, whereas the convolution layer only uses the 10 weights in the kernel and bias.

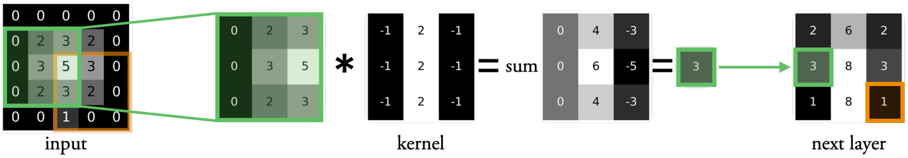


Figure 2.2: Example of convolution operation on a 5x5 input image using a 3x3 kernel and the corresponding 3x3 output image to the next layer.

This type of layer architecture is used in all three papers outlined in this thesis. In Paper I, it is used as the initial layers of the two autoencoders. In Paper II, convolutional layers are used both for the breast cancer classifier network and for the feature extraction in the anomaly detection network. Finally, in Paper III, a CNN is used for yield prediction as a comparison to the graph neural network outlined in the following section.

2.3 Graph Neural Networks

Graph Neural Networks (GNNs) [17] can be considered a generalisation of CNNs. Whereas convolutional layers update pixel values according to a kernel in a neighbourhood around a pixel, GNNs remove the pixel-grid structure and allow for arbitrary definitions of what a neighbourhood is. Instead, it utilizes a graph structure of nodes and edges, where directly connected nodes are considered neighbours.

Each node i in the graph contains a state h_i. The layers of the GNN update the states of all nodes in the graph. The new state of a node after update k is defined through some function f of its current state h_i^{k-1} and the state of all its neighbours $j \in \mathcal{N}_i$.

$$h_i^k = f\left(h_i^{k-1}, \{h_j^{k-1} | j \in \mathcal{N}_i\}\right), \tag{2.7}$$

Specifically, in Paper III, a version of GNNs called Graph Attention Networks (GATs) is used [5]. This network utilizes an attention mechanism that allows the network to emphasize the importance of some of its neighbours over others. The new state of a node is computed through

$$h_i^{k+1} = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \boldsymbol{W}_t h_j^k\right), \tag{2.8}$$

where σ is an activation function, $\boldsymbol{W}_t$ is a matrix of learnable weights and α_{ij} represents the attention between nodes i and j. The attention is defined as the softmax-normalized importance value v_{ij} between nodes,

$$\begin{aligned} \alpha_{ij} &= \underset{j' \in \mathcal{N}_i}{\text{softmax}}(v_{ij}) = \frac{e^{v_{ij}}}{\sum_{j' \in \mathcal{N}_i} e^{v_{ij'}}} \\ v_{ij} &= \boldsymbol{a}^\intercal LeakyReLU(\boldsymbol{W}_s h_i^k + \boldsymbol{W}_t h_j^k), \end{aligned} \tag{2.9}$$

where $\boldsymbol{a}$ and $\boldsymbol{W_s}$ are learnable weights. Importantly, the node i is itself included in its own neighbourhood $\mathcal{N}_i$. For the implementation in Paper III, edges are also attributed features, which results in a slightly modified importance

$$v_{ij} = \boldsymbol{a}^\intercal LeakyReLU(\boldsymbol{W}_s h_i^k + \boldsymbol{W}_t h_j^k + \boldsymbol{W}_e e_{ij}), \tag{2.10}$$

where $\boldsymbol{W}_e$ is a new matrix of learnable weights and e_{ij} are the features of the edge going between node i and j.

Chapter 3

Anomaly Detection

Anomaly detection can be summarized as observing a set of objects, finding some distribution of how normal objects in this set behave, and then determining if newly observed objects are likely to belong to this distribution or not. What is meant by anomaly is therefore ultimately determined by what *normal* means. In real-world applications, this is typically a very subjective task, and one in which something resembling consensus might not exist.

In some cases, one might know exactly what type of deviations one is trying to detect. An example could be looking for green spots on a potato. In these cases, traditional supervised methods can be used, for example, with one class of images containing green potatoes and one containing non-green potatoes. However, in many cases, it is desirable to remain agnostic to the type of anomalies that can occur. Other deviations that can occur in potato harvesting are that the potato could be rotten or split in half. Maybe you're not even looking at a potato, and a carrot has gotten into the system. In the real world, it is often hard to foresee all the ways something unexpected can happen. Constructing a supervised dataset that covers all possible ways in which something can go wrong is therefore often not feasible. Instead, anomaly detection typically relies on unsupervised learning, where the model is trained only on *normal* data and is then tasked with determining whether new observations are also normal or if they represent something new.

Given a probability distribution $p(x)$ describing normal data in some input space X, the set of anomalies $\mathcal{A}$ is simply defined as

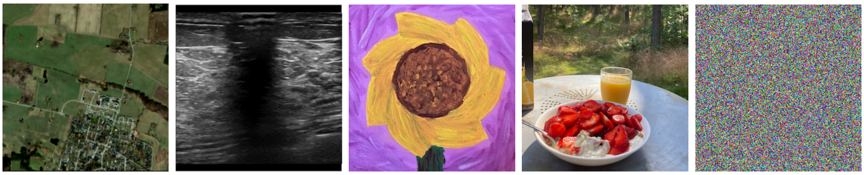
$$\mathcal{A} = \{x \in X \mid p(x) < \tau\}, \tag{3.1}$$

for some threshold τ. The difficulty lies in finding this probability distribution $p(x)$ describing what normal inputs are expected to look like.

3.1 Modelling Image Distributions

There are many ways in which data and information from the real world can be portrayed in formats understandable to a computer. Lists of numbers, embedded text, spectrograms of audio signals, and so on, but none are as visceral as images. As sight is often ranked as our most important sense [8], it comes as no surprise that many machine learning tasks focus on getting computers to understand and interpret images. Convolutional neural networks (CNNs), image diffusion models, depth estimation, and image segmentation are just some of the methods and tasks that have seen incredible advances in recent decades. There are, however, challenges when trying to model and understand images.

The first issue lies in the fact that images are very high-dimensional objects. Even a small colored image of size 256x256 resides in an image space of almost 200,000 dimensions, a fact that reveals the staggering degree of freedom the image space exhibits. This high dimensionality results in a high expressivity, where anything from satellite images of fields and ultrasound images of breast cancer to paintings of sunflowers and pictures of you eating breakfast in 5 years is contained (Figure 3.1). Not to mention the vast amount of images that appear as just random noise.



(a) Satellite image of fields (b) Ultrasound of breast tissue (c) Sunflower painting (d) Breakfast photograph (e) Random noise

Figure 3.1: Samples from the 256x256 image space.

Numerical statistics in high-dimensional spaces is notoriously difficult due to the fact that probabilities rapidly decay to 0 as the number of dimensions increases. The probability density function for the standard multivariate Gaussian can be expressed as

$$p(x) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{||x||^2}{2}}, \tag{3.2}$$

meaning that, at any given distance $||x||$ from the origin, the probability density function decreases exponentially with the dimension n. This means that, ideally, one would like to operate on a distribution in a lower-dimensional space.

The second issue is that one is typically only interested in a very small subset of the full image space. Directly expressing the distribution of this subset is very difficult for a number of reasons. Images that appear very similar to us can be far apart in the image space. Shifting every pixel one step to the right yields an almost identical picture, which is nonetheless very far away in the image space. Furthermore, subsets of the image space constructed from natural categories are typically non-convex. Linear interpolations between images often produce vague, blurry representations that are unlikely to be observed in reality (Figure 3.2). Finally, such distributions are often multi-modal, meaning that the distribution consists of several peaks that are separated by low-probability regions.

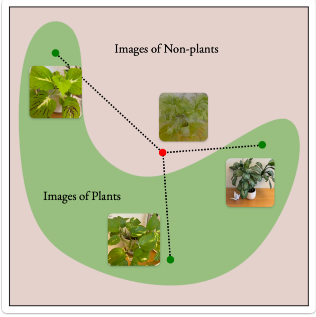


Figure 3.2: Visualization of the non-convex subset of the image space that are images of plants.

The high dimensionality, the non-convex subspace, and the multimodality result in distributions that are hard to model using traditional parametric methods.

3.2 Neural Networks as Parametric Distributions

One thing to remark on is that natural images often contain strong spatial patterns that result in high correlation and non-trivial dependencies between pixels. If one petal of a flower is pink, there is a large likelihood that the other petals will be so as well, but if one petal is yellow, the others are very unlikely to be pink. These high correlations between pixels imply that some distributions might be well-suited for low-rank approximations [15].

As expressed in Paper I, any Gaussian distribution $\mathcal{X} \sim N(\mu, \Sigma)$ of dimension n and with covariance matrix of rank $rank(\Sigma) = d$ can be described as a transformed standard Gaussian distribution $\mathcal{Z} \sim N(0, I_d)$ of dimension d using the transformation

$$\mathcal{X} \overset{d}{=} \mu + A\mathcal{Z}, \tag{3.3}$$

with some $A \in \mathbb{R}^{n\times d}$ and $AA^\intercal = \Sigma$. In theory, this approach solves the first issue of high dimensionality. However, with a simple affine transformation, the issue of non-convexity remains a problem. Tackling this requires more advanced transformations, which can be constructed using neural networks. By recontextualizing a neural network with d input nodes and n output nodes as a non-linear transformation between a d-dimensional vector space and an n-dimensional one, complicated non-convex distributions can be approximated from a single low-dimensional Gaussian. Note that this does not address the issue of multi-modality, something which will be revisited in Section 4.2.

Chapter 4

Variational Autoencoders

Starting with the idea that complex distributions can be approximated as transformations of a standard Gaussian distribution, we introduce the concept of the latent space $Z = \mathbb{R}^d$. This is the generating space in which the variance of the input space $X = \mathbb{R}^n$ arises. In other words, with some transformation $\mathcal{D} : Z \mapsto X$, the distribution of samples in the input space is described as $\mathcal{D}(N(0, I_d))$, where $N(0, I_d)$ is the standard multivariate Gaussian distribution. Notably, $d \ll n$, i.e., the dimension of latent space Z is typically much smaller than the dimension of input space X. The goal is then to approximate the true distribution of inputs $p^*(x)$ as a parameterized distribution $p_\theta(x)$ using the conditional distribution $p_\theta(x|z)$ representing the transformation $\mathcal{D}$.

To learn optimal parameters of $p_\theta(x|z)$ using a machine learning framework, one would need pairs of inputs x and corresponding latent vectors z to train on in a supervised fashion. However, as no latent vectors exist and have not yet been defined, they have to be generated using the posterior distribution $p_\theta(z|x)$. This probability is, however, intractable to compute since calculating $p_\theta(x)$ requires marginalizing z from $p_\theta(x, z)$. Instead, the posterior is approximated with another parameterized transformation $\mathcal{E} : X \mapsto Z$ corresponding to this estimated posterior $q_\phi(z|x) \approx p_\theta(z|x)$ with parameters ϕ.

In a machine learning setting, the two transformations $\mathcal{E}$ and $\mathcal{D}$ are referred to as the encoder and decoder, respectively. They together form the basis of the Variational Autoencoder (VAE). The encoder transforms an input x into a distribution $q_\phi(z|x)$ in the latent space, from which a latent vector z is sampled. The decoder $p_\theta(x|z)$ then generates a reconstructed $\hat{x}$ from this sampled z.

4.1 The Objective of the Variational Autoencoder

A good encoder-decoder pair should generate reconstructions $\hat{x}$ that are similar to the original input x. This objective is embedded in the VAE via a loss function that maximizes the log-likelihood of the dataset inputs.

Consider the log-probability of observing an input x under the parameterized distribution, and then multiplying by 1, through the multiplication and division of $p_\theta(x,z)$ and $q_\phi(z|x)$.

$$\begin{aligned} log[p_\theta(x)] = log\frac{p_\theta(x)p_\theta(x,z)q_\phi(z|x)}{p_\theta(x,z)q_\phi(z|x)} &= log\frac{p_\theta(x)q_\phi(z|x)}{p_\theta(x,z)} + log\frac{p_\theta(x,z)}{q_\phi(z|x)} \\ &= log\frac{q_\phi(z|x)}{p_\theta(z|x)} + log\frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)} \\ &= log\frac{q_\phi(z|x)}{p_\theta(z|x)} + log[p_\theta(x|z)] - log\frac{q_\phi(z|x)}{p_\theta(z)} \end{aligned} \tag{4.1}$$

Note that this equation holds for any latent z. By multiplying both sides by $q_\phi(z|x)$ and integrating with respect to z, the expected value of the right-hand side over $z \sim q_\phi(z|x)$ can be computed, yielding

$$\begin{aligned} log[p_\theta(x)] &= \int q_\phi(z|x)log[p_\theta(x)]dz = \mathbb{E}_{z\sim q_\phi(z|x)}\Big[log[p_\theta(x)]\Big] \\ &= \mathbb{E}_{q_\phi}\left[log\frac{q_\phi(z|x)}{p_\theta(z|x)}\right] + \mathbb{E}_{q_\phi}\Big[log[p_\theta(x|z)]\Big] - \mathbb{E}_{q_\phi}\left[log\frac{q_\phi(z|x)}{p_\theta(z)}\right] \\ &= D_{KL}\Big[q_\phi(z|x)||p_\theta(z|x)\Big] + \mathbb{E}_{q_\phi}\Big[log[p_\theta(x|z)]\Big] - D_{KL}\Big[q_\phi(z|x)||p_\theta(z)\Big], \end{aligned} \tag{4.2}$$

where $D_{KL}(p_1||p_2) = \mathbb{E}_{p_1}[log\frac{p_1}{p_2}]$ is the Kullback–Leibler (KL) divergence, a pseudo-distance between distributions. The KL-divergence is not a proper metric as it is not symmetric ($D_{KL}(p_1||p_2) \neq D_{KL}(p_2||p_1)$) and does not satisfy the triangle inequality [31]. It is, however, non-negative and is 0 only if the two distributions are identical.

As stated before, the conditional probability $p_\theta(z|x)$ in the first KL-divergence is intractable to compute as $\mathcal{D}$ is non-invertible. It is therefore common to subtract this non-negative divergence term from both sides of Equation (4.2). The remaining terms are referred to as the Evidence Lower Bound (ELBO) of $log(p_\theta(x))$.

$$\begin{aligned} log[p_\theta(x)] &\geq log[p_\theta(x)] - D_{KL}\Big[q_\phi(z|x)||p_\theta(z|x)\Big] \\ &= \underbrace{\mathbb{E}_{q_\phi}\Big[log[p_\theta(x|z)]\Big] - D_{KL}\Big[q_\phi(z|x)||p_\theta(z)\Big]}_{ELBO} \end{aligned} \tag{4.3}$$

The intuition behind this is that maximizing the ELBO will simultaneously maximize $log(p_\theta(x))$ and minimize $D_{KL}[q_\phi(z|x)||p_\theta(z|x)]$. This should, ideally, result in a distribution $p_\theta(x)$ that yields a high likelihood of observing the inputs, and an approximated posterior $q_\phi(z|x)$ that is close to the actual posterior $p_\theta(z|x)$. The loss function is then defined as the negative ELBO, consisting of two parts, the reconstruction loss $\mathcal{L}_{rec}$ and the KL-divergence loss $\mathcal{L}_{KL}$,

$$\mathcal{L}(x,\theta,\phi) = \underbrace{-\mathbb{E}_{q_\phi}\Big[log[p_\theta(x|z)]\Big]}_{\mathcal{L}_{rec}} + \underbrace{D_{KL}\Big[q_\phi(z|x)||p_\theta(z)\Big]}_{\mathcal{L}_{KL}}. \tag{4.4}$$

This loss contains three probabilities, $p_\theta(x|z)$ corresponding to the decoder network $\mathcal{D}$, $q_\phi(z|x)$ corresponding to the encoder network $\mathcal{E}$, and the prior $p_\theta(z)$, which the latent distribution is assumed to belong to. In the traditional VAE, $p_\theta(z)$ is assumed to be a standard multivariate Gaussian.

The expected values in the loss are not feasible to compute using integrals over the latent space, and are instead calculated by sampling inputs in the training dataset. In practice, the reconstruction loss is often computed as the Mean Squared Error (MSE) of the reconstruction $\hat{x}$ compared to the input x

$$\mathcal{L}_{rec}(x,\theta,\phi) = ||x-\hat{x}||^2, \tag{4.5}$$

although some models choose to substitute the MSE with other metrics, such as binary cross-entropy. When computing the KL-divergence loss, both distributions are typically assumed to be Gaussian. $p_\theta(z)$ is assumed to be $N(0,I)$, while $q_\phi(z|x)$ is assumed to be $N(\mu(x),\Sigma(x))$, where the mean and covariance is dependent on the input x. This is done by having the encoder output two d-dimensional vectors, one representing the mean and one representing the log-variance of each latent dimension. These two vectors are used to construct the distribution $q_\phi(z|x)$. The KL-divergence then becomes

$$\mathcal{L}_{KL}(x,\theta,\phi) = -\frac{1}{2}\left[d - ||\sigma(x)||^2 - ||\mu(x)||^2 + \sum_{i=1}^{d} log(\sigma_i^2(x))\right]. \tag{4.6}$$

4.2 Conditioning the Encoder

One limitation of the VAE with respect to distributional modeling is multi-modality. That is, transformed Gaussians are still bad at modelling distributions with multiple peaks or *modes*. This is because it is ultimately based on a single Gaussian distribution. Since the decoder is a continuous transformation, the unimodal properties of the latent Gaussian tend to transfer to the output distribution [19].

Papers I and II utilize conditioning of the VAE loss based on class labels. This method modifies the assumption that the prior $p_\theta(z)$ is a standard Gaussian, and instead conditions it based on class information such that $p_\theta(z|c) = N(\mu_c, \Sigma_c)$. This replaces the single global mode with one mode per class, thus allowing for a richer set of distributions to be modelled efficiently. This results in a conditional KL-divergence loss of the form

$$\mathcal{L}_{KL}(x, \theta, \phi) = -\frac{1}{2}\left[d - \left|\left|\frac{\sigma(x)}{\sigma_c}\right|\right|^2 - \left|\left|\frac{\mu(x) - \mu_c}{\sigma_c}\right|\right|^2 + \sum_{i=1}^{d} log\left(\frac{\sigma_i^2(x)}{\sigma_{c,i}^2}\right)\right]. \quad (4.7)$$

4.3 Autoencoders in anomaly detection

The reason for modelling the underlying distributions of data is to aid in detecting novel and anomalous situations in an automatic way using a mathematical description of what normality means. To do this, the VAE model is trained on data assumed to represent the expected and non-anomalous distribution. The learned latent space is then used as a basis for where normative samples should behave. By passing all training data through the model and finding where it tends to cluster in the latent space, one can find anomalies through their presence in low-probability regions of the latent space. Such probability computations are more feasible in the latent space, as it is of lower dimension than the inputs, and because assumptions that the distributions are Gaussian are more plausible compared to distributions in the input space.

Since distributions are modelled in an unsupervised and abstract manner, the idea is that unexpected and unforeseen types of anomalies can be detected with more precision. This reduces the bias toward preconceived notions of what constitutes an anomaly is and bases it more heavily on the dataset itself. With this in mind, it is therefore crucial to have a dataset that is truly representative of all normative cases. Absence of representation in the training set can lead to non-anomalies being incorrectly identified as anomalous. This is especially the case in sensitive and high-risk applications, such as in healthcare and public safety. Awareness of biases caused by dataset distribution and transparency about which domains a model has been verified and tested is crucial to ensure safe and fair deployment.

Figure 4.1 shows a visualization of the conditional VAE. An input image x is encoded into the approximated posterior distribution $q_\phi(z|x)$, which is compared against the conditional prior $p_\theta(z|c)$ using the KL-divergence in the loss $\mathcal{L}_{KL}$. A point z in the latent space is sampled from the posterior $q_\phi(z|x)$, from which the image is reconstructed using the decoder $p_\theta(x|z)$. The input image is compared with the decoded image in the reconstruction loss $\mathcal{L}_{rec}$.

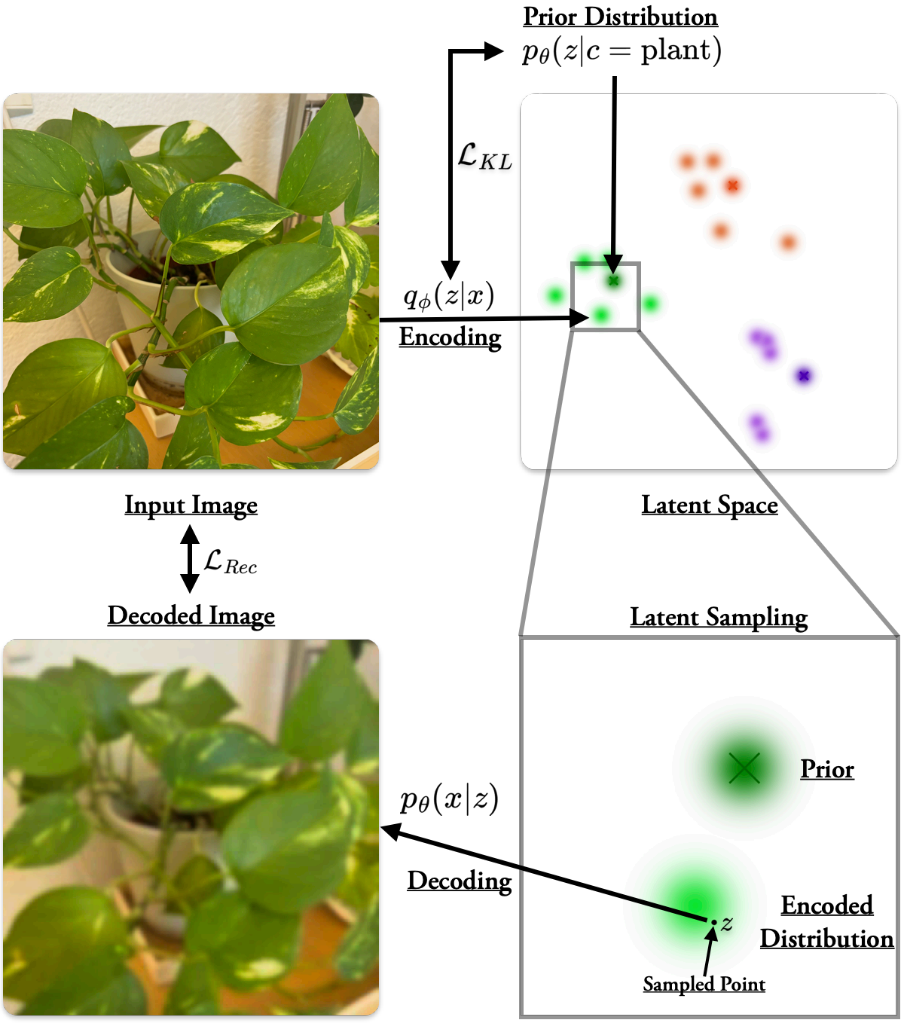


Figure 4.1: Example of a conditional VAE applied to an image of a Golden Pothos plant. The diagram shows the usage of the encoding posterior $q_\phi(z|x)$, conditional prior $p_\theta(z|c)$ and decoder $p_\theta(x|z)$, as well as the two loss terms $\mathcal{L}_{rec}$ and $\mathcal{L}_{KL}$.

Chapter 5

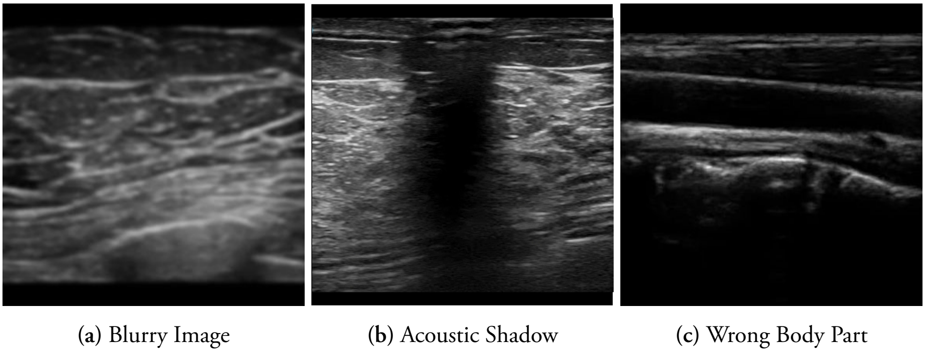
Application: Stability in Breast Cancer Diagnosis

Breast cancer is the most common form of cancer among women, according to a 2022 study by the Global Cancer Observatory (GLOBOCAN) [12]. The relative survival rate of frequently screened populations can be over 100% if the cancer is diagnosed at early, localized stages [20]. The relative survival rate is a measure of how likely a member of a population with the disease is to die compared to members of a population without the disease. A relative survival rate of over 100% can be caused by the ”healthy patient” effect, where members of the screened population are less likely to die from other causes due to increased healthcare visits, supervision, and otherwise higher focus on a healthy lifestyle. If the absolute survival rate is close to 100%, this effect can then push the relative survival rate over 100% [20].

However, while breast cancer has a high survival rate if diagnosed early, more than 660,000 deaths occurred as a result of breast cancer in 2022, out of the close to 2,300,000 reported diagnoses [12]. This represents an almost 30% mortality rate for a disease in which almost all deaths are preventable. This discrepancy is partly caused by the limited availability of screening programs, especially in regions with a lack of healthcare infrastructure [21]. While mammography machines are a reliable way to diagnose cancer, they are expensive and require large-scale healthcare infrastructure to support. This means that there are many regions where mammography is not feasible. An alternative method that has seen some increase in later years is Point-of-care Ultrasound (POCUS) [7]. This method uses small hand-held ultrasound devices to capture images of breast tissue. The portability and low cost of this method allow for screening at the place of the patient, making it well-suited for regions where regular travel to the nearest hospital isn’t an option for parts of the population.

5.1 Point-of-Care Ultrasound Imaging

One major issue with captured POCUS images is that they are often harder to interpret and typically contain higher amounts of noise and image artifacts. Ensuring that images are correctly captured is therefore crucial to ensure accurate diagnoses. In Paper II, distributional modelling for anomaly detection is used to determine whether an ultrasound image was properly acquired. Such deviations can be caused by scanning the wrong body part, image blur due to improper pressure during acquisition, motion blur from movement during image capture, or acoustic shadows caused by obstructing objects or insufficient application of ultrasound gel. Examples of these artifacts taken from Paper II can be seen in Figure 5.1. By modelling a distribution of images without artifacts, Paper II aims to identify and filter out images with deviations that make them unsuitable for diagnosis.



(a) Blurry Image (b) Acoustic Shadow (c) Wrong Body Part

Figure 5.1: Examples of anomalies when capturing POCUS imaging (a) a POCUS image with applied blur; (b) a POCUS image with a contact artifact in the form of an acoustic shadow; (c) an ultrasound image of non-breast tissue (common carotid artery).

Chapter 6

Application:
Variational Patterns in Crop Yield

In agriculture, numerous resources are used to grow food. Water, fertilizer, fuel, machinery, buildings, maintenance cost, human labour, animal labour, land-use, and so on [28]. All of these contribute to the environmental and economic cost of food production. In 2017, agriculture accounted for approximately 20% of total global CO_2-equivalent emissions [11]. As food is a necessary commodity for the function of society, reducing the harmful impacts of food production is an important step toward ensuring long-term societal sustainability.

One contributor to the environmental impact of agriculture is the overuse of fertilizers [13]. This causes increased greenhouse gas emissions through the release of N_2O [32]. It further impacts local ecosystems and water quality through increased eutrophication from fertilizer runoff [1]. Finally, fertilizer itself has a substantial environmental impact through the use of fossil methane in ammonia production, and through the pollution and habitat destruction caused by the mining of mineral fertilizers like phosphorous [26]. Reducing unnecessary use of fertilizers is therefore a key part of lowering the environmental impact of the agricultural sector. In addition, reducing superfluous fertilizer use lowers economic costs for farmers, who typically face very small profit margins [30].

Fertilizer requirements can vary greatly over a field, depending on growth conditions influenced by factors such as slope, soil type, and weather. Treating the growth potential as a uniform distribution over the field leads to fertilizer surplus in low-capacity regions, increasing costs and environmental impact. It also leads to fertilizer deficiency in high-capacity regions, reducing yield. To aid farmers, accurate models of within-field yield variation are needed. Paper III focuses on predicting how yield has been and will be distributed across fields based on historical remote sensing, weather, soil, and harvest data.

6.1 Sentinel 1 and 2

This application predominantly uses remote sensing data from the Sentinel 1 and Sentinel 2 missions by the European Space Agency (ESA). With relatively short revisit times in Sweden of approximately 2 and 5 days for Sentinel 1 and 2, respectively [9, 10], these satellites are useful for real-time applications such as plant monitoring. Additionally, their resolution of 10-60 meters results in high-fidelity maps that can be used to detect spatial variations.

The Sentinel 2 satellite measures the intensity of 13 wavelengths of reflected sunlight from the surface. These wavelength bands span from 443 to 2190 nm, encompassing both visible and short-wave infrared wavelengths [10]. This is a very information-rich data source as it contains the traditional RGB wavelength, allowing for conventional large-scale image methods, but also several wavelengths that are well-suited for detecting specific factors like vegetation, soil moisture, and clouds. These bands can be combined into indices like the Normalized Difference Vegetation Index (NDVI), NDVI $= (B_8 - B_4)/(B_8 + B_4)$, where B_4 is the red band and B_8 is a near infrared band [14]. This is widely used to detect vegetation and moisture. Figure 6.1 shows examples of the NDVI index along with the RGB bands.

The Sentinel 1 satellite uses a C-band Synthetic Aperture Radar (C-SAR) to collect measurements. The C-SAR instrument sends out radar signals toward the planet and measures the reflected signal [9]. The strength of the returned signal is affected by a number of factors, such as surface roughness and moisture. Over land, the C-SAR sends out vertically polarized electromagnetic waves and measures the energy in both the reflected horizontal and vertical components. These two components can be used to construct the SAR Vegetation Index (SARVI) SARVI $= (4\sigma_{VH})/(\sigma_{VV} + \sigma_{VH})$, where σ_{VV} and σ_{VH}, respectively, are the vertical and horizontal components of the reflected vertically polarized light. This index is commonly used in crop growth monitoring. In addition, the SAR signal can penetrate cloud cover but is generally more sensitive to speckle noise, as can be seen in Figure 6.1.

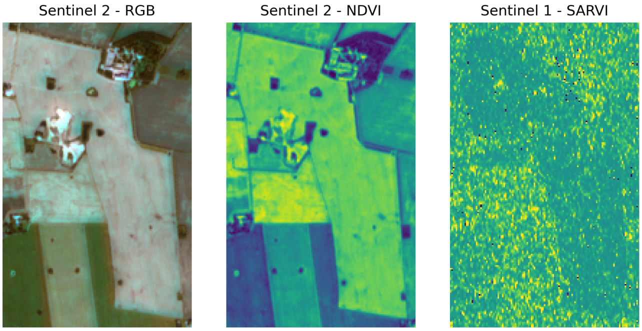


Figure 6.1: Example observations from Sentinel 1 and Sentinel 2 visualized through (left) RGB composite image; (center) NDVI index; and (right) SARVI index.

References

[1] S. O. Akinnawo. Eutrophication: Causes, consequences, physical, chemical and biological techniques for mitigation strategies. *Environmental Challenges*, 12:100733, 2023.

[2] J. An and S. Cho. Variational autoencoder based anomaly detection using reconstruction probability. 2015. URL `https://api.semanticscholar.org/CorpusID:36663713`.

[3] O. Åström and J. Karlsson. Out-of-distribution detection in point-of-care ultrasound breast imaging using variational autoencoders. In *Scandinavian Conference on Image Analysis*, pages 118–130. Springer, 2025.

[4] O. Åström, S. Månsson, I. Lazar, M. Nilsson, J. Ekelöf, A. Oxenstierna, and A. Sopasakis. Predicting intra-field yield variations for winter wheat using remote sensing and graph attention networks. *Computers and Electronics in Agriculture*, 237:110499, 2025.

[5] S. Brody, U. Alon, and E. Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

[6] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao. Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22):4712, 2021.

[7] Q. Dan, T. Zheng, L. Liu, D. Sun, and Y. Chen. Ultrasound for breast cancer screening in resource-limited settings: current practice and future directions. *Cancers*, 15 (7):2112, 2023.

[8] J. Enoch, L. McDonald, L. Jones, P. R. Jones, and D. P. Crabb. Evaluating whether sight is the most valued sense. *JAMA ophthalmology*, 137(11):1317–1320, 2019.

[9] European Space Agency (ESA). S1 mission – overview of sentinel-1 mission. `https://sentiwiki.copernicus.eu/web/s1-mission`, . Accessed: 2025-12-03.

[10] European Space Agency (ESA). S2 mission – overview of sentinel-2 mission. `https://sentiwiki.copernicus.eu/web/s2-mission`, . Accessed: 2025-12-03.

[11] Food and Agriculture Organization of the United Nations. The share of agriculture in total greenhouse gas emissions: Global, regional and country trends 1990–2017, 2020. URL `https://openknowledge.fao.org/handle/20.500.14283/ca8389en`. CA8389EN/1/06.20.

[12] I. A. for Research on Cancer and W. H. Organization. Globocan 2022: World fact sheet. Technical report, International Agency for Research on Cancer, 2024. URL `https://gco.iarc.who.int/media/globocan/factsheets/populations/900-world-fact-sheet.pdf`. Version 1.1.

[13] A. G. Good and P. H. Beatty. Fertilizing nature: a tragedy of excess in the commons. *PLoS biology*, 9(8):e1001124, 2011.

[14] J. Jena, S. R. Misra, and K. P. Tripathi. Normalized difference vegetation index (ndvi) and its role in agriculture. *Agriculture and Food: E-Newsletter*, 1(12):387–389, 2019.

[15] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[16] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[17] T. Kipf. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[18] R. Krusinga, S. Shah, M. Zwicker, T. Goldstein, and D. Jacobs. Understanding the (un) interpretability of natural image distributions using generative models. *arXiv preprint arXiv:1901.01499*, 2019.

[19] F. Lavda, M. Gregorová, and A. Kalousis. Improving vae generations of multimodal data through data-dependent conditional priors. pages 1254–1261, 2020.

[20] A. R. Marcadis, L. G. Morris, and J. L. Marti. Relative survival with early-stage breast cancer in screened and unscreened populations. In *Mayo Clinic proceedings*, volume 97, pages 2316–2323. Elsevier, 2022.

[21] M. E. Martinez, K. M. Schmeler, M. Lajous, and L. A. Newman. Cancer screening in low-and middle-income countries. *American Society of Clinical Oncology Educational Book*, 44(3):e431272, 2024.

[22] P. McCorduck and C. Cfe. *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. AK Peters/CRC Press, 2004.

[23] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[24] A. A. Neloy and M. Turgeon. A comprehensive study of auto-encoders for anomaly detection: Efficiency and trade-offs. *Machine Learning with Applications*, 17:100572, 2024.

[25] E. Norlander and A. Sopasakis. Latent space conditioning for improved classification and anomaly detection. *arXiv preprint arXiv:1911.10599*, 2019.

[26] J. Penuelas, F. Coello, and J. Sardans. A better use of fertilizers is needed for global food security and environmental sustainability. *Agriculture & Food Security*, 12(1): 1–9, 2023.

[27] J. Quiñonero-Candela, M. Sugiama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2008.

[28] H. Ritchie, P. Rosado, and M. Roser. Environmental impacts of food production. *Our World in Data*, 2022. https://ourworldindata.org/environmental-impacts-of-food.

[29] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[30] S. Schaub and N. E. Benni. How do price (risk) changes influence farmers' preferences to reduce fertilizer application? *Agricultural Economics*, 55(2):365–383, 2024.

[31] E. ScienceDirect Topics. Kullback–leibler divergence. URL `https://www.sciencedirect.com/topics/computer-science/leibler-divergence`. Accessed: 2025-12-19.

[32] S. Singh and A. Verma. Environmental review: The potential of nitrification inhibitors to manage the pollution effect of nitrogen fertilizers in agricultural and other soils: A review. *Environmental Practice*, 9(4):266–279, 2007.

[33] J. Yang, K. Zhou, Y. Li, and Z. Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

Scientific Publications

Appendix: Conference posters

Poster 1: Improved Anomaly Detection through Conditional Latent Space VAE Ensembles

Presented at the *Northern Lights Deep Learning Conference* (NLDL 2025) in Tromsø, Norway, January 2025.

Poster 2: Out-of-Distribution Detection in Point-of-Care Ultrasound Breast Imaging using Variational Autoencoders

Presented at the *Scandinavian Conference on Image Analysis* (SCIA 2025) in Reykjavík, Iceland, June 2025.

Improved Anomaly Detection through Conditional Latent Space VAE Ensembles

OSKAR ÅSTRÖM AND ALEXANDROS SOPASAKIS
LUND UNIVERSITY

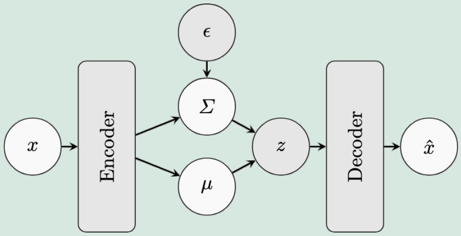
Introduction

Goal: Improve semantic anomaly detection on data with multiple inlier classes.

CL-VAE: Conditional Latent space Variational Autoencoder

- *Multiple Latent Gaussian Distributions*
- *Enforced Radial Separation*
- *Latent Space Ensembling*

Variational Autoencoder



Multiple Latent Gaussian Distributions

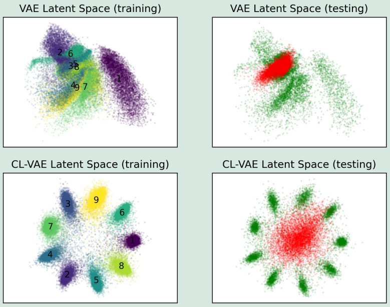
Traditional VAE Loss:

$$\begin{aligned} L(\theta,\phi;x) &= L_{rec} + L_{KL} \\ L_{rec} &= -\mathbb{E}\Big[\log p_\theta(x|z)\Big] \\ L_{KL} &= D_{KL}(q_\phi(z|x)||p_\theta(z)) \end{aligned}$$

By introducing separate gaussians for each inlier class, the diversity in the latent space can be expanded. Each class k is assigned a target mean μ_k.

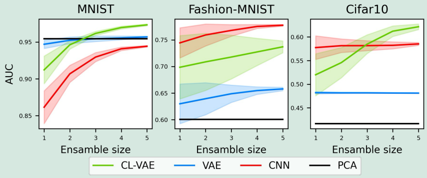
$$\begin{aligned} L_{KL} &= D_{KL}(q_\phi(z|x,k)||p_\theta(z|k)) \\ &= -\frac{1}{2}\left[d - ||\sigma_x||^2 - ||\mu_x - \mu_k||^2 + \sum_{i=1}^{d} \log \sigma_{x,i}^2\right] \end{aligned}$$

Enforced Radial Separation



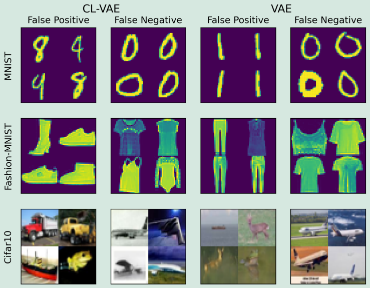
Placing each cluster target μ_k on the surface of a hypersphere forces separation and leaves room in the latent space for anomalies.

Latent Space Ensembling



By randomizing cluster center order, variance is induced between ensemble member, thereby increasing ensembling benefits.

Misclassification Diversity



Conclusions

Allowing separate gaussians for each inlier class induces diversity in the latent space.

Forced separation in the latent space between inlier clusters, space is freed up for semantic anomalies.

Random cluster order creates unique solution spaces that benefit ensembles.

This method of anomaly detection seems to possess improved out-of-distribution detection on inlier data.

Future Work

- Flexible cluster locations
- Learned covariance
- Applications in the wild

Contact

Oskar Åström, Lund University, oskar.astrom@math.lth.se

Implementation at
https://github.com/oskarastrom/CL-VAE

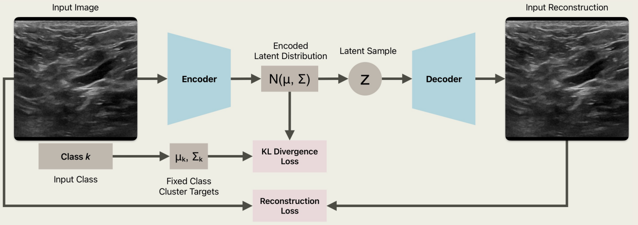
Out-of-Distribution Detection in Point-of-Care Ultrasound Breast Imaging using Variational Autoencoders

JENNIE KARLSSON & OSKAR ÅSTRÖM
DIVISION OF COMPUTER VISION AND MACHINE LEARNING , CENTRE FOR MATHEMATICAL SCIENCES, LUND UNIVERSITY
{JENNIE.KARLSSON, OSKAR.ASTROM}@MATH.LTH.SE

Summary

Recent studies have shown the potential of combining point-of-care ultrasound (POCUS) with a CNN as a support tool in breast cancer diagnostics. However, to ensure trustworthy predictions, it is crucial to detect out-of-distribution (OOD) data. We propose the use of a conditional latent space variational autoencoder (CLVAE) for OOD detection.

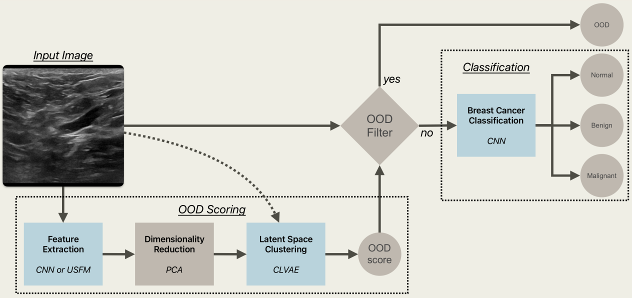
Conditional Latent Space Variational Autoencoder



CLVAE is trained solely on in-distribution (ID) data, i.e. breast ultrasound images labeled into the classes normal, benign, or malignant. Each input image is encoded into a latent distribution. The loss conditions the encoded distributions based on the class. After training, the encoded means of the ID data are used to form three class clusters to which three Gaussians are fitted. The likelihood of a sample belonging to one of these clusters is used as **OOD score**.

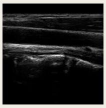
Three different **input pipelines** to the CLVAE were implemented.

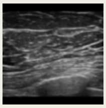
1. **CLVAE** Image directly passed to the CLVAE.
2. **CNN + CLVAE** Image pre-processed by extracting features with the CNN prior to the CLVAE.
3. **USFM + CLVAE** Image preprocessed by extracting features with the ultrasound foundation model (USFM) prior to the CLVAE.

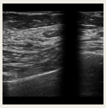


Out-of-Distribution Data

MNIST

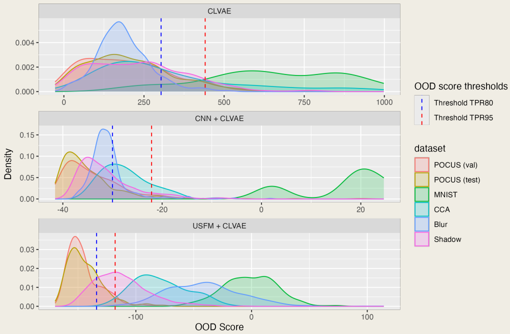

CCA


Blur


Shadow

The CLVAE pipelines were evaluated on four OOD datasets.
MNIST: Images of hand-written digits, i.e. non-ultrasound images.
CCA: Ultrasound images of the common carotid artery, i.e. non-breast tissue.
Blur: Breast ultrasound images with applied blur.
Shadow: Breast ultrasound images with simulated acoustic shadow.

OOD Score Distributions



Distribution of OOD scores for each dataset. The USFM preprocessing shows high separability between the ID dataset (breast POCUS) and OOD datasets.

AUC Results

Pipeline	MNIST	CCA	Blur	Shadow
CLVAE	99.9%	88.1%	80.0%	73.4%
CNN + CLVAE	96.8%	92.0%	93.8%	71.3%
USFM + CLVAE	**100.0%**	**98.7%**	**99.8%**	**88.2%**

Conclusion

- Using the USFM feature extractor improves the OOD detection capabilities of the CLVAE model.
- The Shadow dataset is considerably more difficult to detect as OOD compared to the other datasets.
- This work shows promising results on simulated distortions. Future work should test feasibility on real-world datasets.

Printed by Media-Tryck, Lund 2026 NORDIC SWAN ECOLABEL 3041 0903

LUND
UNIVERSITY

Licentiate Theses in Mathematical Sciences 2026:1
ISBN 978-91-8104-838-4
ISSN 1404-028X