

Representation Learning for Anomaly Detection

Inaugural dissertation

for the attainment of the title of doctor
in the faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

presented by

Rahil Gholamipourfard

from *Iran*

Düsseldorf, June 2022

from the institute of Mathematical Modeling of Biological Systems
at the Heinrich Heine University Düsseldorf

Published by permission of the
Faculty of Mathematics and Natural Sciences at
Heinrich Heine University Düsseldorf

Correspondents:

1. Prof. Dr. Markus Kollmann
2. Prof. Dr. Stefan Conrad

Date of the oral examination: 29.06.2022

*In the memory of my father,
to my mother,
and to those who have inspired me.*

Statement of authorship

I declare under oath that I have produced my thesis independently and without any undue assistance by third parties under consideration of the Principles for the Safeguarding of Good Scientific Practice at Heinrich Heine University Düsseldorf.

Düsseldorf, June 2022
Rahil Gholamipoorfard

Acknowledgements

I would like to express my sincere gratitude to my thesis advisor Prof. Dr. Markus Kollmann for his invaluable guidance during my Ph.D. career. His motivation, enthusiasm and optimism, expertise, and immense knowledge enabled me to grow not only as a researcher but especially as a person. Without his patience, guidance, and persistent help this thesis would not have been possible.

I wish to present my sincere thanks to Prof. Dr. Stefan Conrad for accepting to read the thesis. I am sure that his remarks and suggestions will be so precious to this work.

Some special words of gratitude go to my friends for their unwavering support and love.

Last but not least, I wish to express my heartfelt gratitude to my family members who have always been encouraging me to pursue my passion, trusting in me, and showing me the value of knowledge and diligence. I dedicate this thesis to them.

Düsseldorf, June 2022
Rahil Gholamipoorfard

Abstract

Deep Learning has drawn extensive research interests due to its wide application ranging from autonomous driving to medical diagnosis. One of the most important successes in artificial intelligence is learning generalizable representations from a decent amount of labeled data in an end-to-end fashion. Nevertheless, it is not always feasible to collect a large number of annotated data. Therefore, self-supervised learning has recently drawn increasing attention due to its tremendous performance in various domains, e.g., audio and visual domains. Self-supervised learning is a form of unsupervised learning which learns high-level representations from raw observations without any human supervision which can be broadly used in downstream tasks such as anomaly detection.

Out-of-distribution (OOD) or anomaly detection, i.e., the problem of deciding whether a given test sample is drawn from the same distribution as the training set, is crucial for a reliable learning. Anomaly detection aims at identifying patterns in data that are significantly different to what is expected. Many real-world applications require highly accurate anomaly detection for unassailable deployment, such as in medical diagnosis. There have been many attempts at learning a representation befitting anomaly detection. Inspired by the recent success of self-supervised learning, we aim to make use of its power of representation learning for OOD detection in natural images as well as medical datasets. OOD detection is an important step to improving safety.

In this work, we propose a framework for anomaly detection that does not require any label information. Our framework can be widely applied to OOD detection tasks, including visual and time series data. A main contribution of this work is that our proposed method outperforms supervised and unsupervised methods on challenging OOD detection tasks in the visual domain.

We hope that the provided insights in this work shed light on the challenging problem of anomaly detection and allow for improving decision-making especially in health domain.

Zusammenfassung

Deep Learning hat aufgrund seiner breiten Anwendung, die vom autonomen Fahren bis zur medizinischen Diagnose reicht, großes Forschungsinteresse auf sich gezogen. Das Erlernen verallgemeinerbarer Repräsentationen aus einer angemessenen Menge markierter Daten ist einer der wichtigsten Erfolge in der künstlichen Intelligenz. Dennoch ist es nicht immer möglich, eine große Anzahl kommentierter Daten zu sammeln. Daher hat das selbstüberwachte Lernen in letzter Zeit aufgrund seiner enormen Leistungen in verschiedenen Bereichen, z.B. im Audio und im visuellen Bereich, immer mehr Aufmerksamkeit auf sich gezogen. Selbstüberwachtes Lernen ist eine Form des unüberwachten Lernens, bei dem aus Rohbeobachtungen ohne menschliche Aufsicht Repräsentationen auf hoher Ebene gelernt werden, die in nachgelagerten Aufgaben wie der Erkennung von Anomalien breit eingesetzt werden können.

Out-of-distribution (OOD) oder Anomalie-Erkennung, d.h., das Problem der Entscheidung, ob eine gegebene Testprobe aus der gleichen Verteilung wie die Trainingsmenge stammt, ist entscheidend für ein zuverlässiges Lernen. Die Anomalieerkennung zielt darauf ab, Muster in Daten zu identifizieren, die sich signifikant von dem unterscheiden, was erwartet wird. Viele reale Anwendungen erfordern eine hochpräzise Anomalieerkennung für einen unanfechtbaren Einsatz, wie z.B. in der medizinischen Diagnose. Es gibt viele Versuche, eine geeignete Repräsentation für die Anomalieerkennung zu erlernen. Inspiriert durch den jüngsten Erfolg des selbstüberwachten Lernens wollen wir die Leistungsfähigkeit des Repräsentationslernens für die OOD-Erkennung in natürlichen Bildern sowie in medizinischen Datensätzen nutzen. Die Erkennung von OODs ist ein wichtiger Schritt zur Verbesserung der Sicherheit. In dieser Arbeit schlagen wir einen Rahmen für die Erkennung von Anomalien vor, der keine Etikettinformationen benötigt. Unser Framework kann in großem Umfang auf OOD-Erkennungsaufgaben angewendet werden, einschließlich visueller und Zeitreihendaten. Ein Hauptbeitrag dieser Arbeit ist, dass die von uns vorgeschlagene Methode besser ist als überwachte und unüberwachte Methoden bei anspruchsvollen OOD-Erkennungsaufgaben im visuellen Bereich.

Wir hoffen, dass die in dieser Arbeit gewonnenen Erkenntnisse Licht auf das schwierige Problem der Anomalieerkennung werfen und eine bessere Entscheidungsfindung insbesondere im Gesundheitsbereich ermöglichen.

Contents

1	Introduction	1
1.1	Learning Strategies	2
1.2	Anomaly Detection	3
1.2.1	Anomaly Detection in Medical Health Screening	4
1.3	Thesis Outline	5
2	Basic Information Measures	6
2.1	Entropy	6
2.1.1	Properties of Entropy	7
2.2	Relative Entropy	8
2.2.1	Properties of Relative Entropy	8
2.2.2	Variational Inference	9
2.3	Mutual Information	10
2.3.1	Properties of Mutual Information	11
2.3.2	Data Processing Inequality	13
2.3.3	Information Bottleneck Principle	13
2.3.4	Sufficient Statistics	14
3	Machine Learning Basics	15
3.1	Neural Networks	15
3.1.1	Perceptron	15
3.1.2	Multilayer Perceptron	16
3.2	Machine Learning Algorithms	17
3.3	Generalization	18
3.4	Convolutional NNs vs. Transformers	18
4	Self-Supervised Representation Learning	20
4.1	How to define pretext tasks?	20
4.2	Contrastive Learning	23
4.2.1	Key Properties of Contrastive Loss	25
4.2.2	Different Contrastive Loss Mechanisms	27
5	Self-Supervised Representation Learning in Medical Time Series	28

5.1	Method	28
5.2	Data Preparation	31
5.3	Training Details	31
5.4	Evaluation	31
5.5	Effect of Number of Predicted Latent Steps	32
5.6	Self-Supervised models are more data-efficient	33
5.7	Reliable Detection of Atrial Fibrillation with a Medical Wearable during Inpatient Conditions	34
6	Anomaly Detection in Chest X-ray Images	50
6.1	Pneumonia Detection with Semantic Similarity Scores	51
7	Serious Clinical Complication Detection using Contrastive Learning	57
7.1	Method	58
7.1.1	Learning Statistical Relevant Features	58
7.1.2	Self-Supervised Contrastive Learning	58
7.1.3	Score Function for SCC Detection	59
7.2	Training Details	59
7.3	Detection and Prediction of Serious Clinical Complications with Wearable Based Remote Monitoring and Self-Supervised Contrastive Learning during Intensive Treatment for Hematologic Malignancies	60
8	Anomaly Detection by Negative Sampling	92
8.1	Self-Supervised Anomaly Detection by Self-Distillation and Negative Sampling	93
8.2	Abnormality Detection for Medical Images using Self-Supervision and Negative Samples	106
9	Conclusion and Future Work	117
10	Publications	120
A	Appendix of Chapter 5	122
A.1	Alternative Derivation of I_{NCE}	122

List of Figures

1.1	An illustration to distinguish the supervised, unsupervised and self-supervised learning frameworks. Taken from [1].	3
2.1	Illustrating forward vs reverse KL divergence on a bimodal distribution Taken from [2]. The blue curves are the contours of the true distribution p . The red curves are the contours of the unimodal approximate distribution q . (a) Minimizing forward KL divergence (b-c) Minimizing reverse KL divergence	10
3.1	Illustration of a perceptron. The activation function is denoted by f . $[x_1^n, \dots, x_d^n]^T$ represents feature vector \mathbf{x}_n . w_0 is called bias and represents an external input.	16
3.2	Multilayer perceptron with L hidden layers. Input and output layers have d and c nodes, respectively. The l^{th} hidden layer contains $n^{(l)}$ hidden nodes.	16
4.1	A simple framework for visual representation learning. Each image is augmented twice. Two separate data augmentation operators sampled from the same family of augmentations applied to the image to obtain two correlated views. Taken from [3].	21
4.2	Illustration of the self-supervised task by applying geometric transformations to the input image. The model learns to predict which rotation was applied. Taken from [4].	22
4.3	Illustration of the self-supervised task by predicting relative position between randomly sampling a patch (blue) and one of eight possible neighbors (red). Taken from [5].	22
4.4	Transformations applied to one of the patches extracted from the STL dataset [6]. The original patch is in the top left-hand corner. Taken from [7].	23
4.5	Illustration of self-supervised learning by solving jigsaw puzzle. The tiles marked with green lines are extracted from the image and a puzzle obtained by shuffling the tiles. Taken from [8].	23
4.6	Input grayscale photos and output colorizations. Taken from [9]. . .	23

4.7	(a) Similar examples are mapped to nearby latent features. (b) Uniformity of features distribution on a unit hypersphere. Figures inspired by [10]. CIFAR10 [11] images are used for this demonstration.	24
5.1	Average accuracy of predicting positive samples in the contrastive loss	32
5.2	Evaluating 1-NN classifier performance for the different predicted latent steps in the future.	32
5.3	AUROC for the classification task on the CPC features and on the raw data under varied sizes of label fractions	33

Symbols

Symbol	Typical meaning
X, Y	Random variables are uppercase
x, y	Realizations of random variables are lowercase
τ, λ	Scalars are lowercase
$\mathbf{x}, \mathbf{y}, \mathbf{z}$	Vectors are bold lowercase
\mathbf{W}	Matrices are bold uppercase
$\mathbf{x}^T, \mathbf{W}^T$	Transpose of a vector or a matrix
\mathbb{R}	Real numbers
\mathbb{R}^d	d dimensional vector space of real numbers
N	Number of data points
$\boldsymbol{\theta}$	Parameter vector
$\mathcal{D}, \mathcal{X}, \mathcal{Y}$	Sets

Abbreviations

Artificial Intelligence	AI
Machine Learning	ML
Neural Network	NN
Deep Neural Network	DNN
Deep Learning	DL
Out-of-Distribution	OOD
Mutual Information	MI
Convolutional Neural Network	CNN
One-Class Classifiers	OCC
Kullback-Leibler divergence	KL divergence
Multilayer Perceptron	MLP
Rectified Linear Unit	ReLU
Independent Identically Distributed	iid
Negative log likelihood	NLL
Normalized Temperature-scaled Cross-entropy	NT-Xent
Atrial Fibrillation	AF
Inter-Beat Interval	IBI
Contrastive Predictive Coding	CPC
Area Under the Receiver Operating Characteristics	AUROC
Serious Clinical Complication	SCC

Chapter 1

Introduction

In this chapter, we review some basic concepts playing fundamental role in this thesis.

Humans have been trying to understand their own intelligence and explain it by a few principles instead of heuristics and build intelligent machines [12]. This has provided inspiration for *Artificial Intelligence* research. *Artificial Intelligence* (AI) is any technique which enables computers to mimic human behaviour and focuses on building algorithms to process information. Russell et al. [13], in book *Artificial Intelligence, A Modern Approach*, mention conceivable goals to achieve in AI which are building systems that think like humans, act like humans, think rationally, and act rationally. AI aims at distilling human knowledge and understanding into a suitable form for building machines [14]. *Machine Learning* (ML) is a subfield of AI where focuses on teaching algorithms to process information to inform future decisions without being explicitly programmed to do a task. Traditional ML algorithms typically define a set of features from data. These features are important but hand-engineered and brittle in practice which highlights the weakness of the ML algorithms. For more complex tasks, it is infeasible to know what features should be extracted. For example, if our goal is to detect a face in a given image, we can begin with recognising different parts of face such as nose and ears and after detecting these parts we can say there might be a face in the image. Now the question is how to recognise them? This is where the problem becomes complicated. Most of the ML algorithms have a shallow understanding of the data and are tremendously dependent on representations they are fed as input [15]. Making machine learning algorithms less dependent on feature engineering would be advantageous. This is where deep neural networks (DNNs) make a difference.

Deep Learning (DL) is a subfield of ML which can automatically extract useful pieces of information needed to solve the task at hand. In fact, DNNs derive complex patterns from raw data by composing low level features hierarchically to detect higher level features. For example, in the face detection task we need to take a bunch

of images of faces. DL algorithms attempt to develop hierarchical representations of first detecting low level features such as lines and edges then using these low level features to detect mid-level features like eyes, nose and ears. Finally, composing these features leads to detecting higher level features like facial structure. There has been a gap between learning algorithms and human abilities [12], to bridge the gap towards human-level intelligence, learning generalizable and reliable representations would be desirable. **Representation learning** works by learning a more compact and generalizable representation of the input data, making it easier to find patterns and also giving a better understanding of the data. The learnt representations not only solve the defined task, but because of their generalization properties [15], they can be re-purposed to solve a downstream task of interest. Representation learning methods can be mainly *supervised*, *unsupervised*, or *self-supervised*. In the following section we introduce each of them.

1.1 Learning Strategies

Supervised Learning is a training strategy that relies on labeled data. A model learns a mapping from inputs to outputs. Training highly effective supervised deep learning models usually requires to expose the model to a decent amount of labeled training data. Data annotation is usually time consuming and expensive. On the other hand, data distributions constantly shift which needs more and new data while collecting a large amount of unlabeled data is not difficult. Even with a large number of labeled examples supervised learning has still blind spots in terms of learning useful and rich representations [16].

Unsupervised Learning is a learning approach that allows learning from unlabeled data. Unlike supervised learning that a model is given a set of input and output data, in unsupervised learning only the input data are available and the model must extract patterns from the data. However, unsupervised learning has limited power compared with supervised learning which has access to label information.

Self-supervised Learning is a form of unsupervised learning where raw data (not human) provide the supervision. The self-supervised learning approach aims at learning semantically meaningful features from unlabeled data. Self-supervised tasks, also called pretext tasks, help in learning representations which are beneficial to other downstream tasks such as classification and segmentation. Improving representations require learning features that are not specialized for solving a specific task but rather capturing rich statistics for different downstream tasks. For most of pretext tasks, a part of the data is withheld and the model has to predict it.

A self-supervised task can be realized by predicting a subset of information using the rest, e.g., for a time series sequence the task could be predicting the future from the past [17] or learning the semantic similarities between different patches

of an image [18]. Self-supervised learning has drawn extensive research interests in different domains from language to image and audio [3, 17, 19]. Some believe that self-supervised learning is more close to achieving human-level intelligence and is more human-like in its reasoning¹. Since the representation has to be learnt from the data itself without any additional input, the information-theoretic principles have been formulated. One could be maximizing similarity between different levels of representations measured by *Mutual Information* (MI) [20] which is a measure of dependency. MI captures non-linear dependency between variables and can be considered as a measure of true dependence [21].

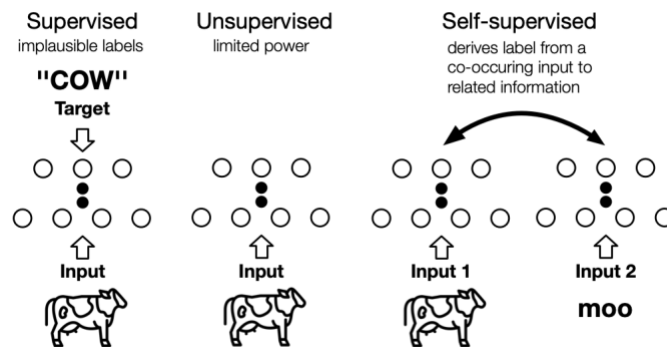


Figure 1.1: An illustration to distinguish the supervised, unsupervised and self-supervised learning frameworks. Taken from [1].

In practice, MI is difficult to compute. The intrinsic difficulty of the MI estimation stems from the fact that the MI is a nonlinear function of a joint probability measure and typically the space of probability measures can be infinitely large [22]. Exact computation is tractable only for discrete variables and a limited family of probability distributions. Estimating MI is challenging but many techniques have been proposed to address this problem. Recent works [17, 18, 23, 24] have combined variational bounds with deep learning to enable tractable estimation of MI.

Most of the prior works aim at improving representations by proposing novel pretext tasks [4, 5, 7]. Self-supervised learning [4, 25] has shown outstanding success in representation learning which can be used in downstream tasks such as classification, object detection, and anomaly detection.

1.2 Anomaly Detection

Anomaly detection or *Out-of-Distribution* (OOD) detection is the problem of deciding whether a given test sample is drawn from the same distribution as the

¹<https://venturebeat.com/2020/05/02/yann-lecun-and-yoshua-bengio-self-supervised-learning-is-the-key-to-human-level-intelligence/>

training data or it belongs to an alternative distribution. This problem can be formulated as a binary classification problem that classifies examples as in-distribution or OOD, given a sufficiently large sample from the in-distribution. An alternative approach is to learn a density model from the training data and compute likelihood of OOD examples [26, 27]. However, in practice, this approach frequently has failed for high-dimensional data [28], where it has been shown that deep generative models can assign higher likelihood to OOD examples than to in-distribution examples. A major challenge in OOD detection is the case where the features of outlier examples are statistically close to the features of in-distribution examples, which is often the case for natural images. Recently numerous self-supervised tasks have been proposed that enable richer feature learning [3, 19] more suited for OOD detection. Particularly, contrastive learning [29] has shown state-of-the-art results on visual representation learning [3, 19]. Contrastive self-supervised learning is an approach to learn useful representations by solving a pretext task which pulls semantically similar examples closer while pushing away from others. Most of the recent state-of-the-art anomaly detection methods have utilized self-supervised contrastive training [30, 31].

1.2.1 Anomaly Detection in Medical Health Screening

Recently, deep learning has benefited medical diagnosis [32–34]. Diagnosing if a sample includes any abnormality can help medical experts with a reliable early treatment and can improve decision making. In health screening, typically the problem of anomaly detection is addressed by training a binary classifier on healthy and unhealthy samples which requires experts to analyse a large amount of data where the dataset is highly imbalanced with the large majority of cases comprising normal samples and a small minority consisting of abnormal samples. In this regard, many studies have been conducted to deal with imbalanced learning [35]. An alternative to fully supervised imbalanced learning is few-shot anomaly detection [36] using an imbalanced training set, containing a large number of normal and a few abnormal samples which is inspired by how humans can generalize from only a handful of examples. Existing unsupervised methods are usually formulated as one-class classifiers (OCC) [37, 38]. Prevailing state-of-the-art OCC methods [37] train a neural network to minimize the volume of a hypersphere that encloses the representations of the in-distribution training data in the representation space. Test images are classified as in-distribution if they fall inside the hypersphere and classified as OOD if they fall outside. Nonetheless, Chen et al. [38] have shown that the OCC models often overfit the training data, especially when the training set is small or contaminated with anomalies. Recently the self-supervised learning has permeated the field of medical image analysis. Due to their efficiency, self-supervised pretext tasks such as predicting geometric transformations [39] or contrastive learning [34] have been designed for OOD detection.

1.3 Thesis Outline

This thesis is principally devoted to experimental investigations concerning the issue of self-supervised representation learning particularly for anomaly detection in different visual and time series domains. This thesis is structured as follows.

In chapter 2, we recapitulate the basics of *Information Theory*, which provides the basis for the machine learning applications in the subsequent chapters. Notations and derivations are inspired by the book *Elements of information theory* written by Cover and Thomas [40].

In chapter 3, we concisely introduce some basic concepts in Machine Learning.

Chapter 4 is devoted to an extensive discussion on self-supervised pretext tasks and contrastive learning. In this chapter, we investigate behaviour of the contrastive loss.

In chapter 5, we study the application of self-supervised representation learning in medical time series. We exploit self-supervised representation learning to detect atrial fibrillation which is the most common arrhythmia having a major impact on mortality.

Chapter 6 addresses pneumonia detection in chest X-ray image. We combine a self-supervised contrastive method with a Mahalanobis distance score function to develop an abnormality detection method that uses only healthy images during training.

The next chapter, chapter 7, shows how we can make use of self-supervised contrastive learning combined with cosine similarity as a score function to detect serious clinical complications in patients receiving oncological treatment for their hematologic malignancies.

In chapter 8, we present a self-supervised method which leverages self-distillation and negative samples for the task of abnormality detection. We show that self-distillation of the in-distribution training data together with contrasting against negative examples derived from shifting transformations can improve OOD detection performance in the visual domain in both natural and medical images.

In chapter 9, we discuss the objectives and achievements in the thesis and proposals for future developments are addressed.

Chapter 2

Basic Information Measures

In this chapter, we introduce basic information measures and theoretical concepts necessary for subsequent chapters. We denote random variables using uppercase letters, e.g. X , and their realizations by the corresponding lowercase letters, e.g. x . For simplicity, we denote probability mass function of a discrete random variable by $p(x)$ instead of $p_X(x)$. We also denote probability density function of a continuous random variable by $f(x)$ rather than $f_X(x)$.

2.1 Entropy

The *entropy* is a measure of uncertainty of a random variable or the average amount of information carried by a random variable.

Let X be a discrete random variable with probability mass function $p(x)$. The entropy of X , denoted by $H(X)$ or $H(p)$, is defined as

$$H(X) = - \sum_x p(x) \log p(x) \quad (2.1)$$

the common values being used for the base of the logarithm are 2 and Euler's number e . The units of entropy are named bits and nats correspondingly. For a discrete random variable X , the entropy is the average number of bits needed to describe the random variable.

Definition. Let $f(x)$ be probability density function of a continuous random variable X . The *differential entropy* of X , denoted by $h(X)$ or $h(f)$, is defined as

$$h(X) = - \int f(x) \log f(x) dx \quad (2.2)$$

since a probability density function can take arbitrarily large values, the differential

entropy can take negative values so the differential entropy can not be a measure of information that the random variable X is carrying. In general, a continuous random variable can carry an infinite amount of information.

Definition. The *conditional entropy* is the uncertainty of a random variable, X , given another random variable Y . The conditional entropy $H(X|Y)$ is defined as

$$H(X|Y) = - \sum_{(x,y)} p(x,y) \log p(x|y) \quad (2.3)$$

2.1.1 Properties of Entropy

In the following, we mention the most important properties of entropy.

1. For any discrete random variable X , $H(X) \geq 0$, because the maximum value a probability mass function can take is 1 while the differential entropy can take negative values.
2. Conditioning always reduces the entropy

$$H(X) \geq H(X|Y) \quad (2.4)$$

we prove this property later using the non-negativity of mutual information.

3. **Chain rule.** The joint entropy can be decomposed as follows

$$H(X, Y) = H(Y) + H(X|Y) \quad (2.5)$$

$$= H(X) + H(Y|X) \quad (2.6)$$

Proof.

$$\begin{aligned} H(X, Y) &= - \sum_x \sum_y p(x, y) \log p(x, y) \\ &= - \sum_x \sum_y p(x, y) \log p(y) p(x|y) \\ &= - \sum_x \sum_y p(x, y) \log p(y) - \sum_x \sum_y p(x, y) \log p(x|y) \\ &= - \sum_y p(y) \log p(y) - \sum_x \sum_y p(x, y) \log p(x|y) \\ &= H(Y) + H(X|Y) \end{aligned}$$

similarly we can show that the entropy of n random variables is the sum of conditional entropies. let X_1, X_2, \dots, X_n be n random variables with joint probability

mass function $p(X_1, X_2, \dots, X_n)$

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) \quad (2.7)$$

4. Let X be a random variable and $g(X)$ be a deterministic function of X

$$H(X) \geq H(g(X)) \quad (2.8)$$

Proof. Expanding $H(X, g(X))$ in two different ways

$$\begin{aligned} H(X, g(X)) &= H(X) + H(g(X)|X) \\ &= H(g(X)) + H(X|g(X)) \end{aligned}$$

$H(g(X)|X) = 0$ because by knowing X there is no uncertainty left about $g(X)$. $H(X) = H(g(X))$ if and only if g is invertible. If we have a unique mapping from X to $g(X)$, both reduce the same amount of uncertainty about each other.

2.2 Relative Entropy

The *relative entropy* or *Kullback-Leibler (KL) divergence* [41] is a quantity that measures distance or dissimilarity between two distributions.

The Relative entropy for discrete probability distributions p and q defined as

$$D_{kl}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \quad (2.9)$$

$$= H(p, q) - H(p) \quad (2.10)$$

for probability density functions the summation is replaced by an integral

$$D_{kl}(f||g) = \int f(x) \log \frac{f(x)}{g(x)} dx \quad (2.11)$$

in the above definition, we set $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$.

2.2.1 Properties of Relative Entropy

In the following, we review some properties of relative entropy.

1. The KL divergence is not symmetric. $D_{kl}(p||q)$ is not necessarily equal to $D_{kl}(q||p)$.

2. For any p and q , $D_{kl}(p||q) \geq 0$.

To prove this property, we need to introduce Jensen's Inequality. If φ is a convex

function on its domain and X is a random variable

$$\mathbb{E}[\varphi(X)] \geq \varphi[\mathbb{E}(X)] \quad (2.12)$$

the proof can be found in [40].

Proof. Let p and q be two discrete probability distributions on the same probability space

$$\begin{aligned} D_{kl}(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= - \sum_x p(x) \log \frac{q(x)}{p(x)} \\ &= - \mathbb{E}_{p(x)} \left[\log \frac{q(X)}{p(X)} \right] \end{aligned}$$

using Jensen's Inequality and this fact that $-\log x$ is a convex function

$$\begin{aligned} D_{kl}(p||q) &\geq - \log \mathbb{E}_{p(x)} \left[\frac{q(X)}{p(X)} \right] \\ &= - \log \sum_x p(x) \frac{q(x)}{p(x)} \\ &= - \log 1 \\ &= 0 \end{aligned}$$

$D_{kl}(p||q) = 0$ if and only if $p \equiv q$.

2.2.2 Variational Inference

Variational inference is a deterministic method vastly used for approximate inference [42]. Variational is a general term used for problems that the inference is reduced to an optimization problem.

Suppose the goal is to find an approximate to a distribution p where the approximate distribution makes inference simpler. Let \mathcal{Q} be a family of tractable distributions where $q \in \mathcal{Q}$ is an approximate to the true distribution p . Suppose q has some free parameters which are optimized to make q as similar as possible to the true distribution p . A good choice for cost function is the KL divergence to be minimized. Forward KL divergence, also known as an **M-projection** or **moment projection** [43] is defined as

$$\begin{aligned} D_{kl}(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= - \mathbb{E}_{p(x)} [\log q] - H[p] \end{aligned} \quad (2.13)$$

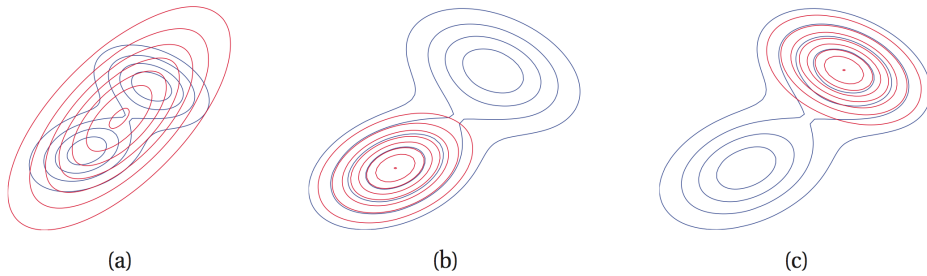


Figure 2.1: Illustrating forward vs reverse KL divergence on a bimodal distribution Taken from [2]. The blue curves are the contours of the true distribution p . The red curves are the contours of the unimodal approximate distribution q . (a) Minimizing forward KL divergence (b-c) Minimizing reverse KL divergence

note that inference on p is required which is assumed to be intractable. The forward KL divergence is stated to be **zero avoiding** [43] for q because if $p > 0$ we must ensure $q > 0$ otherwise $D_{kl}(p||q)$ is infinite.

An alternative is reverse KL divergence, also known as an **I-projection** or **information projection** [43]

$$\begin{aligned} D_{kl}(q||p) &= \sum_x q(x) \log \frac{q(x)}{p(x)} \\ &= -\mathbb{E}_{q(x)}[\log p] - H[q] \end{aligned} \quad (2.14)$$

the family of distributions \mathcal{Q} is chosen such that the expectation w.r.t q is tractable. The reverse KL divergence is stated to be **zero forcing** [43] for q because if $p = 0$ we must ensure $q = 0$ otherwise $D_{kl}(q||p)$ is infinite for $q > 0$.

The difference between these two methods is illustrated in Figure 2.1 taken from [2]. Suppose the true distribution p is a bimodal Gaussian distribution and q is constrained to be a unimodal Gaussian distribution. Minimizing the forward KL divergence yields an approximate distribution q that tends to cover both modes of p and its mode is in a low density region while by minimizing the reverse KL divergence q gets trapped in one of the modes.

2.3 Mutual Information

The *mutual information* (MI) between two random variables X and Y is defined as the relative entropy between the joint distribution and the product of their marginals

$$I(X; Y) = D_{KL}(p(x, y) || p(x)p(y)) \quad (2.15)$$

if X and Y are two discrete random variables with joint probability mass function $p(x, y)$

$$I(X; Y) = \sum_{(x,y)} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (2.16)$$

if X and Y are two continuous random variables with joint probability density function $f(x, y)$

$$I(X; Y) = \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)} dx dy \quad (2.17)$$

$I(X; Y)$ can be equivalently expressed as

$$I(X; Y) \equiv H(X) - H(X | Y) \quad (2.18)$$

$$\equiv H(Y) - H(Y | X) \quad (2.19)$$

$$\equiv H(X) + H(Y) - H(X, Y) \quad (2.20)$$

The mutual information, $I(X; Y)$, can be interpreted as the amount of information one random variable carries about another random variable or the amount of uncertainty reduction about a random variable after observing the other one.

2.3.1 Properties of Mutual Information

In the following we investigate the most significant properties of MI.

1. For any two random variables X and Y : $I(X; Y) \geq 0$. Equality holds if and only if X and Y are independent.

Proof. $I(X; Y) = D_{kl}(p(x, y) || p(x)p(y)) \geq 0$. This non negativity results in $H(X) \geq H(X | Y)$.

2. The mutual information is symmetric that can be inferred from its definition.

$$I(X; Y) = I(Y; X) \quad (2.21)$$

therefore X knows as much about Y as Y knows about X .

3. **Chain rule.**

$$I(X, Y; Z) = I(X; Z) + I(Y; Z|X) \quad (2.22)$$

Proof. Expanding the mutual information and using the chain rule for the entropy

$$\begin{aligned}
I(X, Y; Z) &= H(X, Y) - H(X, Y|Z) \\
&= H(X) + H(Y|X) - H(X|Z) - H(Y|X, Z) \\
&= H(X) - H(X|Z) + H(Y|X) - H(Y|X, Z) \\
&= I(X; Z) + I(Y; Z|X)
\end{aligned} \tag{2.23}$$

Similarly for a set of random variables $X_1, X_2, \dots, X_n; Y$

$$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^n I(X_i, Y|X_1, X_2, \dots, X_{i-1}) \tag{2.24}$$

4. The mutual information is invariant under reparametrization. If f and g are invertible and deterministic functions and $X' = f(X)$ and $Y' = g(Y)$, then

$$I(X; Y) = I(X'; Y')$$

since f is an invertible function, random variables X, X' and (X, X') are equivalent in terms of entropy and provide the same amount of information about random variable Y

$$\begin{aligned}
I(X, X'; Y, Y') &= H(X, X') - H(X, X'|Y, Y') \\
&= H(X') - H(X'|Y') \\
&= I(X'; Y')
\end{aligned}$$

5. Conditioning can either increase or decrease the mutual information. To prove this property we first need to define Markov chain.

Definition. Let X, Y and Z form a *Markov chain*, $X \rightarrow Y \rightarrow Z$, the joint probability mass function can be written as

$$p(x, y, z) = p(x)p(y|x)p(z|y)$$

markovity implies conditional independence

$$\begin{aligned}
p(x, z|y) &= \frac{p(x, y, z)}{p(y)} \\
&= \frac{p(x)p(y|x)p(z|y)}{p(y)} \\
&= \frac{p(y, x)p(z|y)}{p(y)} = p(x|y)p(z|y)
\end{aligned}$$

to prove non-monotonicity property of mutual information, we consider two following cases [40]

- If X, Y and Z form a Markov chain, $X \rightarrow Y \rightarrow Z$, then $I(X; Y|Z) \leq I(X; Y)$.

Proof. The mutual information $I(X; Y, Z)$ can be expanded in two different ways by the chain rule as follows

$$\begin{aligned} I(X; Y, Z) &= I(X; Z) + I(X; Y|Z) \\ &= I(X; Y) + I(X; Z|Y) \end{aligned}$$

by Markovity $I(X; Z|Y) = 0$ and from non-negativity property of the mutual information $I(X; Z) \geq 0$, then

$$I(X; Y|Z) \leq I(X; Y)$$

- If X , Y and Z don't form a Markov chain, it could be possible that $I(X; Y|Z) \geq I(X; Y)$. Let X and Z be two independent random variables, i.e., $I(X; Z) = 0$, then

$$I(X; Y|Z) \geq I(X; Y)$$

2.3.2 Data Processing Inequality

The data processing inequality demonstrates that "no clever manipulation of the data can improve the inferences that can be made from the data" [40] or in the other word "there is no processing of Y , deterministic or random, can increase the information that Y contains about X " [40].

Data Processing Inequality. If X , Y and Z form a Markov chain, $X \rightarrow Y \rightarrow Z$, then $I(X; Y) \geq I(X; Z)$.

Proof. Taking two different expansions of $I(X; Y, Z)$ by the chain rule

$$\begin{aligned} I(X; Y, Z) &= I(X; Z) + I(X; Y|Z) \\ &= I(X; Y) + I(X; Z|Y) \end{aligned}$$

by Markovity $I(X; Z|Y) = 0$ and $I(X; Y|Z) \geq 0$

$$I(X; Y) \geq I(X; Z) \tag{2.25}$$

equality holds if and only if $I(X; Y|Z) = 0$.

2.3.3 Information Bottleneck Principle

In [44], the relevant information in a signal x is defined as the information that x provides about another signal y . One example can be the information that images provide about their labels. Understanding the signal x requires more than just predicting y , it also requires specifying which features of x are involved in the prediction. Tishby et al. [44] formulated this problem as finding a short code for X that captures the maximum information about Y . The information that X provides

about Y is squeezed through a "bottleneck" formed by a compact representation. Let the compressed representation of X be given by random variable T . The information bottleneck proposed by Tishby et al. [44] expresses the trade-off between the mutual information measures $I(X; T)$ and $I(T; Y)$. There is a trade-off between compressing the representation, i.e., minimizing $I(X; T)$, and preserving meaningful information, i.e., maximizing $I(T; Y)$.

2.3.4 Sufficient Statistics

Suppose $\{f_{\theta}(x)\}$ is a family of probability mass functions parameterized by θ and X is a sample drawn from a probability distribution $f_{\theta}(x)$ in this family. Any function of X , $T(X)$, is called a statistic and for any statistic $T(X)$

$$\theta \rightarrow X \rightarrow T(X) \tag{2.26}$$

by the data processing inequality, $I(\theta; X) \geq I(\theta; T(X))$.

Definition. $T(X)$ is called a sufficient statistic relative to the family $\{f_{\theta}(x)\}$ if the conditional probability distribution of X , given the statistic $T(X)$, does not depend on the underlying parameter θ which induces the following Markov chain

$$\theta \rightarrow T(X) \rightarrow X \tag{2.27}$$

sufficient statistic $T(X)$ knows as much information about θ as X knows, so

$$I(\theta; X) = I(\theta; T(X)) \tag{2.28}$$

Chapter 3

Machine Learning Basics

The goal of this chapter is to provide a brief introduction to the most important concepts in machine learning (ML).

3.1 Neural Networks

3.1.1 Perceptron

A *perceptron* is the simplest neural network (NN) possible [45], Figure 3.1. We can design the simplest discriminator by a Perceptron. A discriminator takes an input vector \mathbf{x}_n and assigns it to one of K classes. Suppose we observe N data pairs $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$. Let us assume that the data points $\mathbf{x}_n \in \mathbb{R}^d$, and labels $y_n \in \{-1, +1\}$. The perceptron maintains a weight vector $\mathbf{w} \in \mathbb{R}^d$. For each feature x_i^n in the feature vector \mathbf{x}_n there is a corresponding weight w_i . The Perceptron corresponds to a linear two-class classifier which assigns data point \mathbf{x}_n to class 1 if $\hat{y}_n > 0$ otherwise to class -1 .

$$\hat{y}_n = w_0 + \sum_{i=1}^d x_i^n w_i = w_0 + \mathbf{x}_n^T \mathbf{w} \quad (3.1)$$

where w_0 is bias. If it is desirable to output a number representing a probability we can apply a non-linear activation function such as sigmoid function, $\sigma(\cdot)$, that takes any real value and transforms it to a value between 0 and 1. The classifier can correctly classify all samples, if they are linearly separable.

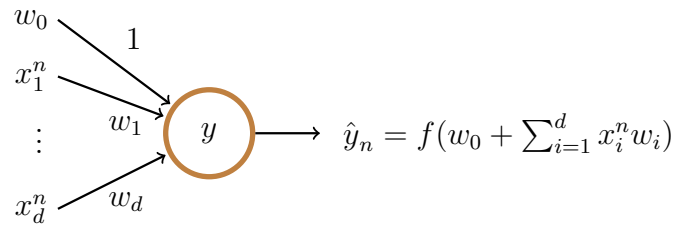


Figure 3.1: Illustration of a perceptron. The activation function is denoted by f . $[x_1^n, \dots, x_d^n]^T$ represents feature vector \mathbf{x}_n . w_0 is called bias and represents an external input.

3.1.2 Multilayer Perceptron

A Multilayer Perceptron (MLP) consists of at least three layers, an input layer, an output layer, and a layer between the input and output layers referred to as “hidden” layer [46]. MLPs are fully connected layers, each node in a layer is connected to every node in the following layer with a corresponding weight. Neural networks imitate the structure of the brain, see Figure 3.2. When a NN contains multiple hidden layers it is considered as a “deep” NN (DNN).

In practice, data are highly non-linear. To introduce non-linearity to NNs, we can take non-linear activation functions such as sigmoid function, rectified linear unit (ReLU) or hyperbolic tangent (tanh). Linear activation functions build linear decision boundaries no matter how large the NN is. Except for the nodes in the input layer, each node is followed by a nonlinear activation function. Introducing non-linearity to the NNs enables approximating arbitrarily complex functions.

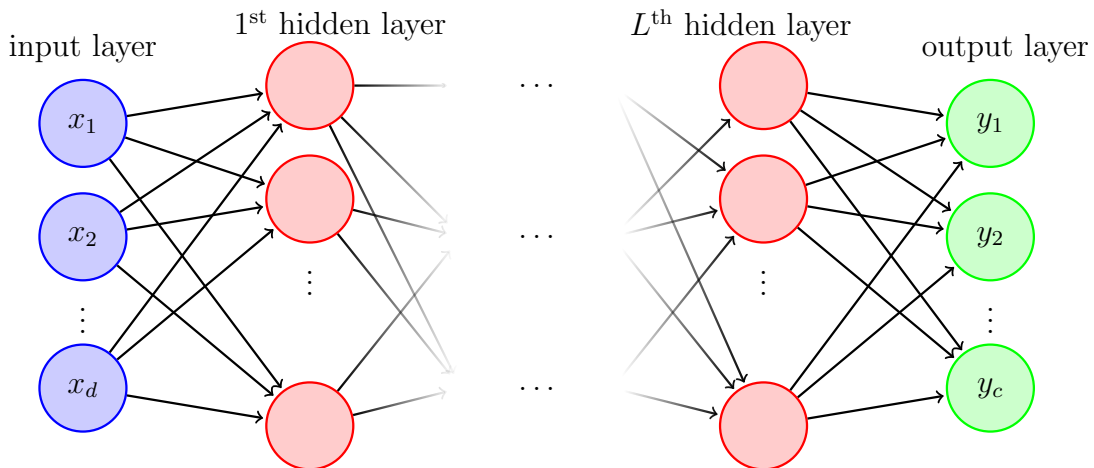


Figure 3.2: Multilayer perceptron with L hidden layers. Input and output layers have d and c nodes, respectively. The l^{th} hidden layer contains $n^{(l)}$ hidden nodes.

3.2 Machine Learning Algorithms

Most machine learning algorithms fall into one of supervised and unsupervised categories. Classification and regression are two examples of supervised learning. In a classification task, aim is to predict a discrete class label and regression is a task of predicting a continuous quantity.

Consider a classification task with K classes, model is presented with a set of observations, $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, along with corresponding target values $\mathcal{Y} = \{y_1, \dots, y_N\}$ where \mathbf{x}_n is a d dimensional feature vector, $\mathbf{x}_n \in \mathbb{R}^d$, and y_n is a K -way categorical random variable. Suppose we aim to predict the class label y given the input vector \mathbf{x} . Let $p_{data}(y|\mathbf{x})$ be the conditional probability distribution over classes and $p(y|\mathbf{x}, \boldsymbol{\theta})$ be an estimation of $p_{data}(y|\mathbf{x})$ parameterized by a neural network. At the output layer of the network, outputs are mapped to a distribution over classes through a softmax function

$$p(y_n = k|\mathbf{x}_n, \boldsymbol{\theta}) = \frac{e^{h_{\boldsymbol{\theta}}^k(\mathbf{x}_n)}}{\sum_{k=1}^K e^{h_{\boldsymbol{\theta}}^k(\mathbf{x}_n)}} \quad (3.2)$$

where $h_{\boldsymbol{\theta}}^k$ is called logit associated with class k . In order to determine parameters $\boldsymbol{\theta}$, we maximize the conditional likelihood of the training data. Given a dataset of N independent, identically distributed (iid) samples, we assume there is no dependency between them, the conditional likelihood function can be constructed as

$$p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) = \prod_{n=1}^N p(y_n|\mathbf{x}_n, \boldsymbol{\theta}) \quad (3.3)$$

to avoid numerical problems such as underflow arising from multiplying small probabilities we maximize the conditional likelihood in log space

$$\begin{aligned} \boldsymbol{\theta}_{ML} &= \arg \max_{\boldsymbol{\theta}} \log \prod_{n=1}^N p(y_n|\mathbf{x}_n, \boldsymbol{\theta}) \\ &= \arg \max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^N \log p(y_n|\mathbf{x}_n, \boldsymbol{\theta}) \end{aligned} \quad (3.4)$$

equivalently

$$\boldsymbol{\theta}_{NLL} = - \arg \min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^N \log p(y_n|\mathbf{x}_n, \boldsymbol{\theta}) \quad (3.5)$$

Negative log likelihood (NLL), Eq. 3.5, is equivalent to cross entropy loss. For a binary classification task, Eq. 3.5 is a binary cross entropy loss and is a categorical cross entropy for multi-class problems.

In unsupervised learning, model is presented with a set of observations, \mathcal{X} , attempting to learn underlying structure of the data. Clustering and density estimation are examples of unsupervised learning. Clustering is a task of grouping samples such that semantically similar ones fall into the same cluster. In the density estimation problem, we try to define a model to estimate the distribution which the samples were drawn from. One way is maximising the likelihood of training data which for N iid samples is defined as $p(\mathcal{X}|\boldsymbol{\theta}) = \prod_{n=1}^N p(\mathbf{x}_n|\boldsymbol{\theta})$.

3.3 Generalization

A NN should fully learn the training data and be able to perform well on unseen data. If a model has enough capacity, e.g., more parameters than training samples, it is able to “memorize” each training sample. The model overfits the training data which yields poor generalization to unseen test data [47]. To alleviate overfitting, different regularization methods have been introduced. Regularization is “any modification we can make to the learning algorithm that is intended to reduce the generalization error but not its training error” [47].

A common way to combat overfitting is to add l_1 -norm or l_2 -norm regularization as a penalty term to loss function which forces parameters to have smaller norms [48]. An alternative form of regularization is early stopping for over-parameterized models, such as large deep networks [49], when model’s performance gets worse on a validation set.

Another way for preventing overfitting is to increase the size of training set by using data augmentation. To keep semantics intact, data augmentation methods need expertise to design augmentations which requires domain adaption. For images, common augmentations include translation, rotation, and sharpening [50]. For a text classification task, the augmentation method can be back-translation [51] which refers to translating a sentence into another language then translating it back which can help train a robust model.

3.4 Convolutional NNs vs. Transformers

Convolutional neural networks (CNNs) have revolutionized computer vision field [52]. CNNs extract visual features and are particularly designed for images that can be computationally expensive which make needs for computationally efficient architectures to achieve state-of-the-art results. Recently, Vision Transformers (ViTs) [53] have emerged as a powerful tool in computer vision, however, CNNs remain dominant. ViTs have been developed based on Transformers [54] originally designed for text-based tasks. An image is divided into multiple patches and then is unrolled into a sequence of patches to be fed to the transformer. ViTs outperform or perform on par with CNNs on many image classification datasets, if trained on large

datasets [53]. ViTs lack some of the inherent spatial inductive biases of CNNs, such as translation equivariance and locality, so they do not generalize well when trained on insufficient amounts of data. Dosovitskiy et al. [53] have shown that ViTs result in excellent performance when pre-trained on adequate amounts of data and can be fine-tuned for tasks with smaller datasets. Unlike CNNs, ViTs are capable of capturing long-range correlations [55], which is crucial for learning high-level semantics.

Zhou et al. [55] have explored the transferability of the learnt representations of CNNs and ViTs on various datasets and showed that ViTs provide more transferable and generalizable representations than CNNs. ViTs are more prone to overfitting which needs strong data augmentation or regularization [56,57] to tackle overfitting.

Chapter 4

Self-Supervised Representation Learning

Self-supervised Learning is a form of unsupervised learning where raw data (not human) provide the supervision. The self-supervised learning is an approach concerned with learning semantically meaningful features from unlabeled data. The self-supervised tasks help in learning representations which are beneficial to other downstream tasks. Improving representations require learning features that are not specialized for solving a specific task but rather capturing rich statistics for different downstream tasks. The self-supervised pretext tasks also require labels (or pseudo-labels) for optimization which are derived from the data alone. For most of pretext tasks, a part of the data is withheld and the model has to predict it. Recent works have focused on designing novel pretext tasks, such as context prediction [5], jigsaw puzzle [8], colorization [9, 58], rotation [4]. Many methods have been proposed for self-supervised representation learning on images, each exploring a different pretext task. In the most common pretext tasks, some transformations applied to inputs while the semantics stay unchanged. In this chapter, we draw on existing self-supervised pretext tasks.

4.1 How to define pretext tasks?

In *SimCLR* [3], geometric transformations and color transformations were applied to images to learn representations. Geometric transformations such as random cropping, scaling, horizontal flipping and vertical flipping don't change the pixel information. Color transformations such as blurring, color distortion and converting to gray-scale can be applied to images such that model learns representations invariant to the color changes. The pretext task is defined as maximizing the agreement between different augmentations of the same image, see Figure 4.1.

In [4], a model is trained to recognize the discrete geometric transformation applied to an image. An image is first rotated by 0, 90, 180, and 270 degrees then the model is trained on a 4-way image classification task to recognize the rotation was applied to the image, see Figure 4.2. They argue that in order for a model to be able to identify the rotation applied to an image the model needs to understand the concepts of the objects in the image such as their location in the image, their type, and their pose.

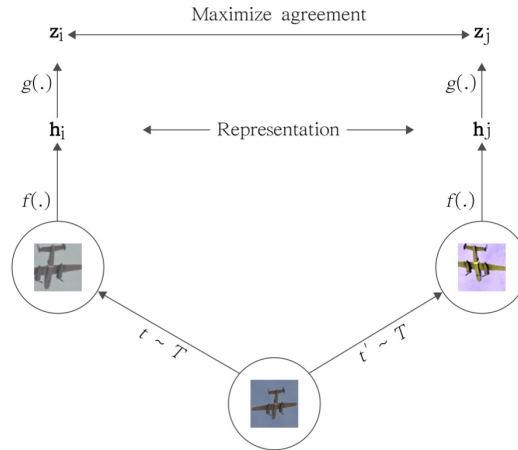


Figure 4.1: A simple framework for visual representation learning. Each image is augmented twice. Two separate data augmentation operators sampled from the same family of augmentations applied to the image to obtain two correlated views. Taken from [3].

Doersch et al. [5] formulated the pretext task as predicting the relative position between two random patches from one image, see Figure 4.3. When designing a pretext task, one must ensure that the task forces the network to extract the desired information such as high level semantics, without taking “trivial” shortcuts. To prevent the model from catching only low level trivial information like textures continuing between patches they introduced some additional noise such as including a gap between patches.

In [7], a model is trained to discriminate between a set of surrogate classes. Each surrogate class is formed by applying a variety of transformations to a randomly sampled image patch. The resulting distorted patches are considered to belong to the same surrogate class, see Figure 4.4.

Noroozi et al. [8] trained a convolutional neural network (CNN) to solve jigsaw puzzles as a pretext task and then learnt representations are repurposed to solve object classification and detection, see Figure 4.5.

In [9], the pretext task is defined as training a model to color a grayscale input image. The model maps the grayscale image to a distribution over quantized color

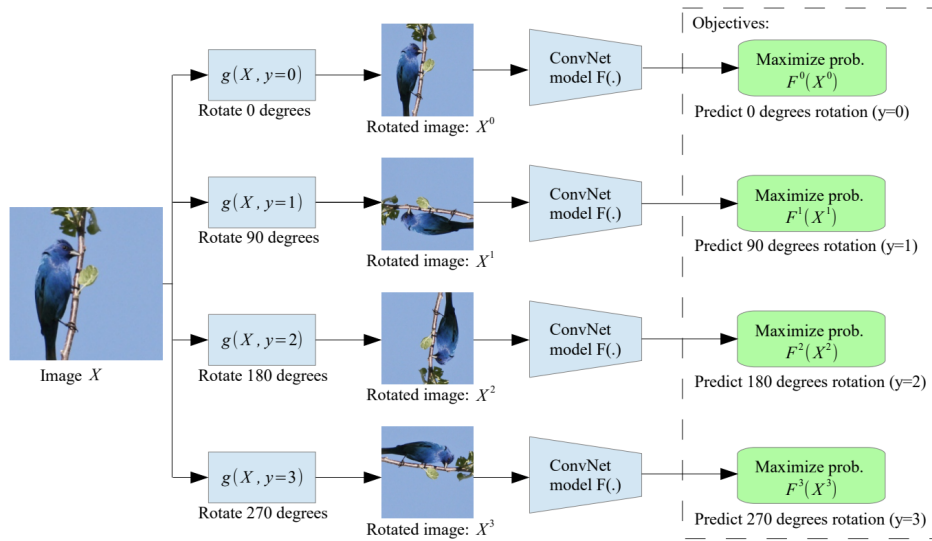


Figure 4.2: Illustration of the self-supervised task by applying geometric transformations to the input image. The model learns to predict which rotation was applied. Taken from [4].

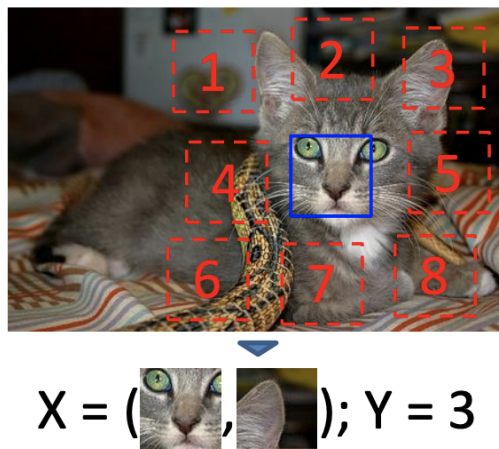


Figure 4.3: Illustration of the self-supervised task by predicting relative position between randomly sampling a patch (blue) and one of eight possible neighbors (red). Taken from [5].

value outputs. It has been shown that colorization can be a powerful pretext task for self-supervised feature learning, see Figure 4.6.

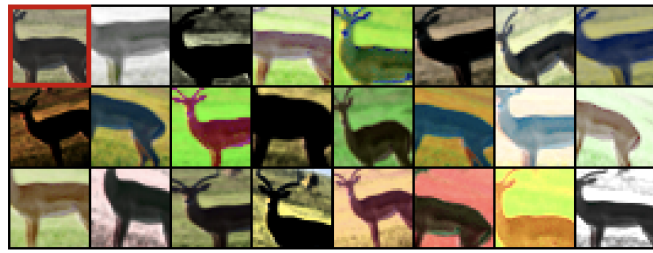


Figure 4.4: Transformations applied to one of the patches extracted from the STL dataset [6]. The original patch is in the top left-hand corner. Taken from [7].

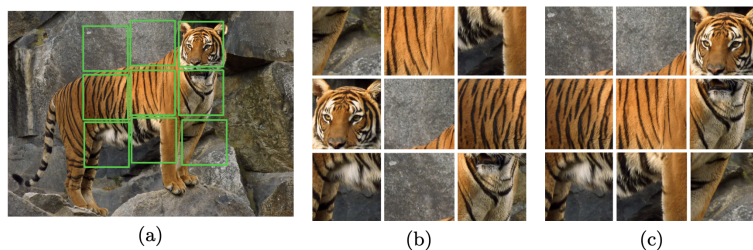


Figure 4.5: Illustration of self-supervised learning by solving jigsaw puzzle. The tiles marked with green lines are extracted from the image and a puzzle obtained by shuffling the tiles. Taken from [8].

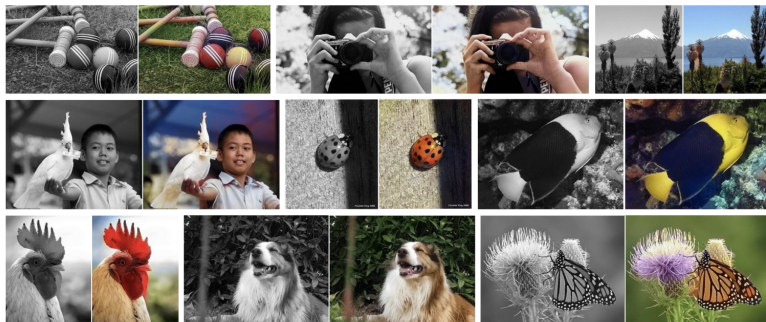


Figure 4.6: Input grayscale photos and output colorizations. Taken from [9].

4.2 Contrastive Learning

Recently, self-supervised methods based on *Contrastive Learning* [59] have been appealing to researchers due to their outstanding performance.

Contrastive Learning is a method that learns to map similar samples, referred to as *positive pairs*, to nearby points in a lower dimensional space and simultaneously pushes dissimilar samples apart, referred to as *negative pairs*. Contrastive learning aims at "learning by comparison" [60]. Let \mathbf{x}_{pos} and \mathbf{x}_{neg} be a positive and a negative sample for an anchor sample \mathbf{x} , contrastive methods aim to learn an encoder f such

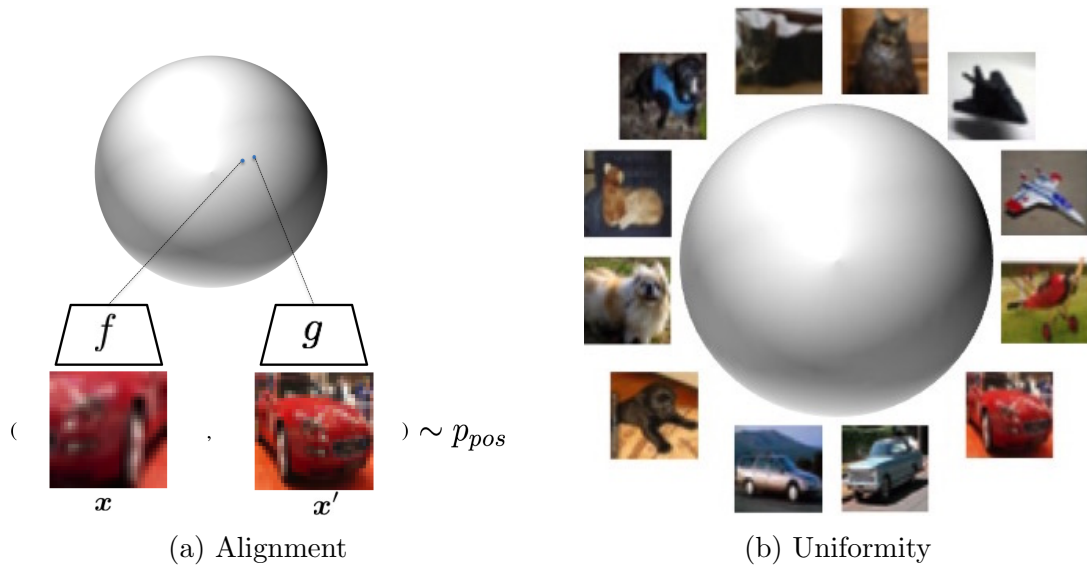


Figure 4.7: (a) Similar examples are mapped to nearby latent features. (b) Uniformity of features distribution on a unit hypersphere. Figures inspired by [10]. CIFAR10 [11] images are used for this demonstration.

that

$$s(f(\mathbf{x}), f(\mathbf{x}_{pos})) > s(f(\mathbf{x}), f(\mathbf{x}_{neg})) \quad (4.1)$$

where s is a score function.

Recently the most competitive methods for self-supervised representation learning have been contrastive. Among them, the state-of-the-art contrastive methods [17, 19, 61–63] learn representations by pulling together different augmented views of the same image and spreading augmented views of different images apart, see Figure 4.7.

The most commonly used self-supervised contrastive loss is Normalized Temperature-scaled Cross-entropy (NT-Xent) introduced in [17] defined as

$$\mathcal{L}_{self} = \sum_i \mathcal{L}_{self}^i = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(f(\mathbf{x}_i), g(\mathbf{x}'_i))/\tau}}{\frac{1}{N} \sum_{j=1}^N e^{s(f(\mathbf{x}_i), g(\mathbf{x}'_j))/\tau}} \quad (4.2)$$

where $0 < \tau < 1$ is a scalar temperature and the summation is over N independent samples $\{(\mathbf{x}_i, \mathbf{x}'_i)\}_{i=1}^N$ drawn from the joint distribution $p_{pos}(\mathbf{x}, \mathbf{x}')$ while $(\mathbf{x}_i, \mathbf{x}'_j)_{i \neq j}$ drawn from $p_{neg}(\mathbf{x}, \mathbf{x}') = p_{pos}(\mathbf{x})p_{pos}(\mathbf{x}')$, where $p_{pos}(\mathbf{x})$ is the marginal of $p_{pos}(\mathbf{x}, \mathbf{x}')$. Intuitively, this loss is the log loss of a N -way softmax-based classifier that tries to classify \mathbf{x}'_i as a positive sample for \mathbf{x}_i . Function $f(\cdot)$ is a feature extractor which maps the images from the input space to a lower dimensional space. The networks g and f can be identical [3, 64], partially shared [17, 18], or different [62]. It can be proved that minimizing NT-Xent loss maximizes a lower bound on

the MI, $I(X; X')$ [24].

An optimal critic for Eq. 4.2 is $s^*(f(\mathbf{x}), g(\mathbf{x}')) = \log p(g(\mathbf{x}')|f(\mathbf{x}))$ [24]. Common choices for critic function s are bilinear critics $s(f(\mathbf{x}), g(\mathbf{x}')) = f(\mathbf{x})^T \mathbf{W} g(\mathbf{x}')$ [17, 62, 65], separable critics $s(f(\mathbf{x}), g(\mathbf{x}')) = \phi_1(f(\mathbf{x}))^T \phi_2(g(\mathbf{x}'))$ [66], and concatenated critics $s(f(\mathbf{x}), g(\mathbf{x}')) = \phi([f(\mathbf{x}), g(\mathbf{x}')])$ [18], here ϕ_1 , ϕ_2 , and ϕ are typically shallow multilayer perceptrons (MLPs).

Tschannen et al. [67] empirically demonstrated that simple critics resulting in loose bounds on the MI can lead to better representations than high-capacity critics. They provide empirical evidence that the success of the contrastive methods is only loosely connected to the MI, instead, it should be more attributed to encoder architecture and the negative sampling strategy [68].

The most commonly used critic function is the cosine similarity which is used to measure the similarity between the features. The cosine similarity is a measure independent of magnitude and can be seen as the l_2 normalized inner product of two vectors. In [69], the necessity of normalization when using feature vectors dot product in a cross entropy loss has been argued and analysed.

4.2.1 Key Properties of Contrastive Loss

In [10], two key properties of contrastive loss *Alignment* and *Uniformity* have been analyzed, see Figure 4.7. Contrastive loss encourages the positive features to be aligned and the embeddings to match a uniform distribution on a unit hypersphere. Alignment favours encoders to map similar samples to nearby latent features and uniformity of (normalized) features distribution preserves maximal information leading separable features. Wang et al. [70] found that the contrastive loss meets a *uniformity-tolerance* dilemma caused by the inherent defect of self-supervised contrastive loss which pushes all different samples ignoring their semantic relations. The contrastive loss doesn't have any constraint on the distribution of negative samples as the objective learns instance discriminative embedding. They showed that an extreme pursuit to the uniformity makes the contrastive loss not tolerant to semantically similar samples which might be destructive for the structure of features useful for the downstream tasks.

Uniformity-tolerance Dilemma. Wang et al. [70] found that the contrastive loss is a hardness-aware loss function. They showed that the contrastive loss potentially focuses on optimizing the negative samples which are penalized according to their hardness controlled by the temperature τ . For small temperatures, the contrastive loss tends to penalize hardest negative samples while they are likely to share the similar semantic content with the anchor sample. As a result, the contrastive loss separates the positive samples close to the anchor sample making the local distribution sparse and the embedding distribution more uniform. For large temperatures, they showed that model tends to be more tolerant to the semantically consistent samples causing embeddings locally clustered and globally separated. On the one

hand, increasing the uniformity of embeddings distribution is achieved by decreasing the temperature, on the other hand making the embedding space tolerant to the similar samples is obtained by increasing the temperature. They argued that a good choice of temperature τ can compromise the two properties.

Another property of contrastive loss is *its intrinsic ability to perform hard positive and negative mining* investigated in [71]. Khosla et al. [71] have analytically shown that the gradient of contrastive loss encourages learning from hard positives and hard negatives. The contrastive loss induces a gradient structure such that gradients with respect to unnormalized representations lead to intrinsic hard positive and negative mining during training and thus there is no need for complicated hard mining algorithms for good performance [72], as hard positives and negatives boost the learning [73, 74]. They demonstrated this property for supervised version of contrastive loss that self-supervised contrastive loss can be a special case.

Another property of the contrastive loss is its concentration on predictable features rather than noise. In the following, we will show unlike maximum likelihood the contrastive loss doesn't incorporate noise.

Let $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ be a set of N independent, identically distributed (iid) training samples drawn from $p_{data}(\mathbf{x}, y)$. Let \mathbf{x} be a d dimensional feature vector, $\mathbf{x}_n \in \mathbb{R}^d$, and y be a K -way categorical random variable. Suppose we are interested in predicting label y given the feature vector \mathbf{x} . Let $p_{model}(y|\mathbf{x}, \boldsymbol{\theta})$ parameterized by a neural network be an estimate of the conditional probability distribution $p_{data}(y|\mathbf{x})$. In order to find parameters $\boldsymbol{\theta}$, we maximize the conditional likelihood of the training data

$$\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^N \log p_{model}(y_n|\mathbf{x}_n, \boldsymbol{\theta})$$

for a large data set, as $N \rightarrow \infty$

$$\boldsymbol{\theta}^* = - \arg \min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x}, y) \sim p_{data}} [\log p_{model}(y|\mathbf{x}, \boldsymbol{\theta})] \quad (4.3)$$

if a sample \mathbf{x} is correctly classified by the model, $-\log p_{model}(y|\mathbf{x}) = 0$ which implies that a correct prediction has a low contribution to the loss while a random prediction has a high contribution, $-\log p_{model}(y|\mathbf{x}) = \log K$ for a balanced dataset. As shown, the loss is determined by what can not be predicted, in other words the maximum likelihood estimation ignores the information can be predicted.

Let us take the contrastive loss to do the prediction task. Eq. 4.4, is a categorical cross entropy loss of classifying the sample \mathbf{x} correctly

$$\boldsymbol{\theta}^* = - \arg \min_{\boldsymbol{\theta}} \sum_{n=1}^N \log \frac{Q(\mathbf{x}_n, y_n)}{\sum_{i=1}^K p_{data}(y_i) Q(\mathbf{x}_n, y_i)} \quad (4.4)$$

as $N \rightarrow \infty$

$$\boldsymbol{\theta}^* = - \arg \min_{\boldsymbol{\theta}} \mathbb{E}_{(\mathbf{x}, y) \sim p_{data}} \left[\log \frac{Q(\mathbf{x}, y)}{\mathbb{E}_{y' \sim p_{data}(y)} Q(\mathbf{x}, y')} \right] \quad (4.5)$$

for a balanced dataset $p_{data}(y) = \frac{1}{K}$. For a correctly classified sample \mathbf{x} , $Q(\mathbf{x}, y)$ takes high values and the contribution of $\{Q(\mathbf{x}, y')\}_{y' \neq y}$ can be ignored to the summation in the denominator. Contribution of a correctly classified sample is $\log K$ while for a random prediction we have $Q(\mathbf{x}, y) \simeq Q(\mathbf{x}, y')$ and it implies a low contribution to the contrastive loss, i.e., $\log \frac{Q}{K \cdot \frac{1}{K} Q} = \log 1 = 0$. It simply means that the contrastive loss is less impacted by what can not be predicted such as noise.

4.2.2 Different Contrastive Loss Mechanisms

Contrastive learning benefits from large batch sizes as larger batch sizes provide more negative samples, accelerating convergence [3]. In *SimCLR* [3], negative samples are taken from a large mini-batch and the mini-batch size is limited by the GPU memory size. A mechanism to maintain negative samples is using a memory bank proposed in [75]. Features for all instances in the dataset are stored in the memory bank and the representation of each sample is updated by a momentum. The momentum update is on the representation of each sample, not the encoder. *Moco* [19] instead maintain negative samples in a queue. The encoded representations of the current mini-batch are enqueued while the oldest are dequeued which keeps the queue as consistent as possible. *Moco* is more memory-efficient, can be trained on large-scale data, and dissociate the batch size from the number of negative samples.

Chapter 5

Self-Supervised Representation Learning in Medical Time Series

Atrial Fibrillation (AF) is the most common arrhythmia and has a major impact on morbidity and mortality; however, detection of asymptomatic AF is challenging. We aim to evaluate the sensitivity and specificity of non-invasive AF detection by a medical wearable. We apply different algorithms to five-minute periods of inter-beat intervals (IBI) for the AF detection. A DNN is trained unsupervised to extract relevant features for AF detection. The training objective is given by maximising the MI between IBI values that are separated by a randomly chosen time point within the five-minute period. Unsupervised feature extraction followed by an unsupervised classification results in higher sensitivity and specificity compared with normalised root mean square of the successive difference (nRMSSD) an established metric for the AF detection.

5.1 Method

To extract relevant features for AF detection we use Contrastive Predictive Coding (CPC) [17], an unsupervised objective, which learns predictable representations. CPC is a general technique that only requires observations to be ordered along, e.g., temporal or spatial dimensions and we can apply it to a variety of different modalities including audio, natural language, and images [65].

An encoder and an autoregressive model are jointly trained to learn generalizable representations of high dimensional data by predicting the representations of future observations from those of past ones [17]. For simplicity, we adopt the same notation as [17]. Let \mathbf{x}_t be an input sequence. An encoder, g_{enc} , parameterized by a neural network nonlinearly maps the input sequence \mathbf{x}_t into a latent space, $\mathbf{z}_t = g_{enc}(\mathbf{x}_t)$. Afterwards, an autoregressive model, g_{ar} , sums up all the information in the la-

tent space for $\mathbf{z}_{\leq t}$ and generates a context latent representation, $\mathbf{c}_t = g_{ar}(\mathbf{z}_{\leq t})$. The k^{th} future feature vector, \mathbf{z}_{t+k} , is predicted by weighted linear combination of context feature vectors, i.e., $\hat{\mathbf{z}}_{t+k} = \mathbf{W}_k \mathbf{c}_t$, with a different prediction matrix \mathbf{W}_k for each step k . The quality of the prediction is assessed by mutual information between \mathbf{z}_{t+k} and \mathbf{c}_t which is modeled by a density ratio proposed in [17] as follows

$$f_k(\mathbf{x}_{t+k}, \mathbf{c}_t) \propto \frac{p(\mathbf{x}_{t+k} | \mathbf{c}_t)}{p(\mathbf{x}_{t+k})} \quad (5.1)$$

the density ratio f can be unnormalized and any positive real score can be used to model f . Same as [17] we choose a log-bilinear model

$$f_k(\mathbf{x}_{t+k}, \mathbf{c}_t) = e^{\mathbf{z}_{t+k}^T \mathbf{W}_k \mathbf{c}_t} \quad (5.2)$$

Both the encoder and autoregressive models are jointly trained to optimize the InfoNCE loss proposed in [17], Eq. 5.3, inspired by Noise-Contrastive Estimation (NCE) [76] loss

$$\mathcal{L}_{NCE} = -\mathbb{E}_{\mathcal{X}} \left[\log \frac{f_k(\mathbf{x}_{t+k}, \mathbf{c}_t)}{\sum_{\mathbf{x}_j \in \mathcal{X}} f_k(\mathbf{x}_j, \mathbf{c}_t)} \right] \quad (5.3)$$

where $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ is a set of N iid random samples containing one positive sample drawn from $p(\mathbf{x}_{t+k} | \mathbf{c}_t)$ and $N - 1$ negative samples drawn from $p(\mathbf{x}_{t+k})$. Rather than sampling negative samples explicitly, given a positive sample we take the rest samples within a mini-batch as negative samples. Intuitively, this loss is the log loss of a N -way softmax-based classifier that tries to classify \mathbf{x}_{t+k} as a positive sample for \mathbf{c}_t . It can be shown that the optimal value for $f_k(\mathbf{x}_{t+k}, \mathbf{c}_t)$ is proportional to $\frac{p(\mathbf{x}_i | \mathbf{c}_t)}{p(\mathbf{x}_i)}$ [17]. Let N samples in \mathcal{X} be iid and \mathbf{x}_i be a positive example for \mathbf{c}_t so $\{\mathbf{x}_l\}_{l \neq i}$ are independent from \mathbf{c}_t , i.e., $p(\mathbf{x}_l | \mathbf{c}_t) = p(\mathbf{x}_l)$ for $l \neq i$

$$p(\mathbf{x}_1, \dots, \mathbf{x}_N | \mathbf{c}_t) = \frac{p(\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{c}_t)}{\sum_{\mathbf{x}_1, \dots, \mathbf{x}_N} p(\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{c}_t)} \quad (5.4)$$

$$= \frac{p(\mathbf{x}_i | \mathbf{c}_t) \prod_{l \neq i} p(\mathbf{x}_l)}{\sum_{j=1}^N p(\mathbf{x}_j | \mathbf{c}_t) \prod_{l \neq j} p(\mathbf{x}_l)} \quad (5.5)$$

$$= \frac{p(\mathbf{x}_i | \mathbf{c}_t) \frac{p(\mathbf{x}_i)}{p(\mathbf{x}_i)} \prod_{l \neq i} p(\mathbf{x}_l)}{\sum_{j=1}^N p(\mathbf{x}_j | \mathbf{c}_t) \frac{p(\mathbf{x}_j)}{p(\mathbf{x}_j)} \prod_{l \neq j} p(\mathbf{x}_l)} \quad (5.6)$$

$$= \frac{\frac{p(\mathbf{x}_i | \mathbf{c}_t)}{p(\mathbf{x}_i)} \prod_l p(\mathbf{x}_l)}{\sum_{j=1}^N \frac{p(\mathbf{x}_j | \mathbf{c}_t)}{p(\mathbf{x}_j)} \prod_l p(\mathbf{x}_l)} \quad (5.7)$$

$$= \frac{\frac{p(\mathbf{x}_i | \mathbf{c}_t)}{p(\mathbf{x}_i)}}{\sum_{j=1}^N \frac{p(\mathbf{x}_j | \mathbf{c}_t)}{p(\mathbf{x}_j)}} \quad (5.8)$$

In the following, we demystify the proof in [17] that minimizing Eq. 5.3 maximizes a lower bound on the MI between \mathbf{z}_{t+k} and \mathbf{c}_t . In Appendix A.1, we provide an alternative proof.

From the data processing inequality, $I(\mathbf{z}_{t+k}; \mathbf{c}_t) \geq I(\mathbf{x}_{t+k}; \mathbf{c}_t)$ induced by the Markov chain, $\mathbf{x}_{t+k} \leftarrow \mathbf{z}_{t+k} \leftarrow \mathbf{c}_t$. Maximizing $I(\mathbf{x}_{t+k}; \mathbf{c}_t)$ maximizes a lower bound on $I(\mathbf{z}_{t+k}; \mathbf{c}_t)$. Substituting $f(\mathbf{x}_{t+k}, \mathbf{c}_t)$ with its optimal value in Eq. 5.3 and splitting \mathcal{X} to the positive sample and negative samples, \mathcal{X}_{neg} ,

$$\mathcal{L}_{NCE}^{opt} = -\mathbb{E}_{\mathcal{X}} \left[\log \frac{\frac{p(\mathbf{x}_{t+k}|\mathbf{c}_t)}{p(\mathbf{x}_{t+k})}}{\frac{p(\mathbf{x}_{t+k}|\mathbf{c}_t)}{p(\mathbf{x}_{t+k})} + \sum_{\mathbf{x}_j \in \mathcal{X}_{neg}} \frac{p(\mathbf{x}_j|\mathbf{c}_t)}{p(\mathbf{x}_j)}}} \right] \quad (5.9)$$

$$= \mathbb{E}_{\mathcal{X}} \log \left[1 + \frac{p(\mathbf{x}_{t+k})}{p(\mathbf{x}_{t+k}|\mathbf{c}_t)} (N-1) \frac{1}{(N-1)} \sum_{\mathbf{x}_j \in \mathcal{X}_{neg}} \frac{p(\mathbf{x}_j|\mathbf{c}_t)}{p(\mathbf{x}_j)} \right] \quad (5.10)$$

since \mathbf{x}_j is a negative example, it is ideally independent of context \mathbf{c}_t , i.e., $p(\mathbf{x}_j|\mathbf{c}_t) = p(\mathbf{x}_j)$, so $\sum_{\mathbf{x}_j \in \mathcal{X}_{neg}} \frac{p(\mathbf{x}_j|\mathbf{c}_t)}{p(\mathbf{x}_j)} = N-1$

$$\mathcal{L}_{NCE}^{opt} \approx \mathbb{E}_{\mathcal{X}} \log \left[1 + \frac{p(\mathbf{x}_{t+k})}{p(\mathbf{x}_{t+k}|\mathbf{c}_t)} (N-1) \right] \quad (5.11)$$

$$= \mathbb{E}_{\mathcal{X}} \log \left[1 - \frac{p(\mathbf{x}_{t+k})}{p(\mathbf{x}_{t+k}|\mathbf{c}_t)} + \frac{p(\mathbf{x}_{t+k})}{p(\mathbf{x}_{t+k}|\mathbf{c}_t)} N \right] \quad (5.12)$$

sample \mathbf{x}_{t+k} was drawn from the conditional distribution $p(\mathbf{x}_{t+k}|\mathbf{c}_t)$ rather than the marginal distribution $p(\mathbf{x}_{t+k})$, so $1 - \frac{p(\mathbf{x}_{t+k})}{p(\mathbf{x}_{t+k}|\mathbf{c}_t)}$ is positive

$$\mathcal{L}_{NCE}^{opt} \geq \mathbb{E}_{\mathcal{X}} \log \left[\frac{p(\mathbf{x}_{t+k})}{p(\mathbf{x}_{t+k}|\mathbf{c}_t)} N \right] \quad (5.13)$$

$$= \mathbb{E}_{\mathcal{X}} \log \left[\frac{p(\mathbf{x}_{t+k})}{p(\mathbf{x}_{t+k}|\mathbf{c}_t)} \right] + \log N \quad (5.14)$$

$$= \sum_{\mathbf{x}_{t+k}} \sum_{\mathbf{c}_t} p(\mathbf{x}_{t+k}, \mathbf{c}_t) \log \left[\frac{p(\mathbf{x}_{t+k})}{p(\mathbf{x}_{t+k}|\mathbf{c}_t)} \right] + \log N \quad (5.15)$$

substituting $p(\mathbf{x}_{t+k}|\mathbf{c}_t)$ by $\frac{p(\mathbf{x}_{t+k}, \mathbf{c}_t)}{p(\mathbf{c}_t)}$

$$\mathcal{L}_{NCE}^{opt} \geq \sum_{\mathbf{x}_{t+k}} \sum_{\mathbf{c}_t} p(\mathbf{x}_{t+k}, \mathbf{c}_t) \log \left[\frac{p(\mathbf{x}_{t+k})}{\frac{p(\mathbf{x}_{t+k}, \mathbf{c}_t)}{p(\mathbf{c}_t)}} \right] + \log N \quad (5.16)$$

$$= \sum_{\mathbf{x}_{t+k}} \sum_{\mathbf{c}_t} p(\mathbf{x}_{t+k}, \mathbf{c}_t) \log \left[\frac{p(\mathbf{x}_{t+k})p(\mathbf{c}_t)}{p(\mathbf{x}_{t+k}, \mathbf{c}_t)} \right] + \log N \quad (5.17)$$

$$= - \sum_{\mathbf{x}_{t+k}} \sum_{\mathbf{c}_t} p(\mathbf{x}_{t+k}, \mathbf{c}_t) \log \left[\frac{p(\mathbf{x}_{t+k}, \mathbf{c}_t)}{p(\mathbf{x}_{t+k})p(\mathbf{c}_t)} \right] + \log N \quad (5.18)$$

$$I(\mathbf{x}_{t+k}, \mathbf{c}_t) \geq \log N - \mathcal{L}_{NCE}^{opt} \quad (5.19)$$

5.2 Data Preparation

We split recordings of IBI values into five-minute periods, which are manually classified into AF and non-AF. Then, we extract periods that have at least 80% reliable IBI values (quality scores of IBI values ≤ 13 , with 1 indicating best quality and 16 worst quality). IBI values are encoded together with the respective quality indices provided by the wearable device into a multi-dimensional vector room, where IBI values with different quality scores are orthogonal to each other. As the dataset contains more non-AF than AF periods, the latter is oversampled to achieve a more balanced dataset. We take all five-minute periods that contain at least 200 IBI values.

5.3 Training Details

Experiments were carried out using a strided convolutional neural network. We use two convolutional layers with strides [1,1], filter sizes [3,3] and 64 hidden units with ReLU activations. We take recurrent neural network, gated recurrent unit (GRU) [77], for the autoregressive model with a 32 dimensional hidden state. We train on sampled windows of length 200. We use AMSGrad optimiser [78] with initial learning rate of 10^{-3} and weight decay of 10^{-4} . We train at batch size 64 for 100 epochs.

To understand the representations extracted by CPC we train a classifier, a fully connected network with two hidden layers, on top of these features. We extract the outputs of the autoregressive model (32 dimensional) as input and train the classifier to predict the labels from the manual classification.

5.4 Evaluation

To evaluate the results, we use Area Under the Receiver Operating Characteristics (AUROC) which has the advantage to be scale-invariant, measures how well predictions are ranked rather than their absolute values and classification-threshold-invariant, it measures how well AF samples are separated from non-AF samples. Because of class imbalance between AF and non-AF in test set we also report values for sensitivity and specificity.

AUROC. AUROC curve shows the trade-off between true positive rate (TPR) and false positive rate (FPR) across different decision thresholds.

Sensitivity (also known as recall, hit rate, or true positive rate). It is the probability that a positive example is correctly identified.

Specificity (also known as selectivity or true negative rate). It is the probability that a negative example is correctly identified.

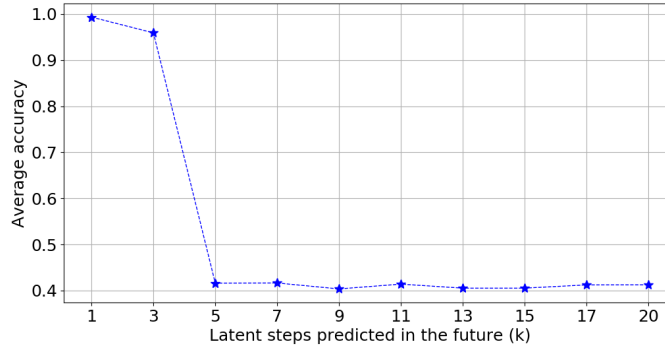


Figure 5.1: Average accuracy of predicting positive samples in the contrastive loss

5.5 Effect of Number of Predicted Latent Steps

To investigate the effect of the number of predicted latent steps, k , in the future, we train the model to predict latents for different timesteps. In Figure 5.1, we report the average number of times that the model correctly classifies ‘future’ representations among a set of unrelated ‘negative’ representations. As expected, the prediction task becomes harder when the target is further away. If the target is easy to predict from the context (e.g., when predicting a single step in the future and the target overlaps with the context) the performance of the model degrades, discouraging the model to further improve the representations. Figure 5.2 shows necessity of learning a good set of representations in order to discriminate two classes AF and non-AF.

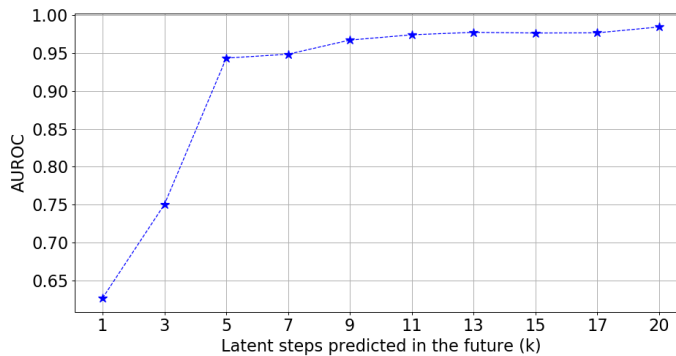


Figure 5.2: Evaluating 1-NN classifier performance for the different predicted latent steps in the future.

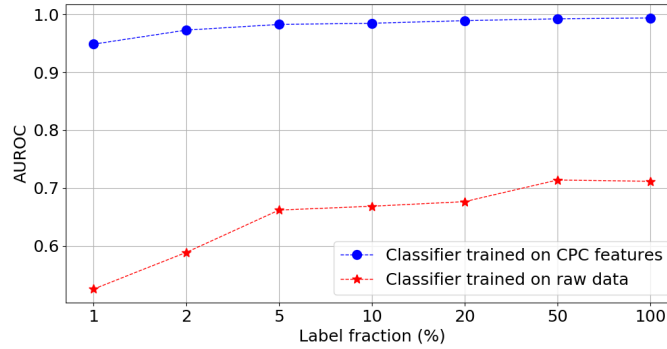


Figure 5.3: AUROC for the classification task on the CPC features and on the raw data under varied sizes of label fractions

5.6 Self-Supervised models are more data-efficient

To investigate the CPC representations enable generalization from few labels, we train two classifiers on the CPC features extracted by the autoregressive model and on the raw data for different fractions of labeled training data. We investigated using 1%, 2%, 5%, 10%, 20%, 50%, and 100% of the dataset. Figure 5.3 shows how the performance varies for different available label fractions. We observe that self-supervised model can significantly help with label efficiency and generalization for the classification task. The results suggest that fine-tuning with a few labeled examples yields significantly higher gain compared with the supervised baseline.

5.7 Reliable Detection of Atrial Fibrillation with a Medical Wearable during Inpatient Conditions





Jacobsen M, Dembek TA, Ziakos AP, Gholamipoor R, Kobbe G, Kollmann M, Blum C, Müller-Wieland D, Napp A, Heinemann L, Deubner N, Marx N, Isenmann S, Seyfarth M. Reliable Detection of Atrial Fibrillation with a Medical Wearable during Inpatient Conditions. *Sensors*, 2020.

Status: Published.

Contributions: The author contributed with designing and implementation of deep neural network, training, evaluation, and visualization. The author contributed with writing parts related to DNN-based algorithm under the supervision of Prof. Dr. Markus Kollmann.

Article

Reliable Detection of Atrial Fibrillation with a Medical Wearable during Inpatient Conditions

Malte Jacobsen ^{1,2,*} , Till A. Dembek ³ , Athanasios-Panagiotis Ziakos ^{1,4} ,
Rahil Gholamipoor ⁵, Guido Kobbe ⁶, Markus Kollmann ⁷, Christopher Blum ⁷,
Dirk Müller-Wieland ², Andreas Napp ², Lutz Heinemann ⁸, Nikolas Deubner ^{1,4} ,
Nikolaus Marx ², Stefan Isenmann ^{1,9} and Melchior Seyfarth ^{1,4}

- ¹ Faculty of Health, University Witten/Herdecke, 58448 Witten, Germany; athanasios-panagiotis.ziakos@helios-gesundheit.de (A.-P.Z.); nikolas.deubner@helios-gesundheit.de (N.D.); stefan.isenmann@st-josef-moers.de (S.I.); melchior.seyfarth@helios-gesundheit.de (M.S.)
 - ² Department of Internal Medicine I, University Hospital Aachen, RWTH Aachen University, 52074 Aachen, Germany; dirmueller@ukaachen.de (D.M.-W.); anapp@ukaachen.de (A.N.); nmarx@ukaachen.de (N.M.)
 - ³ Department of Neurology, Faculty of Medicine, University of Cologne, 50937 Cologne, Germany; till.dembek@uk-koeln.de
 - ⁴ Department of Cardiology, Helios University Hospital of Wuppertal, 42117 Wuppertal, Germany
 - ⁵ Department of Computer Science, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany; rahil.gholamipoorfard@hhu.de
 - ⁶ Department of Hematology, Oncology, and Clinical Immunology, University Hospital Düsseldorf, Medical Faculty, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany; kobbe@med.uni-duesseldorf.de
 - ⁷ Department of Biology, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany; markus.kollmann@hhu.de (M.K.); christopher.blum@hhu.de (C.B.)
 - ⁸ Science-Consulting in Diabetes, 41462 Neuss, Germany; l.heinemann@science-co.com
 - ⁹ Department of Neurology, St. Josef Hospital, 47441 Moers, Germany
- * Correspondence: mjacobsen@ukaachen.de; Tel.: +49-173-560-6980

Received: 31 August 2020; Accepted: 24 September 2020; Published: 26 September 2020



Abstract: Atrial fibrillation (AF) is the most common arrhythmia and has a major impact on morbidity and mortality; however, detection of asymptomatic AF is challenging. This study aims to evaluate the sensitivity and specificity of non-invasive AF detection by a medical wearable. In this observational trial, patients with AF admitted to a hospital carried the wearable and an ECG Holter (control) in parallel over a period of 24 h, while not in a physically restricted condition. The wearable with a tight-fit upper armband employs a photoplethysmography technology to determine pulse rates and inter-beat intervals. Different algorithms (including a deep neural network) were applied to five-minute periods photoplethysmography datasets for the detection of AF. A total of 2306 h of parallel recording time could be obtained in 102 patients; 1781 h (77.2%) were automatically interpretable by an algorithm. Sensitivity to detect AF was 95.2% and specificity 92.5% (area under the receiver operating characteristics curve (AUC) 0.97). Usage of deep neural network improved the sensitivity of AF detection by 0.8% (96.0%) and specificity by 6.5% (99.0%) (AUC 0.98). Detection of AF by means of a wearable is feasible in hospitalized but physically active patients. Employing a deep neural network enables reliable and continuous monitoring of AF.

Keywords: clinical trial; wearable sensors; atrial fibrillation; photoplethysmography; deep neural network

1. Introduction

Atrial fibrillation (AF) is the most common arrhythmia with rising incidence and prevalence [1,2]; the current prevalence is estimated to be between 2% to 4% [3]. AF is more common in males and shows an increasing prevalence with age [4]. There are a number of modifiable known risk factors for AF, including obesity, hypertension, diabetes mellitus, and smoking, as possible contributors to the development and progression of AF [5].

AF is associated with a broad spectrum of clinical events, including ischemic stroke. The proportion of time in AF associated with a significant risk for complications is unknown, thus requiring further evaluation [6]. Due to the paroxysmal and often asymptomatic occurrence of AF, ECG Holter monitoring is frequently employed to detect episodes of silent AF [7]. However, ECG Holter monitoring has limitations: Carrying an ECG Holter limits patients in their daily activities and restricts monitoring to relatively short periods of time. Additionally, ECG Holters are prone to movement artifacts, and thus, not reliable during phases of physical activity [8].

Wearables that are used as medical devices (defined as having a regulatory approval like a *Conformité Européenne* (CE) mark for Europe) offer an affordable non-invasive screening option for AF [9–13]. Photoplethysmography (PPG) is frequently employed in such wearables [14]. It is an optical method to measure volume changes in the tissue. PPG is used to calculate clinically relevant parameters, e.g., heart rate, inter-beat intervals (IBI—the interval between two pulse waves in milliseconds) [15]. Intervals between heartbeats are a parameter often used for the detection of AF. PPG derived IBI show a high correlation to the ECG derived heart rate intervals (gold-standard) [16]. Technologies employed in wearables and evaluated for the detection of AF are most often based on single-lead ECG or PPG and can be separated into active and passive approaches: Active monitoring requires that the patient initialized a recording, e.g., individuals have to place their fingers on the electrodes of a smartphone like device. In contrast, wearables with a passive monitoring approach do not require patient intervention. With this approach, measurements are performed continuously or semi-continuously (e.g., every 5 min). In a previous clinical trial with an active approach, wearable detection of AF was possible with a sensitivity of 91.5% and specificity of 99.6% [13]. In clinical trials with a passive approach, equivalent results were shown in patients where physical activity was restricted while recording. However, there was a risk of missing asymptomatic episodes of AF. When such wearables are used for ECG recordings, usage of adhesives or bandages is needed, and there are limitations regarding diagnostic adherence [9]. In the Huawei Heart Study, more than one-third of individuals with suspected AF were primarily detected with a periodical passive PPG approach [12]. However, a recent trial using a passive approach showed that there is a gap in detecting AF under controlled and uncontrolled conditions, most likely due to periods of physical activity with an increase in heart rate and movement artifacts [17]. Some wearables under evaluation had varying sampling rates, with considerable risk of missing AF [10].

A novel upper arm medical wearable (Everion[®], Biovotion AG, Switzerland) employs a passive PPG approach, allowing reliable long-term, high-resolution data recording. This device records of patients' physical activities during recording and provides information about the proportion of the automatically interpretable time.

The aim of this study was to evaluate the performance of a medical wearable by means of employing a PPG technology for AF detection in patients with paroxysmal or persistent AF study during inpatient conditions.

2. Materials and Methods

This study was an open-label, single-arm, inpatient, single-center trial. The clinical investigation plan was approved by the Ethical Committee of the University Witten/Herdecke, Germany, and was registered in the German clinical trials register (DRKS00014821).

Patients were recruited consecutively at the Department of Cardiology, University Hospital Wuppertal between September and December 2018 (Figure 1).

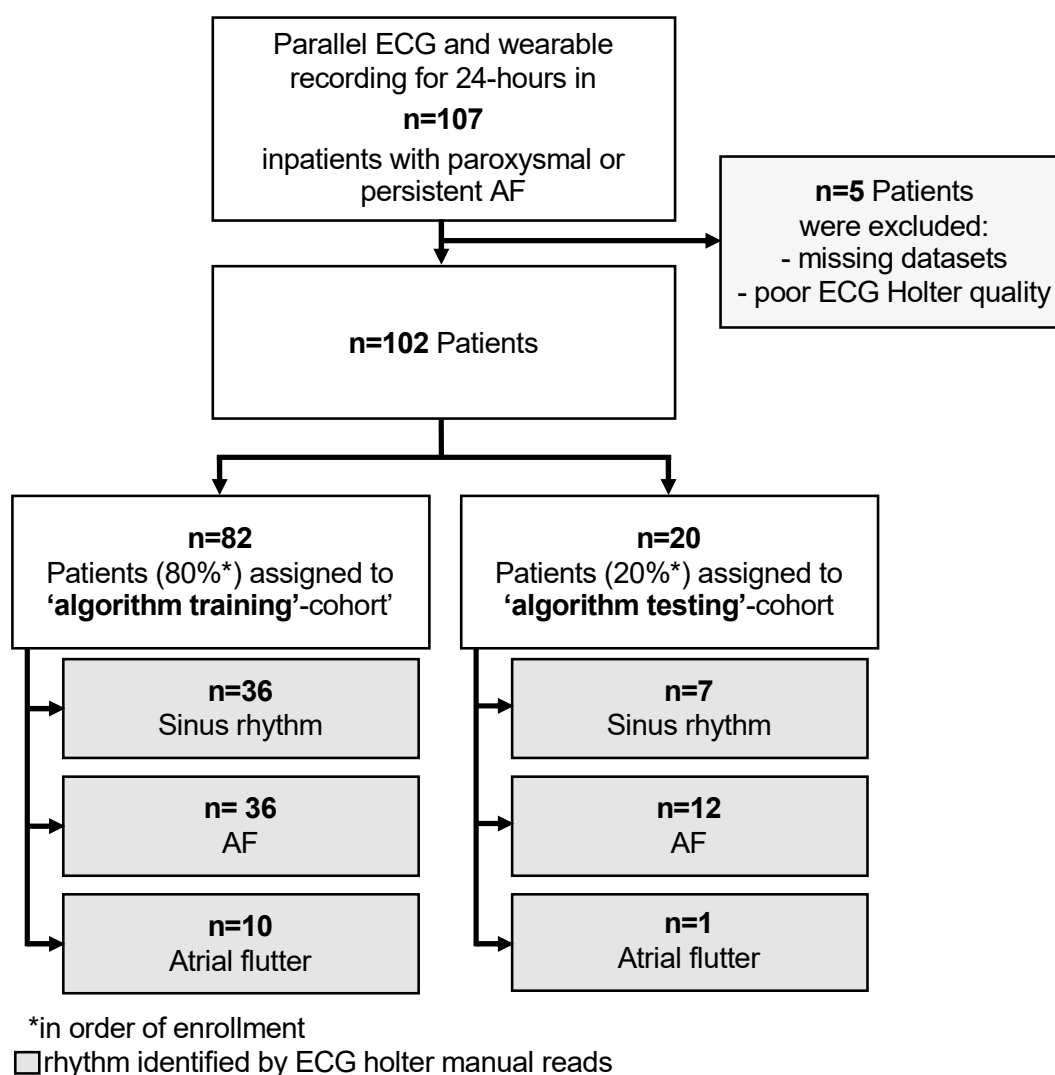


Figure 1. Flow-chart of patient disposition for algorithm development and group classification for the trial.

The primary outcome of this trial was the evaluation of sensitivity and specificity of non-invasive AF detection by a medical wearable at rest and during moderate physical activity. The secondary outcome was the determination of the proportion of recording time interpretable by algorithms.

All patients gave written informed consent prior to enrolment in this trial. Admitted patients with documented AF (e.g., prior to electrical cardioversion) or known paroxysmal AF were screened for eligibility for trial participation. Inclusion criteria were patients admitted for AF by their treating cardiologist and emergency room show ups with age ≥ 18 years and an indication for ECG Holter monitoring. Exclusion criteria were any cardiac implants or conditions which might impair measurements (e.g., upper arm tattoos, skin diseases).

Patients had no restrictions on their physical activity. At the end of the monitoring period, a safety assessment was performed. Patients answered a short questionnaire at the study end to evaluate wearable usage (discomfort, pain, sense of safety, design, willingness to perform inpatient, and outpatient monitoring).

In line with the standard of care in the hospital, patients carried a three-lead ECG Holter (Lifecard CF, Spacelabs Healthcare GmbH, Germany) for detection of AF over 24 h. ECG Holter data were reviewed for atrial arrhythmias by two cardiologists independently using a standard of care software tool (Sentinel 10, Spacelabs Healthcare, Snoqualmie, WA, USA). In the case of differing diagnoses, a third

cardiologist was consulted. Heart rhythm was classified into either sinus rhythm, AF, or atrial flutter, and this classification served as the gold standard for further analysis (Figure 2). ECG datasets were discarded if more than 50% of recorded data was not interpretable as defined by our independent raters.

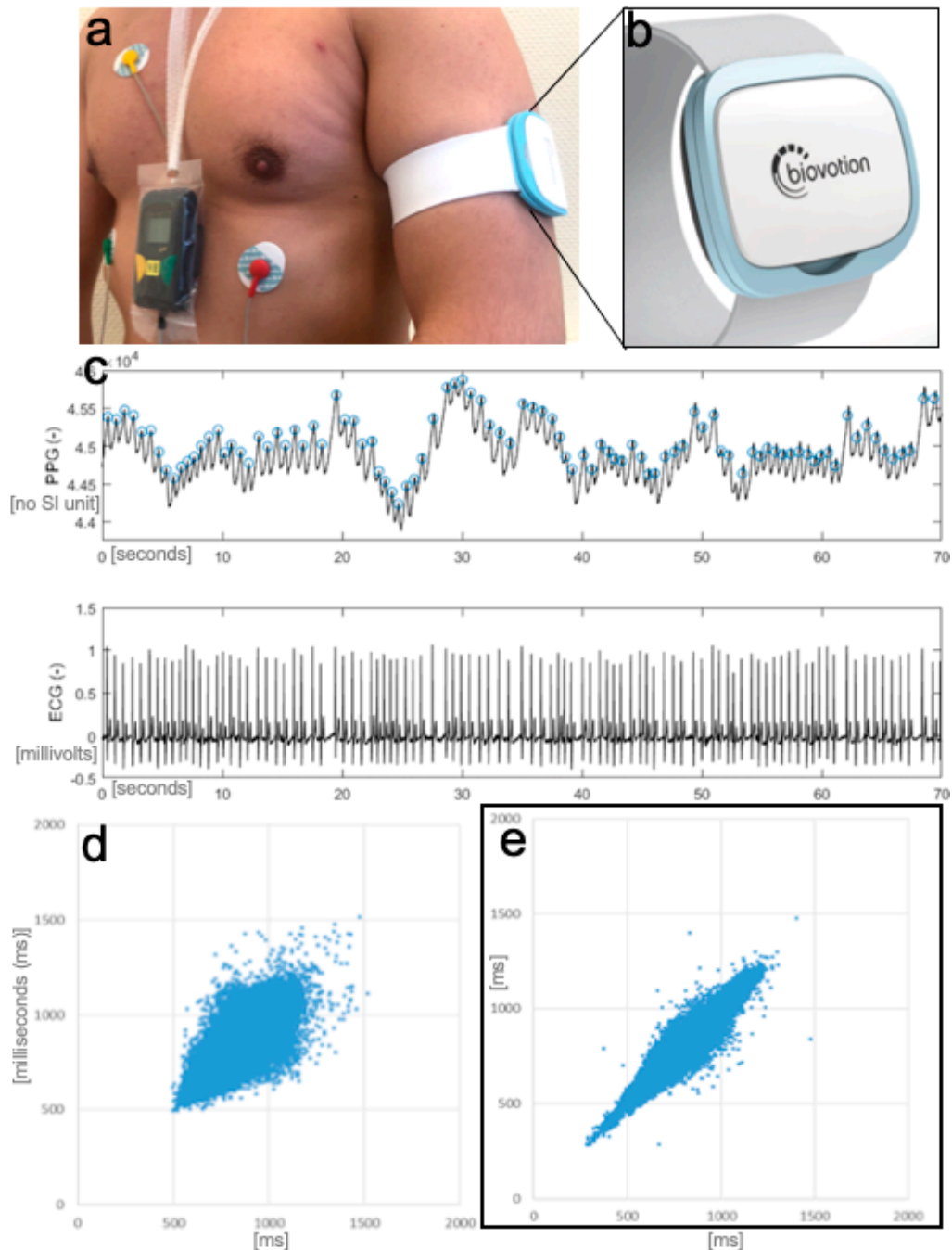


Figure 2. Recordings setup with the medical wearable attached to the left upper arm and ECG Holter (a); wearable (b); recorded signals (c) (first-row photoplethysmography (PPG), second-row ECG; showing an atrial fibrillation (AF) recording); Poincaré plot of PPG derived inter-beat intervals in AF (d) and sinus rhythm (e).

In parallel, a commercially available medical wearable (Everion, Biovotion AG, Switzerland) was worn by the patients. The wearable was attached to the preferred upper arm of the patients by the investigator. The time base of the wearable was synchronized to the ECG Holter. The wearable is a CE marked medium-risk device (class IIa), according to the Directive 93/42/EEC (firmware used was for clinical investigation only). It has different sensors for non-invasive monitoring of vital signs (e.g., PPG, accelerometry, gyroscope), memory storage of 16 MB Flash and a battery life of up to 32 h. Parameters, such as heart rate, IBI, the morphology of the pulse wave, and a physical activity index (based on the accelerometry data), are calculated using proprietary algorithms of the manufacturer implemented in the firmware. PPG-Signals were acquired with a sampling rate of 51.2 Hz. IBI were calculated permanently and stored approximately every 40 s. The device also provides recording quality indices for each data point. Data stored in the wearable were downloaded via a Bluetooth connection.

Two different approaches for detecting AF from the downloaded data were investigated: First, an established metric for AF detection, the normalized root mean square of successive differences (nRMSSD) of the IBIs, to differentiate between sinus rhythm and AF was used [11]. Second, a deep neural network (DNN) to detect episodes of AF was applied. Data with an insufficient quality based on the point-in-time accuracy estimate in the pre-processed data were excluded. Sufficient quality was defined when such an estimate for the IBI values could be calculated.

nRMSSD classification: Data was split into successive five-minute periods, and nRMSSD was calculated for all of these. For determining the optimal nRMSSD threshold, the dataset was split into a 'training cohort' consisting of the first 80% of the recruited patients and a 'testing cohort' consisting of the remaining 20%. Receiver operating characteristics (ROC) were calculated for five-minute periods in the 'training cohort', and the threshold with the highest Youden's J statistic was determined. This threshold was then applied to calculate the sensitivity and specificity of nRMSSD based AF detection in the 'testing cohort'. Algorithms presented were not trained to differentiate between AF and atrial flutter, only to discriminate AF.

Deep neural network classification: As data source, the same five-minute periods of IBI values were used as for the nRMSSD-model described above. As the dataset contained significantly more non-AF periods than AF periods, oversampling was performed by replicating the randomly selected samples to achieve a balanced dataset. The IBI values were encoded together with their associated quality scores into a multi-dimensional vector space, where IBI values with different quality scores are taken orthogonal to each other. A DNN was trained unsupervised on the dataset to extract the relevant features for AF detection. The training objective was given by maximizing the mutual information between IBI values that were separated by a randomly chosen time point within the five-minute period. The algorithmic details for computing of mutual information can be found in the appendix (see Appendix B) [18]. The unsupervised classification was carried out by one-nearest neighbor classification (Figure 3). Additionally, a second DNN (classifier) was trained on the extracted features from unsupervised learning using annotated data. The evaluation of the DNNs were carried out by randomly splitting the pre-processed data into the train (80%)/validation (10%)/test datasets (10%). Subsequently, sensitivity and specificity were calculated using ten-fold cross-validation. For testing, the unbalanced original data was used.

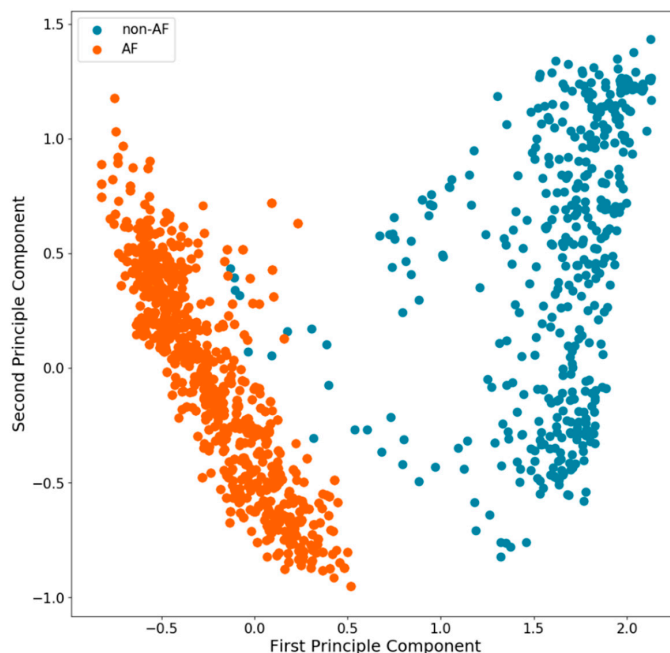


Figure 3. First two principle components of the latent space from the unsupervised Deep Learning approach for five-minute periods. Results of one-nearest neighbor classification for individual periods are shown that would be interpreted as AF (orange) or non-AF (blue).

A prerequisite for reliable detection of AF over time in clinical practice is sufficient data quality and that the recording time is maximal, e.g., a given patient might carry a wearable for 24 h; however, the proportion of recording time automatically interpretable by algorithms (=interpretable time) may be decisively less. [19,20] From the data obtained, the percentage of ‘good quality data’ was assessed by aggregating the time periods during which data were available that enabled an automatic IBI analysis. Others have used a cut-off value of 90% analyzable data for each five-minute period in resting patients; however, in order to apply a pragmatic approach in potentially active patients value of 80% was used for this trial. A threshold of $\geq 80\%$ of the interpretable time was considered to be sufficient for clinical monitoring. Logistic regression analysis was used to evaluate which factors have an impact on the analyzable time. To evaluate the success of wearable data recording, the total recording time, as well as total interpretable time (time with accepted quality indices), were calculated.

Due to the known effect of patients’ physical activity on the detection of AF, an activity index over time was calculated for each patient. Based on the activity classification provided by the wearable, any classification besides ‘resting’ was considered as physical activity (e.g., walking flat). From the activity data provided by the wearable subsequent five-minute periods were labeled as ‘active’ or ‘resting’. The activity index is expressed as a percentage of each hour of recording. It was analyzed if detection of AF was possible with the wearable used during periods with and without physical activity.

For accuracy testing of heart rate estimation by the wearable in patients with different underlying heart rhythms, in each patient, one hour of ECG recording with a low rate of artifacts was selected manually (see Appendix B). Accuracy evaluation was performed as described elsewhere [21]. For data analysis, a standard software tool was used (MATLAB R2018b; MathWorks, Natick, MA, USA). Statistical Analysis

The confidence interval was set to 95% for all statistical analyses. Non-parametric categorical distributed variables were tested with a 2-tailed Fishers exact test or Chi-Square test. Continuous variables were tested with the Mann-Whitney test. For analyses of variables that have an impact on interpretable time, logistic regression was performed. For the primary outcome of AF detection Receiver Operating Characteristics (ROC) analysis for nRMSSD was performed, and the area under the curve (AUC) of the ROC-analysis was calculated.

3. Results

Five of the 107 patients enrolled were excluded, due to missing data or poor ECG Holter data quality. The 102 patients analyzed (age 71.0 ± 11.9 years; 52% male) had a mean CHA₂DS₂-VASc-Scores of 2.7. Demographical data, comorbidities, and concomitant medication of these patients are given in Table 1.

Table 1. Demographics, comorbidities, concomitant medication, and CHA₂DS₂-VASc Score of patients enrolled.

Patient Characteristics	No. (%)
Sex	
Male	53 (52.0)
Female	49 (48.0)
Age [years]	71.0 ± 11.9
Height [cm]	176.6 ± 10.8
Weight [kg]	86.1 ± 20.0
BMI [kg/m ²]	28.8 ± 5.4
Arm circumference [cm]	29.6 ± 3.7
Comorbidities	
Arterial hypertension	82 (80.4)
Diabetes mellitus	20 (19.6)
Stroke/ Myocardial infarction	21 (20.6)
Reduced left ventricular ejection fraction	32 (31.4)
Peripheral vascular disease	2 (1.9)
CHA ₂ DS ₂ -VASc-Scores	
0	8 (7.8)
1	15 (14.7)
2	17 (16.7)
3	30 (29.4)
4	23 (22.5)
5	9 (8.8)
>5	0 (0.0)
Mean	2.7 ± 1.4
Concomitant medication	
Anticoagulants	90 (88.2)
Antiplatelet	14 (13.7)
Beta-blocker	82 (80.4)
Calcium channel blocker	23 (22.5)
Renin-angiotensin system inhibitors	68 (66.7)
Other antihypertensive drugs	52 (51.0)
Other antiarrhythmic drugs	16 (15.7)
Glycosides	9 (8.8)
Heart rhythm by ECG Holter reads	
Sinus rhythm	43 (42.2)
Atrial fibrillation	48 (47.0)
Atrial flutter	11 (10.8)

By means of ECG Holter recording the patients were diagnosed (Cohens kappa 0.87) as having: Only sinus rhythm ($n = 43$, 42.2%), AF ($n = 48$, 47.0%), or atrial flutter episodes ($n = 11$, 10.8%). Patients with sinus rhythm were younger compared to those with AF ($p = 0.026$). There were no significant differences between patients with different heart rhythms with respect to comorbidities and concomitant medication.

The mean data recording time was 23.0 ± 3.3 h, comprising 2306 h of total recording time. In 62 out of the 102 patients (60.8%), the interpretable time was $\geq 80\%$; for the algorithms applied 1781 h (77.2%; average of 17.7 h) were evaluable (Tables 2 and 3); however, the time varied considerably among patients (SD 23.2%).

Table 2. Mean interpretable time, sensitivity, specificity, positive predictive value, negative predictive value, and AUC of ROC-analysis for detection of AF by using PPG analysis overall and during moderate physical activity and the average sensitivity/ specificity with SD estimated with 1-nearest neighbor classification and a deep neural network trained on five-minute periods on different training and validation test splits.

Method	Sensitivity [%]	Specificity [%]	PPV [%]	NPV [%]	AUC
nRMSSD	95.2	92.5	70.1	97.8	0.97
-periods in physical activity	92.9	85.5	63.1	97.7	-
1-nearest neighbor classification	96.0 ± 0.4	99.0 ± 0.2	94.7 ± 0.6	99.3 ± 0.0	0.98 ± 0.2
-periods in physical activity	96.8 ± 0.6	96.9 ± 0.5	94.3 ± 0.4	99.3 ± 0.1	-
DNN	97.0 ± 0.3	95.0 ± 0.4	81.0 ± 1.3	99.3 ± 0.1	0.99 ± 0.2
(classifier trained on annotated data)	97.0 ± 0.3	95.8 ± 0.4	83.8 ± 1.2	99.3 ± 0.1	-
-periods in physical activity					

nRMSSD = normalized root mean square of the successive difference, DNN = deep neural network, PPV = positive predictive value, NPV = negative predictive value, AUC = area under the receiver operating characteristics curve, ROC = receiver operating characteristic.

Table 3. Differences in demographics, medical characteristics, concomitant medication, and measurement conditions (below the bold line) of patients with interpretable time $< 80\%$ and $\geq 80\%$. (Significant differences are marked in bold, Continuous variables are given as mean \pm SD).

Characteristics	Interpretable Time $< 80\%$	Interpretable Time $\geq 80\%$	<i>p</i> Value
	No. (%)		
Count	40 (39.2)	62 (60.8)	
Sex			
Male	18 (45.0)	35 (56.5)	0.312
Female	22 (55.0)	27 (43.5)	
Age [years]	74.3 ± 9.8	68.9 ± 12.8	0.023
Height [cm]	168.6 ± 10.2	175.1 ± 10.6	0.003
Weight [kg]	86.8 ± 23.5	85.7 ± 17.6	0.619
Arterial hypertension	35 (87.5)	47 (75.8)	0.203
Diabetes mellitus	9 (22.5)	11 (17.7)	0.614
Stroke/myocardial infarction	10 (25.0)	11 (17.7)	0.454
Reduced left ventricular ejection fraction	17 (42.5)	15 (24.2)	0.080
Peripheral vascular disease	1 (2.5)	1 (1.6)	nA

Table 3. Cont.

Characteristics	Interpretable Time < 80%	Interpretable Time \geq 80%	<i>p</i> Value
	No. (%)		
CHA ₂ DS ₂ -VAsC-Scores			
0	2 (5.0)	6 (9.7)	0.172
1	2 (5.0)	13 (21.0)	
2	6 (15.0)	11 (17.7)	
3	13 (32.5)	17 (27.4)	
4	12 (30.0)	11 (17.7)	
5	5 (12.5)	4 (6.5)	
Anticoagulants	36 (90.0)	54 (87.1)	0.760
Antiplatelet	4 (10.0)	10 (16.1)	0.557
Beta-blocker	37 (92.5)	45 (72.6)	0.020
Calcium channel blocker	13 (32.5)	10 (16.1)	0.088
Renin-angiotensin system inhibitors	32 (80.0)	36 (58.1)	0.031
Heart rhythm			0.225
Sinus rhythm	14 (35.0)	29 (46.8)	
Atrial fibrillation	23 (57.5)	25 (40.3)	
Atrial flutter	3 (7.5)	8 (12.9)	
Arm circumference [cm]	29.9 \pm 2.9	29.9 \pm 4.7	0.559
Activity index (median)	14.7%	14.9%	0.204

1 nRMSSD-based algorithm

Detection of AF in the algorithm testing dataset was possible with a sensitivity of 95.2% and a specificity of 92.5% (Table 2) based on nRMSSD algorithm. Data obtained with the ECG Holter contained 5156 five-minute periods of AF. For 4469 of these episodes, simultaneous wearable data of sufficient quality was available. Of these 4,469 periods (algorithm training and algorithm testing), 4141 were correctly classified (true positive) as AF. In total, 1905 periods were classified false-positive, 328 periods were false-negative. Of the 1905 false-positive periods, 88 (4.6%) had a positive activity index. During 3,464 five-minute time periods with physical activity, AF was present in 755 (21.8%) periods. Of these, 701 periods were correctly classified as AF with a minor decrease in sensitivity (92.9%) and specificity (85.5%).

2 DNN-based algorithm

Further improvement in the detection of AF was achieved by means of the DNN (Table 2). On 10 different training /validation splits, the best model achieved a sensitivity of 96.9% and specificity of 95.4% (AUC 0.99). With ten-fold cross-validation of the models applied to the test set resulted in an average sensitivity of 96.9 \pm 0.3% and a specificity of 95.0 \pm 0.4% (AUC 0.99 \pm 0.1). Applying a fully unsupervised approach to the complete datasets resulted in a sensitivity of 96.7% and a specificity of 98.6% (AUC 0.98). With the same cross-validation methods applied to the test set on average, a sensitivity of 96.0 \pm 0.4% and a specificity of 99.0 \pm 0.2% (AUC 0.98 \pm 0.2) was achieved. In the five-minute periods with a positive physical activity index, sensitivity (96.8 \pm 0.6), and specificity (96.9 \pm 0.5; AUC 98.9 \pm 0.1) of AF detection remained unchanged with the DNN.

3 Further analysis

Patients with an interpretable time \geq 80% were allocated to one group, and differences in comorbidities, concomitant medication, arm circumference, and activity index were compared to

those with an interpretable time < 80% (Table 3). Descriptive characteristics between the two groups differed as follows: Patients with an interpretable time < 80% were older ($p = 0.023$) and used more antihypertensive agents (beta-blockers, $p = 0.020$; renin-angiotensin system inhibitors, $p = 0.031$). Logistic regression analysis showed that age ($p = 0.039$, OR 0.95, CI 0.904–0.997) had a negative impact on interpretable time. In contrast, height had a positive impact ($p = 0.002$, 1.10, 1.034–1.162).

Measurement conditions in both groups with respect to heart rhythm, side of recording, arm circumference, and activity index were comparable (Table 3). The physical activity level of all patients during 24 h was 16.1% based on a positive activity index in five-minute periods.

The activity index showed peaks after breakfast and in the afternoon (Figure 4).

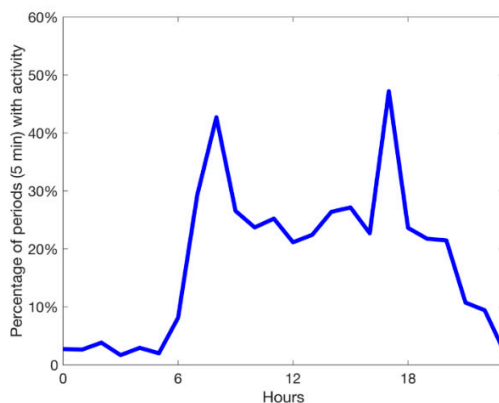


Figure 4. Percentage of five-minute periods with a positive physical activity index of all patients for 24 h.

Carrying the wearable did not induce any discomfort or pain in 97.5% of the patients. More than 70% of the patients could envisage using such a wearable for home monitoring. No serious adverse effects were observed during the trial; however, one device-related adverse effect was observed; a skin irritation after wearing the device was fully reversible after six days.

4. Discussion

Our study suggests that reliable detection of AF in high-risk patients for AF is possible with the medical wearable used, also during time periods with physical activity. The deep neural network approach showed an even better ability of AF detection than the established nRMSSD algorithm. The DNN approach enables a reliable computer-based analysis, and thereby, the option of a real-time AF detection. Using a passive measurement approach, a high interpretable time proportion (77.2%) was achieved.

The high-risk population studied was comparable with respect to age and cohort distribution in terms of heart rhythm to the population of the multicenter trial of Brasier et al. [13]; however, the population in their trial had a higher mean CHA₂DS₂-VASc-Scores reflecting a higher prevalence of comorbidities.

Detection of AF with nRMSSD in five-minute periods showed higher sensitivity, but lower specificity than in other studies conducted with an active measurement approach; however, our results were obtained in a not physically restricted population [11,13]. Detection of AF within periods of physical activity represents a challenge for wearables (also with ECG Holter monitoring). In some trials, there was a gap in the detection of AF during physical active vs. restricted physical conditions [17]. In other trials, like the Apple Heart study, no measurements were performed while participants were physically active [10]. In our trial, the overall physical activity index (as provided by the wearable) observed probably does not reflect real-world physical activity, since only inpatients were enrolled. Nevertheless, also in periods with a positive activity index, detection of AF was feasible with good reliability. However, it is a limitation that the activity index used was not assessed with a standardized

reference method in parallel. There is a difference between the number of available five-minute periods in AF in ECG Holter and wearable data, i.e., due to the interpretable time of wearable data.

The deep learning setup applied—consisting of unsupervised feature extraction followed by unsupervised classification—showed higher sensitivity and specificity in detecting AF. These results were comparable to Tison et al. [17]; additionally, they were achieved with unlabeled data. Large amounts of unlabeled data were accurately classified with no cumbersome annotation of data performed. Furthermore, no data-pre-processing steps were needed, such as rescaling mean and variance of IBI values, and noise is mostly discarded in the encoding IBI values with respective quality indices. DNNs are preferably trained on raw-data, as they can extract information from data that human observers would miss; however, even with the use of pre-processed data, such approaches improve detection of AF. The wearable utilized in this study employed proprietary algorithms and only provided pre-processed data. This might impact the information content originally contained in the raw data. Especially in a medical context, it should be mandatory to perform context-related accuracy testing when using pre-processed data (see Appendix A). Testing the pre-processed data revealed a comparable correlation for ECG and PPG derived heart rate estimation [21]. For practical application of such medical wearables, utilization of pre-processed data may represent the more frequent use case.

In this trial, the recording time was identical to the monitoring time (=time device was used by patients), driven by the fact that the wearable was attached and dismantled by the investigator. However, this might be different in daily practice, as patients might, e.g., wear the device while the battery is empty. It is of interest to note that in other studies, no clear time definitions and data are provided, e.g., in the Apple Heart and Huawei Heart study [10,12]. An analysis of variables that have an impact on interpretable time in this trial is at least partly in accordance with published data [22]. The impact of age and height on the interpretable time shown by the logistic regression was modest.

The European Society of Cardiology guidelines on the management of AF recommends screening for silent AF with ECG-based devices in selected patient populations [4]. New technologies, such as smart watches (ECG and PPG based), are not yet recommended in the guidelines as no formal evaluation of these devices has been performed yet. Passive monitoring approaches with wrist-worn smartwatches (as those used in the Apple Heart and Huawei Heart studies) showed an acceptable diagnostic performance in a non-risk population. However, the performance of such devices is not sufficient for screening for silent AF, due to their low interpretable time with respect to a 24 h measurement. It is known that recording of ECG Holter is hampered by noisy measurements and/or artifacts induced by physical activity, thus potentially leading to under-diagnosis of AF episodes. In a recent analysis, an elimination rate of 30% (i.e., not interpretable ECG recording time) of data was observed [23]. A disadvantage of conventional adhesive ECG-patches used until now is the limited adherence of patients, due to discomfort, visibility, and skin reactions.

The wearable used in this trial was chosen because of a tight upper arm fit in order to reduce artifacts induced by probe-tissue movement, e.g., due to physical activity [15]. In this respect, it is worth mentioning that the activity index had no significant impact on interpretable time. Moreover, ambient light emitted by external sources interference is minimized by a sensor location most often covered by clothing. The medical wearable could be connected to secure web-based services, and thus, provide immediate feedback. The respective results of the questionnaire used in this trial showed that the patients appreciated the non-invasive wearable; however, it was used for one day only. It remains to be studied if patients are willing to wear such a medical wearable for long-term monitoring (=high adherence rate) as it is not a 'lifestyle-device'. Nevertheless, patients might favor the comfort of such a wearable in contrast to other options.

Till today it is still under discussion which duration of AF burden is associated with an increased risk for clinical complications, such as ischemic stroke [6]. Considering the commercially available wearables and the studied device, data acquisition is based on a block-wise approach (i.e., five-minute time periods). It is not clear which time resolution (=number of data points per time unit) is needed in order to be able to detect all AF episodes with sufficient diagnostic accuracy.

In summary, medical wearables with such specifications offer the option of permanent surveillance, i.e., live monitoring of the patients by health care professionals. The workload of specialized clinics may be reduced if live remote patient monitoring was enabled by modern wearables. This could be a contributively brick for structured disease management applications [24].

This trial evaluated only hospitalized patients at high risk for AF in a proof of concept approach. Sensitivity and specificity have to be further evaluated in a population at a lower risk for AF. For this study, we evaluated a population with a high risk of AF. Importantly, patients at ‘moderate’-risk of AF might represent the most relevant population, in whom longer monitoring times are required to detect AF episodes. Compliance and adherence were only tested in patients carrying the wearable for 24 h. It remains to be studied how good the acceptance of the wearable is over longer time periods. Such studies would also help to see how limits of the current version of this wearable can be handled. If these and other exogenous factors can be overcome, this would achieve a high interpretable time to maximize high-resolution data. A limitation of our study was that we had to rely on data acquisition and raw data analysis that was implemented in the wearable and on proprietary quality indices. It is acknowledged that the signal quality index is critical for AF detection, as noisy sinus rhythm might be mis-detected as AF. A preliminary accuracy testing was performed (see Appendix B).

5. Conclusions

In conclusion, detection of AF with a medical wearable attached to the upper arm is a feasible and reliable approach, also during physical activity for remote monitoring purposes. The results presented encourage the performance of long-term clinical trials with a focus on everyday conditions. Assuming a positive outcome of such studies, monitoring of patients with AF might move away from Holter ECG towards medical wearables.

Author Contributions: Conceptualization, M.J., T.A.D., A.-P.Z., N.D., M.S., G.K., S.I. and L.H.; methodology, M.J., T.A.D., R.G., M.K. and C.B.; software, T.A.D., R.G., and C.B.; validation, M.J., T.A.D., R.G., M.K. and C.B.; formal analysis, M.J., T.A.D., R.G., M.K. and C.B.; investigation, M.J., A.-P.Z., N.D., S.I., M.S.; resources, M.K., S.I., M.S.; data curation, M.J., T.A.D.; writing—original draft preparation, M.J., T.A.D., L.H., S.I., and M.S.; writing—review and editing, G.K., D.M.-W., A.N. and N.M.; visualization, M.J., T.A.D., R.G. and C.B.; supervision, D.M.-W., N.M. and S.I.; project administration, M.J., L.H. and M.S.; funding acquisition, M.J., S.I., and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding from the internal grant program (PhD Program Biomedicine) of the Faculty of Health at Witten/Herdecke University, Germany and by a grant of HELIOS Kliniken GmbH (Grant-ID 047476), Germany.

Acknowledgments: We would like to thank A. Caduff, A. Uhde, P. Vettel, R. Amacher, G. Haas, the clinical staff of the Department of Cardiology, Helios University Hospital of Wuppertal and the Centre for Clinical Studies (P. Thürmann, K. Graf, W. Eglmeier, R. Geißen, S. Schmiedl, F. Hohmann), University Witten/Herdecke for their support.

Conflicts of Interest: The author declares that there is no conflict of interest. This trial was an Investigator Initiated Trial. This study used the wearable “Everion”-Device provided by Biovotion AG, Switzerland. Biovotion did not provide any financial support for the research and had no impact on writing of the manuscript. Biovotion did not participate in the analysis of the data or influence the conclusions in any sense.

Appendix A

Table A1. Accuracy of heart rate measurements of the upper-arm wearable (PPG based) compared to ECG Holter recording. AF, atrial fibrillation.

Cohort	Rho	r ²	% ±10-Beat
All	0.89	0.66	0.88
Sinus rhythm	0.94	0.85	0.97
AF	0.80	0.51	0.80
A-flutter	0.83	0.64	0.88

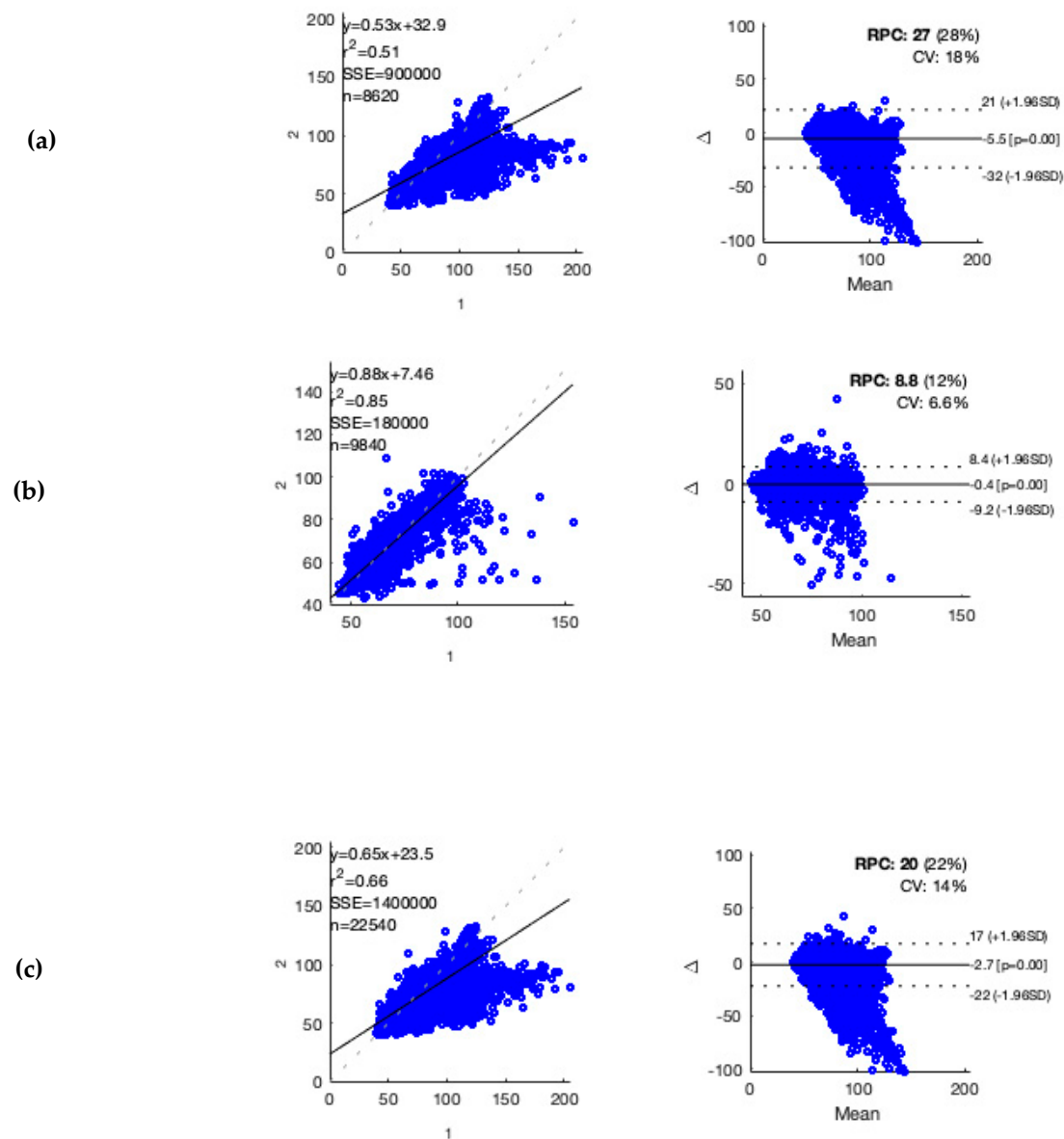


Figure A1. (a–c) The accuracy of heart rate estimation by the medical wearable compared to ECG was calculated by using the Spearman correlation (left figure). Additionally, a Bland–Altman graph was plotted with 95% LoA (Limits of Agreement) (right figure). Automated ECG processing was performed using an open-source algorithm (Sedghamiz H. Complete Pan Tompkins Implementation ECG QRS detector. In: MATLAB Central File Exchange 2019.).

Appendix B. Deep Neural Network

Data preparation: We split recordings of inter-beat intervals (IBI) values into five-minute periods, which are manually classified into AF, A-flutter, and non-AF. We extract periods that have at least 80% reliable IBI values (quality score of IBI values ≤ 13 , with 1 indicating best quality and 16 worst quality) and split them into a training set and validation set. We encode the IBI values together with their quality score into a 16 dimensional vector, e.g., $\vec{z}_t = [0, \dots, \text{IBI-value}, \dots, 0]$, where IBI values with different quality score are orthogonal to each other. As the dataset contains more non-AF examples than AF examples, we simply oversample the AF class to make the dataset balanced. For the analysis, we took all five-minute periods that contain at least 200 IBI values and could be uniquely assigned to the non-AF or AF class.

Neural Network Model: We trained deep neural networks to maximize the mutual information $I(\vec{c}_t; \vec{z}_t)$ between the first t IBI values and the subsequent k IBI values within a five-minute periods, with t a randomly chosen time point. Computation of the mutual information is realized by first encoding the sequence of IBI vectors $(\vec{z}_1, \dots, \vec{z}_t)$ by recurrent neural network to encode the information into a vector \vec{c}_t and make use of the InfoNCE objective [18] to estimate the mutual information between \vec{c}_t and \vec{z}_{t+k} for $k \in \{1, 2, 3, 4\}$. As a result, \vec{c}_t contains all information that can be used to predict the next 4 IBI values. Using mutual information as objective for unsupervised learning has the advantage that no data-pre-processing steps are needed, such as rescaling mean and variance of IBI values, and that noise is mostly discarded in the encoding of \vec{c}_t . To classify the training set into AF and non-AF we train a fully connected network with two hidden layers that takes \vec{c}_t at the end of the period as input and predicts the labels from the manual classification. This deep learning setup (consisting of unsupervised feature extraction followed by a classifier on the relevant features) is especially valuable for cases where large amounts of unlabeled data can be recorded easily, and accurate classification is expensive or time-consuming.

References

1. Chugh, S.S.; Havmoeller, R.; Narayanan, K.; Singh, D.; Rienstra, M.; Benjamin, E.J.; Gillum, R.F.; Kim, Y.-H.; McAnulty, J.H.; Zheng, Z.-J.; et al. Worldwide Epidemiology of Atrial Fibrillation: A Global Burden of Disease 2010 Study. *Circulation* **2013**, *129*, 837–847. [[CrossRef](#)] [[PubMed](#)]
2. Zoni-Berisso, M.; Lercari, F.; Carazza, T.; Domenicucci, S. Epidemiology of atrial fibrillation: European perspective. *Clin. Epidemiol.* **2014**, *6*, 213–220. [[CrossRef](#)] [[PubMed](#)]
3. Benjamin, E.J.; Blaha, M.J.; Chiuve, S.E.; Cushman, M.; Das, S.R.; Deo, R.; de Ferranti, S.D.; Floyd, J.; Fornage, M.; Gillespie, C.; et al. Heart Disease and Stroke Statistics–2017 Update: A Report From the American Heart Association. *Circulation* **2017**, *135*, e146–e603. [[CrossRef](#)] [[PubMed](#)]
4. Kirchhof, P.; Benussi, S.; Kotecha, D.; Ahlsson, A.; Atar, D.; Casadei, B.; Castellá, M.; Diener, H.-C.; Heidbuchel, H.; Hendriks, J.; et al. 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *Eur. J. Cardio-Thorac. Surg.* **2016**, *50*, e1–e88. [[CrossRef](#)]
5. Shamloo, A.S.; Dagues, N.; Arya, A.; Hindricks, G. Atrial fibrillation: A review of modifiable risk factors and preventive strategies. *Rom. J. Intern. Med.* **2019**, *57*, 99–109. [[CrossRef](#)]
6. Jones, N.; Taylor, C.J.; Hobbs, F.R.; Bowman, L.; Casadei, B. Screening for atrial fibrillation: A call for evidence. *Eur. Hear. J.* **2019**, *41*, 1075–1085. [[CrossRef](#)]
7. Svennberg, E.; Engdahl, J.; Al-Khalili, F.; Friberg, L.; Frykman, V.; Rosenqvist, M. Mass Screening for Untreated Atrial Fibrillation. *Circulation* **2015**, *131*, 2176–2184. [[CrossRef](#)]
8. Wachter, R.; Gröschel, K.; Gelbrich, G.; Hamann, G.F.; Kermer, P.; Liman, J.; Seegers, J.; Wasser, K.; Schulte, A.; Jürries, F.; et al. Holter-electrocardiogram-monitoring in patients with acute ischaemic stroke (Find-AF_{RANDOMISED}): An open-label randomised controlled trial. *Lancet Neurol.* **2017**, *16*, 282–290. [[CrossRef](#)]
9. Steinhubl, S.R.; Waalen, J.; Edwards, A.M.; Ariniello, L.M.; Mehta, R.R.; Ebner, G.S.; Carter, C.; Baca-Motes, K.; Felicione, E.; Sarich, T.; et al. Effect of a Home-Based Wearable Continuous ECG Monitoring Patch on Detection of Undiagnosed Atrial Fibrillation. *JAMA* **2018**, *320*, 146–155. [[CrossRef](#)]
10. Perez, M.V.; Mahaffey, K.W.; Hedlin, H.; Rumsfeld, J.S.; Garcia, A.; Ferris, T.; Balasubramanian, V.; Russo, A.M.; Rajmane, A.; Cheung, L.; et al. Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. *N. Engl. J. Med.* **2019**, *381*, 1909–1917. [[CrossRef](#)]
11. Krivoshei, L.; Weber, S.; Burkard, T.; Maseli, A.; Brasier, N.; Kühne, M.; Conen, D.; Huebner, T.; Seeck, A.; Eckstein, J. Smart detection of atrial fibrillation. *Europace* **2017**, *19*, 753–757. [[CrossRef](#)] [[PubMed](#)]
12. Guo, Y.; Wang, H.; Zhang, H.; Liu, T.; Liang, Z.; Xia, Y.; Yan, L.; Xing, Y.; Shi, H.; Li, S.; et al. Mobile Photoplethysmographic Technology to Detect Atrial Fibrillation. *J. Am. Coll. Cardiol.* **2019**, *74*, 2365–2375. [[CrossRef](#)]
13. Brasier, N.; Raichle, C.J.; Dörr, M.; Becke, A.; Nohturfft, V.; Weber, S.; Bulacher, F.; Salomon, L.; Noah, T.; Birkemeyer, R.; et al. Detection of atrial fibrillation with a smartphone camera: First prospective, international, two-centre, clinical validation study (DETECT AF PRO). *Europace* **2018**, *21*, 41–47. [[CrossRef](#)] [[PubMed](#)]

14. Jacobsen, M.; Dembek, T.A.; Kobbe, G.; Gaidzik, P.W.; Heinemann, L. Noninvasive Continuous Monitoring of Vital Signs With Wearables: Fit for Medical Use? *J. Diabetes Sci. Technol.* **2020**. [[CrossRef](#)] [[PubMed](#)]
15. Allen, J. Photoplethysmography and its application in clinical physiological measurement. *Physiol. Meas.* **2007**, *28*, R1–R39. [[CrossRef](#)] [[PubMed](#)]
16. Schäfer, A.; Vagedes, J. How accurate is pulse rate variability as an estimate of heart rate variability? *Int. J. Cardiol.* **2013**, *166*, 15–29. [[CrossRef](#)]
17. Tison, G.H.; Sanchez, J.M.; Ballinger, B.; Singh, A.; Olgin, J.E.; Pletcher, M.J.; Vittinghoff, E.; Lee, E.S.; Fan, S.M.; Gladstone, R.A.; et al. Passive Detection of Atrial Fibrillation Using a Commercially Available Smartwatch. *JAMA Cardiol.* **2018**, *3*, 409–416. [[CrossRef](#)]
18. Oord, A.V.D.; Li, Y.; Vinyals, O. Representation Learning with Contrastive Predictive Coding. *arXiv* **2018**, arXiv:1807.03748v2. Available online: <https://arxiv.org/abs/1807.037482018> (accessed on 22 January 2019).
19. Bumgarner, J.M.; Lambert, C.T.; Hussein, A.A.; Cantillon, D.J.; Baranowski, B.; Wolski, K.; Lindsay, B.D.; Wazni, O.M.; Tarakji, K.G. Smartwatch Algorithm for Automated Detection of Atrial Fibrillation. *J. Am. Coll. Cardiol.* **2018**, *71*, 2381–2388. [[CrossRef](#)]
20. William, A.D.; Kanbour, M.; Callahan, T.; Bhargava, M.; Varma, N.; Rickard, J.; Saliba, W.; Wolski, K.; Hussein, A.; Lindsay, B.D.; et al. Assessing the accuracy of an automated atrial fibrillation detection algorithm using smartphone technology: The iREAD Study. *Hear. Rhythm.* **2018**, *15*, 1561–1565. [[CrossRef](#)]
21. Koshy, A.N.; Sajeev, J.K.; Nerlekar, N.; Brown, A.J.; Rajakariar, K.; Zureik, M.; Wong, M.C.; Roberts, L.; Street, M.; Cooke, J.; et al. Smart watches for heart rate assessment in atrial arrhythmias. *Int. J. Cardiol.* **2018**, *266*, 124–127. [[CrossRef](#)] [[PubMed](#)]
22. Shcherbina, A.; Mattsson, C.M.; Waggott, D.; Salisbury, H.; Christle, J.W.; Hastie, T.; Wheeler, M.T.; Ashley, E.A. Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *J. Pers. Med.* **2017**, *7*, 3. [[CrossRef](#)] [[PubMed](#)]
23. Lee, J.; McManus, D.D.; Merchant, S.; Chon, K.H. Automatic motion and noise artifact detection in Holter ECG data using empirical mode decomposition and statistical approaches. *IEEE Trans. Biomed. Eng.* **2011**, *59*, 1499–1506. [[CrossRef](#)] [[PubMed](#)]
24. Kotecha, D.; Chua, W.W.L.; Fabritz, L.; Hendriks, J.; Casadei, B.; Schotten, U.; Vardas, P.; Heidbuchel, H.; Dean, V.; Kirchhof, P. European Society of Cardiology smartphone and tablet applications for patients with atrial fibrillation and their health care providers. *Europace* **2017**, *20*, 225–233. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Chapter 6

Anomaly Detection in Chest X-ray Images

X-ray images have been widely used for medical diagnoses of cardiothoracic and pulmonary abnormalities due to its noninvasiveness. Advancement in computer-aided diagnostic technologies, such as deep supervised methods, can help radiologists with a reliable early treatment and reduce diagnosis time. Nevertheless, these methods are prone to the small number of labeled samples and are limited to a specific abnormality.

We combine a self-supervised contrastive learning framework for X-ray anomaly detection trained only with the normal (i.e., healthy) images to make our method future-ready for yet unknown anomalies. The self-supervised representations are highly effective for the task of anomaly detection in our framework. We define an anomaly detection score based on Mahalanobis distance applicable for detecting anomalies. We found that our approach outperforms all previous unsupervised methods on a pneumonia detection challenge dataset. This work may allow for improving radiology work-flow and clinical decision-making.

6.1 Pneumonia Detection with Semantic Similarity Scores

R. Gholamipoor, N. Rafiee, M. Kollmann. Pneumonia Detection with Semantic Similarity Scores. *ISBI*, 2022.

Status: Published.

Contributions: The research and preparation of this manuscript were done jointly by R. Gholamipoor and N. Rafiee under the supervision of Prof. Dr. Markus Kollmann.

PNEUMONIA DETECTION WITH SEMANTIC SIMILARITY SCORES

Rahil Gholamipoor^{*1}, *Nima Rafiee*^{*1}, *Markus Kollmann*^{1,2}

Department of Computer Science¹, Department of Biology²
Heinrich Heine University, Düsseldorf, Germany
{rahil.gholamipoorfard, nima.rafiee, markus.kollmann}@hhu.de

X-ray images have been widely used for medical diagnoses of cardiothoracic and pulmonary abnormalities due to their noninvasiveness. Advancement in computer-aided diagnostic technologies, such as deep supervised methods, can help radiologists with a reliable early treatment and reduce diagnosis time. Nevertheless, these methods are prone to the small number of labeled samples and are limited to a specific abnormality. In this paper, we combined a self-supervised contrastive method with a Mahalanobis distance score to develop an abnormality detection method that uses only healthy images during the training procedure. We were able to outperform previous unsupervised methods for the task of Pneumonia detection. We show that representation learned by the self-supervised method improves the supervised tasks for Pneumonia detection.

1. INTRODUCTION

Chest X-ray has been used for medical screening in order for the detection of cardiothoracic and pulmonary abnormalities, which are one of the causes of mortality worldwide. Radiologists widely use chest X-ray images to diagnose lung-related diseases such as pneumonia. A computer-aided diagnostic approach would be very helpful to allow radiologists to detect potential abnormalities in chest X-ray images for early care and treatment. Recently supervised deep learning approaches have achieved promising results in abnormality detection for these images. Hendrycks et al. [1] proposed the maximum value of posterior distribution from the classifier as a baseline method to detect anomalies and Liang et al. [2] improved performance using temperature scaling and input pre-processing. However, these approaches [3] require large, annotated datasets for training which is not always feasible. Additionally, it is in general challenging to acquire enough supervised data for rare pathologies. To address these problems, many approaches have exploited unsupervised or semi-supervised frameworks to use unlabeled data for extracting generalizable features in medical images [4, 5]. Among unsupervised approaches, reconstruction-based methods assume that anomalies cannot be represented and reconstructed accu-

rately by a model trained only on normal data. However, in practice, these models can also reconstruct abnormal samples fairly well and thus fail to detect them [5, 6]. To overcome this problem, Mao et al. [7] trained an autoencoder model to not only reconstruct the corresponding normal version of any input but also estimate the uncertainty of reconstruction at each pixel to enhance the performance of anomaly detection. In [8], an autoencoder is trained while a constraint is additionally imposed on the lower-dimensional representation of the data in which features of the same X-ray images under random data augmentations are invariant, while the features of different images are scattered.

Recently the effectiveness of self-supervised contrastive learning has been proven in different domains, e.g. the visual domain [9, 10], which enables learning of robust representations through unlabeled data. Azizi et al. [11] investigated the effect of self-supervised pre-training on the classification downstream task on the CheXpert dataset [12]. Zhang et al. [13] improved on supervised-based pneumonia detection using a contrastive-based pre-training and leveraging image description as an extra modality. In this paper, we utilize a self-supervised contrastive method to construct an anomaly detection score based on Mahalanobis distance for anomaly detection. To the best of our knowledge, we achieved state-of-the-art results for anomaly detection among all methods that can be applied to unlabeled data.

2. METHOD

2.1. Contrastive Learning

Given unlabeled training data, self-supervised contrastive representation learning aims to train a feature extractor, g_θ , to discriminate similar samples from dissimilar ones. Using image transformations that keep the semantics, each image is augmented twice, referred to as positives. The function g_θ is optimized to pull semantically similar samples together while pushing away from other images, referred to as negatives. Assuming that (x_i, x_j) is a positive pair for the i^{th} image from a batch of N images, τ is a scalar temperature parameter and $\text{sim}(u, v) = \frac{u^T v}{\|u\| \|v\|}$ denotes the dot product between l_2 normalized u and v (i.e. cosine similarity). Contrastive learning

*Equal contribution

minimizes the following loss for a positive pair of examples (i, j) , referred to as Normalized Temperature-scaled Cross-entropy (NT-Xent):

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(\text{sim}(z_i, z_k)/\tau)} \quad (1)$$

where $\mathbb{1}_{k \neq i}$ is an indicator function evaluating to 1 iff $k \neq i$. z_i denotes the output feature of the contrastive layer. Intuitively, this loss is the log loss of a $(2N)$ -way softmax-based classifier that tries to classify x_j as a positive sample for x_i . One can define the contrastive feature $z(x)$ directly from the encoder g_θ , i.e., $z(x) = g_\theta(x)$ [10], or apply an additional projection layer f_ϕ , i.e., $z(x) = f_\phi(g_\theta(x))$ [9]. The contrastive loss (Eq.1) can be minimized by different mechanisms that differ in how the negative samples are maintained. Chen et al. [9] take negatives from the same batch but it requires a large batch size to provide a large set of negative pairs. Alternatively, Eq.1 can be minimized with sufficient number of negative pairs without using large batch sizes by maintaining negatives in a queue [10]. The encoded representations of the current mini-batch are enqueued while the oldest are dequeued. Unlike [9] in which only one encoder is used, following [10] we use two encoders, a query encoder and a slowly progressing key encoder, implemented as a momentum-based moving average of the query encoder.

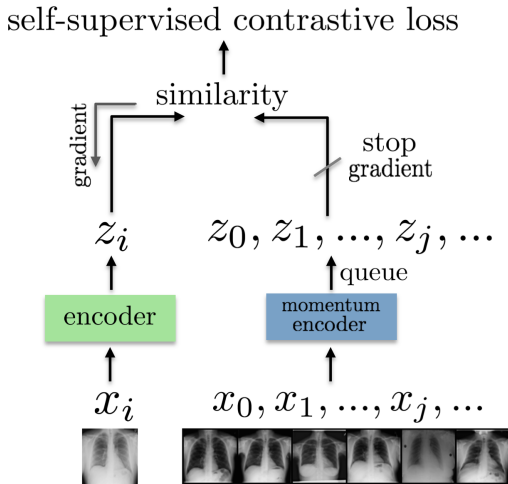


Fig. 1. The query encoder is updated end-to-end by back-propagation while the key encoder maintains a queue and is updated with momentum-based moving average. We got our best results when the model is pre-trained on ImageNet dataset.

2.2. Score Function for Anomaly Detection

Mahalanobis distance-based confidence score We use Mahalanobis distance on feature space $h(x)$ of the trained contrastive encoder as a score function for anomaly detection.

Mahalanobis distance achieved promising results for supervised anomaly detection. Lee et al. [14] show that with a well-trained softmax classifier, applying Mahalanobis distance on feature space using the class means and the feature covariance matrix can reach the state of the art results on supervised anomaly detection. To measure the Mahalanobis distance for a given test sample x first, we apply K-means clustering with $K = 1$ on the feature space $h(x)$ of training data. This clustering helps to reduce computation time as we only compare the distance with the cluster mean. The anomaly score $s(x)$ for a test sample x is given by the Mahalanobis distance

$$s(x) := (h(x) - \mu_m)^T \Sigma_m^{-1} (h(x) - \mu_m) \quad (2)$$

where μ_m and Σ_m are the mean and covariance of the feature vectors from the training data. The reason to use the Mahalanobis distance is to remove the dominance of larger eigenvalues in euclidean distance metric as shown in [15] eigenvalues have an approximately inverse correlation with anomaly detection performance.

3. EXPERIMENTAL SETUP

3.1. Dataset

RSNA¹. The Radiological Society of North America (RSNA) Pneumonia Detection Challenge dataset [16] is a publicly available dataset of frontal view chest radiographs. Each image was labeled as "Normal", "No Opacity/Not Normal" or "Opacity". The Opacity group consists of images with opacities suspicious for pneumonia, and images labeled "No Opacity/Not Normal" may have lung opacity but no opacity suspicious for pneumonia. The RSNA dataset is a subset of the National Institutes of Health (NIH) Chest X-Ray dataset [17]. It contains 26,684 X-rays with 8,851 normal, 11,821 no lung opacity/not normal and 6,012 lung opacity.

3.2. Self-supervised Contrastive Training

Experiments were carried out using ResNet50 neural network architecture. Following [9], two fully connected layers are used to map the output of ResNet to a 128-dimensional embedding space where the contrastive loss is applied. We perform training on RSNA with initialization from ImageNet self-supervised pre-trained weights. We train at batch size 128 for 100 epochs using SGD optimiser. The temperature τ in Eq.(1) is set as 0.07. At training time, we apply the following augmentations: (1) a 224×224 -pixel crop is taken from a randomly resized image (2) random rotation by an angle sampled from the uniform distribution $U(-20, 20)$ (3) random horizontal flip with probability 0.5 (4) brightness and contrast adjustments.

¹<https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data>

3.3. Evaluation Methodology

We evaluate the results using Area Under the Receiver Operating Characteristic curve (AUROC), which has the advantage to be scale-invariant "it measures how well predictions are ranked, rather than their absolute values" and classification-threshold-invariant "it measures how well anomaly samples are separated from the normal samples".

4. EXPERIMENTAL RESULTS

4.1. Self-Supervised Anomaly Detection

For Mahalanobis distance, the highest performance achieved from the last layer, the output after the average pooling layer, before the MLP head [15]. On RSNA dataset, to detect anomalies, we consider three different cases: "Normal" vs. "Opacity"; "Normal" vs. "No Opacity/Not Normal" and "Normal" vs. all "Opacity and No Opacity". In Table 1, we compare our method with both supervised methods and unsupervised methods trained on only healthy images. We averaged AUROC values over 5 different train/test splits.

Table 1. OOD detection performance (AUROC).

Methods	Opacity	No Opacity	All
<i>Methods making use of label information</i>			
Automated Abnormality Classification [18]	0.980	-	0.949
Pneumonai Detection using Radiomic Features[19]	0.923	-	-
ConVIRT [13]	-	-	0.908
<i>Unsupervised methods trained on normal samples</i>			
UAE[7]	0.89	0.78	0.83
Deep Anomaly Detection[20]	0.838	0.704	0.752
Generative Adversarial one-class classifier[5]	0.802	-	0.841
Ours	0.940	0.828	0.866

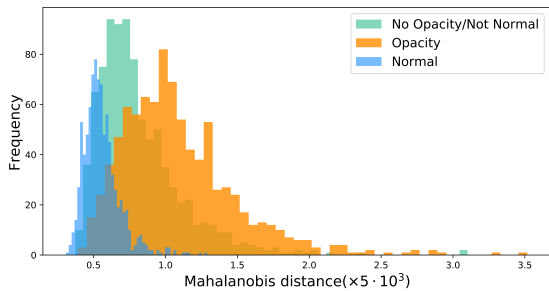


Fig. 2. Distributions over the anomaly detection score trained only on Normal samples and applied to the test sets of Normal as in-distribution, "Opacity" and "No Opacity/Not Normal" as out-distributions.

4.2. Pre-training and Label Efficiency of Multilabel Classification

In addition to the self-supervised anomaly detection task, we evaluate the learned representation by its performance in

KNN accuracy and the impact it has on multilabel classification downstream task. For the pre-training task, we use the same data split statistics as in [18] including 21, 152 training samples (14, 159 abnormal and 6, 993 normal samples). We use the same optimization config as for the anomaly detection task. The self-supervised pre-trained model achieved 1-NN accuracy of 79.01%. For the classification task, we replace the projection head of the contrastive encoder with a classification head, projecting the data into a one-dimensional scalar value and fine-tune the whole model with binary cross-entropy loss and same optimization config as in [18]. To see the effect of self-supervised pre-training, we start with a small fraction of training data and compare model AUROC performance on test data for different case studies. Figure 3 shows that self-supervised pre-training can significantly help with label efficiency and causes a considerable performance improvement when we have a small fraction of labeled samples for the downstream task. We achieve an AUROC score of 94.4% when fine-tuning with all labeled training data and an AUROC score of 82.97% when using only 100 labeled samples which are selected randomly from training data.

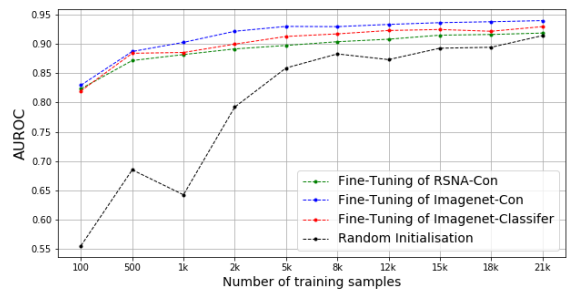


Fig. 3. Self-supervised pre-training increases the downstream classification task performance with small fraction of training samples. RSNA-Con and Imagenet-Con are fine-tunings of models with different model initialisation in self-supervised pre-training as follows: randomly initialised and initialised with Imagenet. Imagenet-Classifer stands for fine-tuning an already trained imagenet classifier and Random Initialisation is performing classification with random weight initialisation.

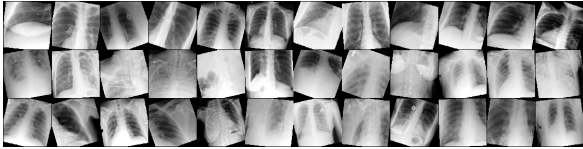
5. ABLATION STUDIES

5.1. Data Augmentation Details

In our setting, to train the self-supervised contrastive encoder, we utilize random crop (resize to 224×224), random rotation (image rotation by angle θ from range $(-20, 20)$), random horizontal flip, brightness and contrast adjustments as the data augmentations. Brightness and contrast adjustments are composed by color jittering. The details of these augmentations are provided in Table 2.

Table 2. Data augmentation used for contrastive training

Transformation	PyTorch snippet
Cropping	<code>transforms.RandomResizedCrop(224, scale=(0.08, 1.0))</code>
Rotation	<code>transforms.RandomRotation(20)</code>
Horizontal Flip	<code>transforms.RandomHorizontalFlip(p = 0.5)</code>
Color Jitter	<code>transforms.ColorJitter(0.4, 0.4, 0, 0)</code>
Normalization	<code>transforms.Normalize()</code>

**Fig. 4.** Examples of augmented images from RSNA dataset

5.2. Ablation on Batch Size

Training with small batches. Table 3 confirms that large batches are not necessary for a good performance in our anomaly detection problem. We scale the learning rate linearly with the batch size [21]. We averaged AUROC values for "Normal" vs. "Opacity" over 5 different train/test splits for each batch size.

Table 3. Effect of batch size

Batch size	AUROC
256	0.926
128	0.940
64	0.916

5.3. Ablation on Data Augmentation

Because of the less diverse nature of X-ray images, in our experiments, we used strong data augmentations in order to prevent over-fitting and improve anomaly detection performance. In Table 4, we change the strength of each augmentation individually while keeping the rest unchanged. We averaged AUROC values for "Normal" vs. "Opacity" over 5 different train/test splits.

Table 4. Effect of data augmentation

Transformation	AUROC
<code>transforms.RandomResizedCrop(224, scale=(0.4, 1.0))</code>	0.934
<code>transforms.RandomRotation(10)</code>	0.928
<code>transforms.ColorJitter(0.25, 0.25, 0, 0)</code>	0.926

5.4. Fine-tuning Implementation Details

To do the fine-tuning, we use the same data augmentation as used in [18]. Table 5 shows the augmentation details together

with related PyTorch code. We use a batch size of 128 for all experiments where the training samples are more than 1000 images and 64 where we have 100 and 500 training samples. Other optimisation hyper-parameters are the same for all experiments. Table 6 summarises the optimisation hyper-parameters used for fine-tuning.

Table 5. Data augmentation used for fine-tuning

Transformation	PyTorch snippet
Resize	<code>transforms.Resize(256)</code>
Cropping	<code>transforms.CenterCrop(224)</code>
Horizontal Flip	<code>transforms.RandomHorizontalFlip(p = 0.5)</code>
Color Jitter	<code>transforms.ColorJitter(0.3, 0.3, 0, 0)</code>
Random Affine	<code>transforms.RandomAffine(15, translate=(0.1, 0.1), scale=(0.9, 1.1))</code>
Normalization	<code>transforms.Normalize()</code>

Table 6. Hyper-parameters for fine-tuning training

Hyper-parameter	Default value
Number of epochs	50
Learning rate	10^{-2}
Weight decay	10^{-4}
Optimizer	SGD
Momentum of SGD	0.9

6. CONCLUSION

In this work, we proposed a self-supervised contrastive learning framework for X-ray anomaly detection trained only with the normal images to make our method future-ready for yet unknown anomalies. The self-supervised representations are highly effective for the task of anomaly detection in our framework. We define an anomaly detection score based on Mahalanobis distance applicable for detecting anomalies. We find that our approach outperforms all previous unsupervised methods on the RSNA pneumonia detection challenge dataset. This work may allow for improving radiology workflow and clinical decision-making.

7. REFERENCES

- [1] Dan Hendrycks and Kevin Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," 2018.
- [2] Shiyu Liang, Yixuan Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," 2020.
- [3] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," 2018.

- [4] Diana Davletshina, Valentyn Melnychuk, Viet Tran, Hitansh Singla, Max Berrendorf, Evgeniy Faerman, Michael Fromm, and Matthias Schubert, “Unsupervised anomaly detection for x-ray images,” 2020.
- [5] Yuxing Tang, Youbao Tang, Mei Han, Jing Xiao, and Ronald M. Summers, “Abnormal chest x-ray identification with generative adversarial one-class classifier,” 2019.
- [6] Erdi Çalli, Ecem Sogancioglu, Bram van Ginneken, Kicky G van Leeuwen, and Keelin Murphy, “Deep learning for chest x-ray analysis: A survey,” *Medical Image Analysis*, p. 102125, 2021.
- [7] Yifan Mao, Feifei Xue, Ruixuan Wang, Jianguo Zhang, Wei-Shi Zheng, and Hongmei Liu, “Abnormality detection in chest x-ray images using uncertainty prediction autoencoders,” in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part VI*. 2020, vol. 12266 of *Lecture Notes in Computer Science*, pp. 529–538, Springer.
- [8] Behzad Bozorgtabar, Dwarikanath Mahapatra, Guillaume Vray, and Jean-Philippe Thiran, “SALAD: self-supervised aggregation learning for anomaly detection on x-rays,” in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part I*. 2020, vol. 12261 of *Lecture Notes in Computer Science*, pp. 468–478, Springer.
- [9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, “A simple framework for contrastive learning of visual representations,” 2020.
- [10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, “Momentum contrast for unsupervised visual representation learning,” 2020.
- [11] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi, “Big self-supervised models advance medical image classification,” 2021.
- [12] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng, “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” 2019.
- [13] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz, “Contrastive learning of medical visual representations from paired images and text,” 2020.
- [14] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” 2018.
- [15] Vikash Sehwal, Mung Chiang, and Prateek Mittal, “SSD: A unified framework for self-supervised outlier detection,” in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. 2021, OpenReview.net.
- [16] George Shih, C. C. Wu, S. Halabi, M. Kohli, L. Prevedello, T. Cook, Arjun Sharma, J. Amorosa, V. Arteaga, M. Galperin-Aizenberg, R. Gill, M. Godoy, Stephen Hobbs, J. Jeudy, A. Laroia, P. Shah, D. Vummidi, K. Yaddanapudi, and Anouk Stein, “Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia,” *Radiology. Artificial intelligence*, vol. 1 1, pp. e180041, 2019.
- [17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers, “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3462–3471.
- [18] Yu-Xing Tang, You-Bao Tang, Yifan Peng, Ke Yan, Mohammadhadi Bagheri, Bernadette A Redd, Catherine J Brandon, Zhiyong Lu, Mei Han, Jing Xiao, and Ronald M Summers, “Automated abnormality classification of chest radiographs using deep convolutional neural networks,” *npj Digital Medicine*, vol. 3, no. 1, pp. 1–8, 2020.
- [19] Yan Han, Chongyan Chen, Ahmed Tewfik, Ying Ding, and Yifan Peng, “Pneumonia detection on chest x-ray using radiomic features and contrastive learning,” in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 247–251.
- [20] Takahiro Nakao, Shouhei Hanaoka, Yukihiro Nomura, Masaki Murata, Tomomi Takenaga, Soichiro Miki, Takeyuki Watadani, Takeharu Yoshikawa, Naoto Hayashi, and Osamu Abe, “Unsupervised deep anomaly detection in chest radiographs,” *Journal of Digital Imaging*, pp. 1–10, 2021.
- [21] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He, “Accurate, large minibatch sgd: Training imagenet in 1 hour,” 2018.

Chapter 7

Serious Clinical Complication Detection using Contrastive Learning

Today's conventional monitoring techniques of patients during intensive treatment protocols for aggressive hematologic malignancies are cumbersome and often fail to detect life threatening complications early. Continuous monitoring of vital signs by means of medical wearables will potentially lead to an earlier diagnosis and better treatment. The major challenge is to extract clinically relevant information from a large amount of data provided by wearables. Oncologic treatment for patients with hematologic malignancies is associated with a high incidence of (post-)treatment complications, such as infections resulting in severe morbidity and mortality. In fact, nearly every patient on such treatment protocols experiences at least one serious clinical complication (SCC) requiring treatment. Early diagnosis of SCC is not only of high clinical relevance for the safety and well-being of the patients as it enables a more rapid treatment of SCC, but it would also potentially help reduce the number of hospitalisations. In this work, we aim to evaluate whether wearable-based monitoring enables detection of SCC with sufficient reliability. To do so, we take an anomaly or OOD detection method to identify whether defined episodes of vital signs are "regular" (= absence of SCC) or not in order to detect clinical complications.

7.1 Method

7.1.1 Learning Statistical Relevant Features

The problem of how to extract generalizing features from data – typically all what is semantically meaningful – is often difficult in practice and strongly depends on the particular machine learning task for what these features are needed. Generalizing features refer to those that are found in the training data but also in any example that comes from same distribution as the training data. For example, in case of object classification as computer vision task, the generalizing features are typically defined as those that are invariant under image transformations that keep the semantics of the shown object, such as horizontal flip, cropping, and slight changes in colouring. For the time series data used in this work, we define generalizing features as the information shared between two time series samples, \mathbf{x} and \mathbf{x}' , where each of the two samples (positive pair) consists of a randomly selected time interval of 1000 seconds (period) within the same hour. To extract these features, we map each time series sample, \mathbf{x} , to a d dimensional feature vector \mathbf{h} , with the help of a deep convolutional neural network, $\mathbf{h} = f_{\theta}(\mathbf{x})$, as a feature extractor. The network $f_{\theta}(\mathbf{x})$ is trained by a Self-Supervised Contrastive Learning objective, which approximately maximises the mutual information for all sampled positive pairs across all recorded hours.

7.1.2 Self-Supervised Contrastive Learning

To learn generalizable representations, self-supervised contrastive learning is used by ensuring that in the representation space embeddings of similar inputs are pulled closer while simultaneously embeddings from dissimilar inputs are pushed apart. The feature extractor, f_{θ} , is trained to extract the necessary information to discriminate similar samples, referred to as positive pairs, from dissimilar ones, referred to as negative pairs. Negative examples are not sampled explicitly; instead, given a positive pair, other examples within a mini-batch are treated as negative examples. The encoder, f_{θ} , maps the inputs to feature vectors in a d dimensional feature space where contrastive loss is applied, $\mathbf{h}_i = f_{\theta}(\mathbf{x}_i)$.

Let $\mathbf{h}_i = f_{\theta}(\mathbf{x}_i)$ and $\mathbf{h}_j = f_{\theta}(\mathbf{x}_j)$ be feature vectors of \mathbf{x}_i and \mathbf{x}_j and $(\mathbf{x}_i, \mathbf{x}_j)$ form a positive pair and $sim(\mathbf{u}, \mathbf{v}) := \frac{\mathbf{u}^T \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$ denote cosine similarity between feature vectors. In self-supervised contrastive representation learning, a contrastive loss can be defined as

$$\mathcal{L}_{i,j} = -\log \frac{\exp(sim(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{k \neq i} \exp(sim(\mathbf{h}_i, \mathbf{h}_k)/\tau)} \quad (7.1)$$

where $0 < \tau < 1$ is a scalar temperature parameter. The dataset of vital signs and activity data is represented as $\{\mathbf{X}_n\}_{n=1}^N$, with $\mathbf{X}_n \in \mathbb{R}^{D \times T}$, where N is the number of hours across all patients, D is the input dimension and T is the number of consecutive time points within one hour. We take $T = 3000$, which is less than the

expected $T = 3600$ seconds for an hour, as we frequently observed interruptions and therefore shorted T to keep most of the consecutive time series data. Let \mathbf{x} represent a window of time series of length w . To generate a positive pair, we randomly sample two windows \mathbf{x} and \mathbf{x}' of length 1000 time points with a time gap (500 – 1000 time points) from the same hour \mathbf{X}_n . Before passing the intervals through the encoder, f_{θ} , each input feature value is encoded with the respective quality index provided by the wearable device into a multi-dimensional vector space, where values with different quality scores are orthogonal to each other.

7.1.3 Score Function for SCC Detection

From the set of representations for the training examples, $\mathcal{D}_{train} = \{\mathbf{h}_m\}_{m=1}^M$, with $M = KN$, a score function can be defined to evaluate whether a given sample is SCC or not. For a given test sample, $\mathbf{h}_{test} = f_{\theta}(\mathbf{x}_{test})$, the cosine similarity is calculated to the nearest training sample in \mathcal{D}_{train} as a score for detecting SCC samples. The cosine similarity based SCC score is then defined as

$$SCC(x_{test}) := \frac{1}{K} \sum_{k=1}^K \left[1 - \max_{\mathbf{h}_m \in \mathcal{D}_{train}} sim(\mathbf{h}_m, \mathbf{h}_{test}^k) \right] \quad (7.2)$$

We take $K = 6$ and \mathbf{h}_{test}^k is randomly sampled from the same hour. The test sample, \mathbf{x}_{test} , is classified as SCC if the SCC score is above a threshold.

7.2 Training Details

The neural network f is a one-dimensional ResNet [79]. The input features were stored with sample rate of 1 Hz by the wearable device. The temperature, τ , in Eq. 7.1 was set as 0.07. Adam optimiser with initial learning rate of 10^{-3} and weight decay of 10^{-4} was used. The model was trained at batch size 128 for 500 epochs. For a set of B randomly chosen samples, the corresponding batch used for training consists of $2B$ pairs, 1 positive pair and $2B - 2$ negative pairs per sample. The encoder, f_{θ} , maps inputs to a 128 dimensional embedding. The outputs of this network are normalized to lie on a unit hypersphere, which enables using an inner product to measure the cosine similarity.

7.3 Detection and Prediction of Serious Clinical Complications with Wearable Based Remote Monitoring and Self-Supervised Contrastive Learning during Intensive Treatment for Hematologic Malignancies

Malte Jacobsen, Rahil Gholamipoor, Till A. Dembek, Pauline Rottmann, Marlo Verket, Julia Brandts, Paul Jäger, Ben Niklas Baermann, Mustafa Kondakci, Lutz Heinemann, Anna L. Gerke, Nikolaus Marx, Dirk Müller-Wieland, Kathrin Moellenhoff, Markus Kollmann, Melchior Seyfarth, Guido Kobbe. 2022.

Status: Submitted to *Lancet Digital Health*.

Contributions: The author contributed with the methodology, implementation of the anomaly detection algorithm, evaluation, visualization and describing the algorithm under the supervision of Prof. Dr. Markus Kollmann.

Detection and Prediction of Serious Clinical Complications with Wearable Based Remote Monitoring and Self-Supervised Contrastive Learning During Intensive Treatment for Haematological Malignancies

Malte Jacobsen MD^{1,2}, Rahil Gholamipoor MSc³, Till A. Dembek MD⁴,
Pauline Rottmann⁵, Marlo Verket MSc², Julia Brandts MD², Paul Jäger MD⁵,
Ben-Niklas Bärman⁵, Mustafa Kondakci MD⁶, Lutz Heinemann PhD⁷, Anna L. Gerke⁵,
Nikolaus Marx MD², Dirk Müller-Wieland MD², Kathrin Möllenhoff PhD⁸,
Markus Kollmann PhD⁹, Melchior Seyfarth MD^{1,10}, Guido Kobbe MD⁵

¹Faculty of Health, University Witten/Herdecke, 58448 Witten, Germany;

²Department of Internal Medicine I, University Hospital Aachen, RWTH Aachen University, 52074 Aachen, Germany;

³Department of Computer Science, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany;

⁴Department of Neurology, Faculty of Medicine, University of Cologne, 50937 Cologne, Germany;

⁵Department of Haematology, Oncology, and Clinical Immunology, University Hospital Düsseldorf, Medical Faculty, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany;

⁶Department of Oncology and Haematology, St. Lukas Hospital Solingen, 42697 Solingen, Germany

⁷Science-Consulting in Diabetes, 41564 Kaarst, Germany;

⁸Mathematical Institute, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

⁹Department of Biology, Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany;

¹⁰Department of Cardiology, Helios University Hospital of Wuppertal, 42117 Wuppertal, Germany.

Conflict of interest

The author declares that there is no conflict of interest. This study was an Investigator Initiated trial. This research received funding from the internal grant program (PhD and Dr. rer. nat. Program Biomedicine) of the Faculty of Health at Witten/Herdecke University, Germany and by a grant from the Leukämie Lymphom Liga e.V, Germany.

Disclosures

This study used the wearable 'Everion'-Device provided by Biovotion AG (now: biofourmis AG), Switzerland. Biovotion did not provide any financial support for the research and had no impact on writing of the manuscript. Biovotion did not participate in the analysis of the data or influence the conclusions. The funding source had no involvement in conducting the study.

Registration

The study was registered in the German clinical trials register (DRKS00014782).

Abbreviations:

ADL – Activities of daily living; AML – Acute myeloid leukaemia; ALL – Acute lymphocytic leukaemia; BMI – Body Mass Index; BP – Blood pressure; CE – Conformité Européenne; CI – Confidence interval; CML – Chronic myeloid leukaemia, CLL – Chronic lymphocytic leukaemia; CTCAE – Common Terminology Criteria for Adverse Events; ECG – Electrocardiogram; GGT – Gamma-Glutamyltransferase; HL – Hodgkin lymphoma; IC – Inpatient cohort; IV – Intravenous; MDS – Myelodysplastic syndrome, MPN – Myeloproliferative neoplasms; NHL – Non-Hodgkin lymphoma; OC – Outpatient cohort; PMF – primary myelofibrosis; RPM – Remote patient monitoring; SCC – Serious clinical complication; TPN – Total parenteral nutrition; ULN – Upper limit of normal

Keywords

Artificial intelligence, neural network, wearables, oncology, remote patient monitoring, continuous vital sign monitoring, complication management

Acknowledgement:

We would like to thank E. Hein, K. Graf, A. Caduff, the clinical staff of the Department of Oncology and Clinical Immunology and the "Koordinierungszentrum für Klinische Studien" (KKS) at the University Hospital Düsseldorf, the Heine Center for Artificial Intelligence and Data Science (HeiCAD), as well as the Leukämie Lymphom Liga e.V., Düsseldorf, Germany for their support.

Abstract

BACKGROUND: Serious clinical complications (SCC; CTCAE grade ≥ 3) occur frequently in patients treated for haematological malignancies. Early diagnosis and treatment of SCCs are essential and improve outcomes. This study assessed an AI-model-derived SCC-Score to detect and predict SCCs from time series data recorded continuously by a medical wearable.

METHODS: In this single-arm, single-centre observational cohort study, vital signs and physical activity were recorded by a wearable for a total of 30,182 hours in 79 patients (54 Inpatient Cohort (IC) / 25 Outpatient Cohort (OC)) receiving therapy for haematological malignancies. Hours with normal physical functioning without evidence of SCCs ('regular') were presented to a deep neural network to train a self-supervised contrastive learning model to extract features from the time series that are typical in regular periods. The model calculated a so-called SCC-Score based on dissimilarity to regular features (higher values indicate higher risk for a SCC). Detection and prediction performance of the SCC-Score compared to the clinical documentation of SCCs was evaluated by employing the area under the receiver operating characteristic curve (AUROC).

FINDINGS: Hundred-twenty-four clinically documented SCCs occurred in the IC and 16 in the OC. Detection of SCC by the AI model was achieved in the IC with a sensitivity 84.8% and specificity 66.6%, providing an AUROC of 0.80 (95%CI 0.78-0.83) (OC: 93.2%/46.9% (0.75 (0.68-0.81))). Up to 2 days prior to the clinical diagnosis, prediction of infectious SCC by the SCC-Score was possible (AUROC 0.86 at -24 hours and 0.81 at -48 hours).

INTERPRETATION: This study shows a reliable detection and prediction of SCCs in patients receiving therapy for haematological malignancies using an AI model. This encourages further exploration of remote patient monitoring by a wearable to enable pre-emptive complication management.

FUNDING: This study was funded by the Leukämie Lymphom Liga e.V.

Research in context

Evidence before this study

PubMed, EMBASE and MEDLINE were searched for randomised and non-randomised clinical trials, observational studies published before 03.01.2022 with no language restrictions. MeSH and free terms related to the disease terms ('malignancy'/'oncology'/'cancer'/'tumor') and ('wearable'/'wearable sensor'/'biosensor'/'actigraphy'/'accelerometry'/'smart watch') were included, or variations of these terms thereof. Articles were screened by title and abstract to identify relevant studies. Reference lists of eligible articles were also searched for additional studies.

There is limited evidence for applying medical wearables to detect complications in patients with malignancies. Thirty-six studies demonstrated that wearable-based assessment of physical activity levels is correlated to relevant clinical outcomes, such as symptom burden and hospitalisations. To date, linking vital signs recorded by a wearable for detection of complications is missing. Currently, detection and management of complications in patients receiving treatment for haematological malignancies in the outpatient setting rely on their self-assessment and specialist evaluation at high-frequency clinical visits. In the inpatient setting, detecting complications relies on clinical examinations and diagnostics, such as daily laboratory sampling. For Health Care Professionals (HCP), clinical scores such as the Multinational Association for Supportive Care in Cancer Score allow for a priori risk stratification for complications during oncological therapies. However, no systematic tools are available to facilitate on-treatment surveillance in these patients, and the frequency with which treatment-related complications occur is largely determined by the patient's condition and the selected treatment approach.

For patients, the occurrence of potentially life-threatening complications represents a high psychosocial and psychological burden. Hence, early detection of treatment-associated complications in patients with haematological malignancies is challenging for patients and HCP.

Therefore, the development and application of AI-enhanced wearable-based remote patient monitoring (RPM) could improve detection and management of clinical deterioration.

Added value of this study

This study suggests that wearable-based RPM in combination with AI analytics enables personalized detection and prediction of serious clinical complications in patients with haematological malignancies during treatment. The applied pragmatic trial design provides a large data set of different vital signs and physical activity in this patient population, parallel to extensive clinical documentation of complications. To our knowledge, this is the first trial evaluating a wearable-based monitoring approach to detect clinical complications in patients with haematological malignancies.

Implications of all the available evidence

In summary, our results indicate that patients and HCPs with haematological malignancies can benefit from RPM with a medical wearable in combination with a suitable analytics model that can identify subtle and early symptoms. Usage of such a non-obtrusive approach in clinical practice might also allow to optimise complication management, i.e. reduce the workload of specialised health care professionals while also improving patient care. To realise the full potential of wearable-based RPM, the ability for real-time detection of complications needs to be shown in an adequately designed prospective study.

Background

Treatment of patients with haematological malignancies is associated with a high incidence of clinical complications, such as infections, cardiac events and immunologic dysregulations.^{1,2} These potentially life-threatening complications require early recognition and therapeutic intervention, as it is known that delayed intervention is associated with increased morbidity and mortality.^{3,4} Recent diversification of oncological treatment options, including e.g. CAR-T cell therapy, increase therapeutic options but add to the spectrum of complications, such as 'cytokine release syndrome'. Today's management of complications depends on the setting of oncological treatment: Under hospital conditions (= inpatients), the management of complications relies on intermittent recordings of vital signs, daily clinical examinations, and laboratory tests by health care professionals (HCP). An increasing number of oncological treatments are applied in the outpatient setting.⁵ There, complication detection relies primarily on patient self-assessment.⁶ Early on detection of (subtle) symptoms indicating complications is challenging and is often delayed. To avoid 'late show ups', outpatients are routinely admitted to their treatment centre without evidence of complications, which burdens both patients and HCP.⁷ Therefore, there is a need for innovative concepts for early and reliable detection of treatment-associated complications.⁸

Remote patient monitoring (RPM) with medical wearables represents a novel option for non-invasive and continuous real-time monitoring of vital signs and physical activity.^{9,10 11} Medical wearables provide longitudinal and high-resolution health data that expand monitoring options.^{12,13} Ideally, the large data sets recorded by wearables should be combined among patients to increase the statistical power of the prediction algorithm that is used to evaluate the patient specific vital signs in real-time. The challenge is to resolve the seemingly contradicting situation of employing a single prediction algorithm that at same time can make different predictions for similar vital signs, if these vital signs originate from different patients. Such personalized predictions under a single prediction model can be realised for a very large number of patients by making use of recently developed concepts of self-supervised deep learning.

Aim of this study was to evaluate if a wearable-based RPM approach in combination with AI analytics is able to detect and predict complications with sufficient reliability in in- and outpatients during their oncological treatment for haematological malignancies.

Methods

Study design and setting

This was an open-label, single-arm, single-centre, investigator-initiated cohort study covering patients with a haematological malignancy receiving an oncological treatment (chemotherapy alone or in combination with radiotherapy and/or haematopoietic stem cell transplantation) (figure supplementary 1). Details of the study have been described elsewhere.¹⁴ In brief, the study was conducted at the Department of Haematology, Oncology and Clinical Immunology of the University Hospital Düsseldorf, Germany. The study was approved by the Ethics Committee of the Medical Faculty of the Heinrich Heine University Düsseldorf and was registered in the German clinical trials register (DRKS00014782).

Participants

Inclusion criteria were patients' age ≥ 18 years and an indication for a treatment protocol with expected haematotoxicity according to Common Terminology Criteria for Adverse Events (CTCAE) grade 4 alone or in combination with stem cell transplantation. Exclusion criteria were medical or mental conditions impairing the ability to continuously wear the wearable (e.g. dementia, upper arm tattoos, skin diseases) and active implants, which might impair recordings.

All patients provided written informed consent. During visits, the following data were obtained for each individual (e.g. medical history, comorbidities, symptoms, physiological parameters, laboratory values, results of physical examination). A convenience sample of 79 patients was recruited for an intensive treatment protocol: 54 patients were treated in hospital [inpatient cohort (IC)] and 25 patients for an outpatient-based treatment [outpatient cohort (OC)] (table supplementary 1-2).

Patients and clinical staff were blinded for wearable data (analysed retrospectively). Prior to study participation, patients were informed that they would not derive immediate individual benefit from study participation.

Wearable

The commercially available wearable (Everion, Biovotion AG, Switzerland) employed is a CE marked medium-risk device (class IIa) according to the Directive 93/42/EEC (firmware used was for clinical investigation only). Different sensors implemented in this wearable are used for non-invasive monitoring of vital signs and physical activity (e.g. photoplethysmography, temperature probe, accelerometer). Longitudinally recorded parameters, such as heart rate, temperature, respiratory rate and physical activity, and if applicable, respective quality indices were calculated with proprietary methods implemented in the firmware (table supplementary 3). Raw signals were acquired with a frequency of >30 Hz, calculated parameters were stored with a rate of 1 Hz (= 3,600 data points/hour). The battery of the wearable had to be recharged daily for 90 min.

Two wearables for alternate use were assigned to each patient at the baseline visit prior to starting treatment to enable continuous wearable-based monitoring of vital signs and physical activity in these patients. Frequency of subsequent study visits (app. every 90 hours for device swap) was determined by the limited data storage capacity of the wearable.

Identification of Serious Clinical Complications based on clinical documentation

Non-haematological SCCs were defined by meeting the criteria of CTCAE (v4.03) grade ≥ 3 .¹⁵ Clinical documentation (visit entries, laboratory results, diagnostic results) was independently and retrospectively reviewed by two investigators (PR, MJ) for the occurrence of SCC. For each clinically documented SCC, a starting time point was noted. Infectious SCCs with no focus of origin were classified as 'Infections and infestations – other'. Recovery from a SCC was defined as absence of documented clinical symptoms, absence of pathological laboratory and diagnostic results. With respect to varying time courses of different types of SCCs, e.g. a hypertensive crisis with a rapid onset vs. an infection, which develops over several hours/days.

Input data for AI model

Time series input data recorded by the wearable in patients of the total cohort, the IC and OC were presented separately to the AI model. No predefined quality constraints were used, as the trained AI model is able to extract features of informational value. Data sets were split into hours according to their timestamps, and only hours with $\geq 3,000$ data points were included to ensure enough input data and discriminative information within each hour. The investigators annotated the hours (figure 1) as 'regular' if no SCC was recorded in the clinical documentation. Hours with documented SCC were annotated as 'non-regular', with a special annotation for infectious SCCs for subgroup analyses.

Since changes in vital signs and physical activity may already occur before SCC criteria are fulfilled, e.g. a hospital visit or lab results, a time buffer was introduced, i.e. recordings in the 48 hours prior to the timestamp of SCC onset and 24 hours post-recovery from a SCC were annotated as 'non-regular'.

AI model

For this study, a self-supervised contrastive learning approach was used, a subset of unsupervised learning in AI technology that learns to extract generalising features from complex data.¹⁶ Generalizing features are features that are specific to the distribution of patients where the training set has been derived from and as a consequence are useful for any test example drawn from the same distribution. Details of the AI model development and training are given in the supplementary material. The features were extracted using a deep neural network that was trained on a dataset generated by randomly collecting 90% of the 'regular' hours for each patient.¹⁷ Remaining 10% of the 'regular' hours were used to establish a null-distribution for statistical testing. These 'regular' and 'non-regular' hours (together: 'test set') were used for evaluating the sensitivity of the approach to detect SCCs. After testing,

Following training the AI model, it is able to identify deviations from the 'regular' hours data set and therefore detect SCCs.

AI model output

The generalizing features for an hour of vital signs and physical activity are represented as high dimensional vector of unit length, for which a cosine similarity score can be calculated between them by taking the scalar product. For each hour of vital signs and physical activity in the test set, the highest similarity to a 'regular' hour in the training set was identified. A patient specific evaluation can be realized by narrowing down the best-match search to the training hours of one patient. Based on this best-match of similarity scores, we introduce the 'SCC-Score', which is defined as one minus the highest cosine similarity to the training set (figure 1 and Methods supplementary). SCC-Scores ranged from zero to one. A higher SCC-Score indicates a larger deviation from what is expected to be a 'regular' hour. SCC-Scores for the 10% of 'regular' hours (which were not shown to the AI model during training) are used to establish the null-distribution under the null hypothesis that an hour is 'regular'. Therefore, the null hypothesis would be rejected for any hour with a SCC-Score above a pre-specified significance level, with the result that the assessed hour is classified as 'non-regular'. The significance level has to be pre-specified to meet clinical requirements and can be understood as a 'decision boundary'. It can be obtained from calculating the corresponding quantile of the null distribution. The SCC-Scores of the test set were evaluated per hour but also per day to address inter-hourly variability.

Outcomes and statistical analysis

Primary outcomes were the detection and prediction of clinically documented SCCs by the SCC-Score. Subgroup analysis evaluated infectious SCCs. For statistical analysis, differences between means of hours annotated as 'regular' and 'non-regular' obtained from SCC_{IC} -Score, SCC_{OC} -Score and SCC_{Total} -Score were tested for significance using a two-sided t-test, adjustment for multiple comparisons was performed by using Bonferroni correction. To address overfitting, ten-fold cross-validation was performed. Statistical significance was tested by an ANOVA between the cross-validation splits of 'regular' and 'non-regular' (figure supplementary 3). Receiver Operating Characteristics analysis (Area under the curve of the ROC-analysis (AUROC)) was computed to assess primary outcomes. The cut-point that optimises the detection of true-positive results (sensitivity) and false-positive results (1-specificity) is reported by the Youden index. For clinical requirements (not missing a SCC), specificity was reported at a sensitivity of approximately 95%.

To assess SCC-Score prediction capabilities for infectious SCC (table 1), the performance of the score in the 120 hours prior to and after the time stamp of diagnosis ($t = 0$ hour) were analysed. For AUROC-analysis, 95% confidence intervals (CI) are reported.¹⁸ A p-value <0.05

was considered significant. For data and statistical analysis, an open-source software tool was used (Python, version 3.6.5).

Findings

A total of 140 SCCs were extracted from the clinical documentation of 79 patients in the two cohorts (table 1; data of two patients without 'regular' hours and early study withdrawal were excluded from further analysis): Cumulative incidence of SCCs in the IC was 90.7% and of those in the OC was 48.0%. More than one SCC occurred in 30 patients in IC and 3 patients in OC. Most SCCs occurred within the first 15 days after starting treatment (IC 82.2%, OC 93.8%, figure supplementary 4). Infectious SCCs accounted for 65.0% of the total SCCs and were the most frequent SCCs in both cohorts (IC 63.7%, OC 75.0%).

Wearable data were recorded for 24,100 hours in the 54 patients in the IC (figure supplementary 2), the median recording time per patient was 457.4 (IQR 324.3-538.5) hours. The 25 patients in the OC had 7,215 hours total recording time, with a median participation time of 315.5 (227.4-340.8) hours per patient. Hours meeting data constraint (>3,000 data points) were 23,227 hours (96.4%) in the IC and 6,955 hours (96.3%) in the OC.

SCC-Scores were significantly higher in 'non-regular' hours and days, indicating a higher risk for SCC prevalence in comparison to sets of 'regular' hours (table 2). This observation was stable with ten-fold cross-validation (figure supplementary 3). The mean SCC-Score levels differed between patients in the IC and OC, i.e. the average scores for 'regular' days were 0.197 ± 0.052 and 0.176 ± 0.045 for IC and OC, respectively. For the infectious SCC-Score, the absolute difference was even more pronounced (0.169 ± 0.047 in IC and 0.122 ± 0.035 in OC). The SCC-Scores for different types of SCCs differ (table 1; last column), with scores from 0.114 for immune system disorders (n=1) to 0.290 for paroxysmal atrial tachycardia (n=3). For the most common SCCs ('Infections and infestations – other'), the mean score was 0.249 (n=66). Training of the AI model with different numbers of data sets increased the AUROC (figure supplementary 5).

AI model application to the test data set of 'regular' hours and 'non-regular' hours of accumulated 12,128 hours for SCC_{IC} , 2,227 hours for SCC_{OC} , and SCC_{Total} revealed a significant mean difference in the SCC-Scores (table 2 and figure 2).

At the Youden index, sensitivity for detecting 'non-regular' hours was 69.0% for IC and 57.0% for OC. Specificity for those hours for IC and OC was 63.5% and 69.4%, respectively. With SCC-Scores applied to 'hours for testing', an AUROC for IC of 0.72 and for OC 0.68 was observed (figure 2). Calculating a mean daily score showed increased sensitivity (IC 84.8% and OC 93.2%). Specificity in IC was stable at 64.9%, while in OC, specificity decreased to 46.9%. Assessment of hourly SCC-Score for the infectious SCCs showed equivalent performance in terms of sensitivity, specificity and AUROC. The highest AUROC (IC 0.83, OC 0.82) was observed with detecting days containing infectious SCCs. Reporting specificity at a sensitivity of approximately 95.0% showed lower specificity (IC 23.9% and OC 19.4%) with

hourly SCC-Scores. When reporting specificity at approximately 95·0% sensitivity on daily SCC-Scores, a moderate decline in specificity was observed (IC 51·7% and OC 41·6%).

To evaluate the SCC-Score's prediction capabilities, the clinical diagnosis of infectious SCC was set as 0 hour. The SCC-Scores calculated by the AI model for the hours before and after each SCC showed an increase in Score starting up to 48 hours prior to diagnosis (figure 3a and table supplementary 4). An equivalent course was observed in IC and OC; however, the overall SCC-Score level differed between the two cohorts. The decline in SCC-Score is reflected by an increase in AUROC over the same time periods with both cohorts and a subsequent less steep slope in the 120 hours post diagnosis (figure 3b). The observed change in AUROC in the two days prior to clinical SCC diagnosis may allow prediction of infectious SCCs.

Interpretation

Results

Wearable-based RPM combined with an AI model enables calculation of a SCC-Score that allows detection and prediction of SCCs in patients receiving intensive treatment for haematological malignancies up to 48 hours before clinically documented SCCs were diagnosed. This study can be regarded as a 'proof of principle' for wearable-based RPM during oncological treatment where patients are at high risk for life-threatening complications.

As to be expected, the different SCCs observed in this study were heterogeneous in type and severity. As the course over time of the SCC is diverse, the induced changes in vital signs and physical activity vary to a different degree. For example, an infection may develop over the course of hours and days, whereas a hypertensive crisis or cardiac arrhythmias can both occur and resolve itself from one moment to the other. The SCC-Scores represent this diversity, i.e. allergic skin rash scored lower than infections and arrhythmia.

The overall levels of the SCC-Scores observed for 'regular' hours and 'non-regular' hours in the two cohorts (IC vs OC) were different; however, relative recording times in both cohorts were comparable.¹⁴ Yet, the cause for this difference is not clear; it might reflect the higher physical activity levels of patients in the OC or the different oncological treatment conditions. However, the degree of change in the SCC-Score induced by SCCs is similar in IC and OC. This results in a comparable AUROC analysis outcome (table 2), pronouncing the robustness of our method. Further improvement in performance was achieved by applying daily SCC-Score instead of hourly. Given the different levels in SCC-Scores between the cohorts implicates the necessity to record data in a precise clinical context.¹⁹ For clinical application, an automated SCC detection based on daily SCC-Scores would be a convenient RPM tool for the HCP.

In the subgroup analysis for infectious SCC, the SCC-Score showed an increase prior to the clinical diagnosis of the SCC (at t=0 hour) in both cohorts, with a steeper slope prior to the diagnosis than the decrease in the hours post-diagnosis (figure 3). This increase could be driven by the uninhibited evolvement of an infection, whereas the decline is probably associated with the therapeutic intervention initiated. This phenomenon allows for the speculation that treatment success of a SCC or failure may also be tracked by RPM.

Methods

In line with the aim to assess changes in the recorded vital signs and physical activity induced by SCCs, 'regular' and 'non-regular' hours during treatment were compared. Therefore, in contrast to other studies, pre-treatment recordings were omitted, as it can be proposed that vital signs and physical activity differ relevantly between pre-treatment and during treatment, even in the absence of SCC.²⁰

Training of the AI model with 'Big Data' improves its performance to a given extent, i.e., with the used amount of data sets, the slope of the performance improvement in AUROC decreases (figure supplementary 5).

A relevant advantage of the self-supervised contrastive learning AI model is that it is able to extract informational value from large and noisy data sets with artefacts and data gaps.

Furthermore, the presented self-supervised approach is data-efficient, as small quantities of recordings of a given patient in a diverse data set are sufficient to calculate a SCC-Score. Usually, individual detection of SCCs, based on a risk score for a given patient, requires a large quantity of data from this specific patient to first train an individual AI model. For real-world adaption the trained AI model can be implemented on a smartphone, as the computationally demanding training of the AI model can be done remotely.

Patients' individual responses in vital signs and physical activity to SCC of any kind are unknown. Instead of rigid thresholds for a single parameter, the totality of provided vital sign and physical activity measures was used to calculate a similarity score. Applying this SCC-Score to unseen 'regular' and 'non-regular' hours, mean values differed significantly between both groups.

From a clinical point of view, it is desirable to minimise the risk of missing SCCs. Therefore, the decision boundary was set corresponding to a sensitivity of 95%. Of note, this choice is somewhat arbitrary and needs to be discussed according to clinical context.²⁰ Depending on the situation under consideration and prior knowledge (e.g. given by a pre-test probability), clinicians can individually choose the decision boundary such that a certain balance of sensitivity and specificity is achieved. This decision boundary, which is directly related to the significance level of the statistical test, may also be adapted during real-world application when more information becomes available.²¹ In general, the SCC-Score calculated represents a single value that can be translated into actionable clinical information.

Generalisability

In the future, automated SCC detection by a wearable-based RPM in clinical oncology offers the option of permanent patient surveillance and may thereby improve complication management. Ideally, recorded data would be analysed in real-time to provide actionable information for early and effective treatment. This may improve clinical pathways, e.g. implementation of demand-driven visits, which could reduce physicians' and nurses' workload in specialised clinics.²² Furthermore, a decrease in the frequency of blood sampling during treatment of patients for their haematological malignancy is possible as recent research indicated a good correlation of wearable recorded vital signs with laboratory measurement results.²³ This approach may reduce treatment and disease burden by enabling optimal timing of interventions to counter SCC.

Limitations

The sample size evaluated in this exploratory study is limited; however, to our knowledge, this is the largest trial employing RPM in patients treated for haematological malignancies.¹⁰ Comparability between IC and OC may be limited due to the unequal number of patients in both cohorts. However, no major performance differences were observed.

Limitations of the wearable used in this study are described elsewhere.¹⁴ In terms of suboptimal data availability concerning the participation time of patients in the study, one can foresee that more convenient wearables with longer-lasting battery life will lead to higher practicability of the approach.

Grading of SCC with CTCAE grade ≥ 3 may influence vital signs and physical activity differently. Using this grading threshold for SCC omits lower grade complications, which, however, may already be of therapeutic relevance and affect patient's wellbeing. Nevertheless, detecting severe SCCs as a primary medical need was chosen because CTC grade ≥ 3 definitely requires medical intervention. Not all SCCs may affect vital signs and physical activity to the degree that they are likely to be detected by a wearable-based RPM approach; infection-induced SCCs might lead to a stronger 'signal' than some other SCCs and may therefore be an ideal target for RPM. However, it is unclear which sets of parameters are required for optimal SCC detection. This question must be addressed in subsequent evaluations.

Annotation of data recorded by the wearable during 'regular' hours was based on clinical SCC documentation, and there is a probability that SCCs weren't documented properly. Consecutively, the AI model would be trained on incorrectly annotated input data. Adding a time buffer to any clinical documentation of SCC time to avoid inadequate training of the AI model was a pragmatic approach; however, it is not obvious what the optimal buffer duration is given the heterogeneity of SCCs.

Conclusion

In summary, this study provides proof of principle that SCCs in a vulnerable patient population of patients receiving treatment for haematological malignancies can be detected and predicted with an innovative approach, based on continuously recorded wearable data combined with a self-supervised AI model. Prospective confirmatory studies are needed to document the clinical benefit of this approach in clinical practice.

Table 1

Serious clinical complications (SCC) based on adverse events classification (Common Terminology Criteria for Adverse Events v4.0 (2009)) sorted in order of frequency of occurrence in Inpatient cohort (IC), Outpatient cohort (OC), and Total. Exemplary SCC_{Total}-Score (last column) for the respective SCC.

No.	Adverse event	Criteria for Grade 3 in Common Terminology Criteria for Adverse Events	SCC [n]			SCC _{Total} -Score
			IC	OC	Total	
1	Infections and infestations - other*	Severe or medically significant but not immediately life threatening; hospitalization or prolongation of existing hospitalization indicated; disabling; limiting self-care ADL	55	11	66	0.249
2	Lung infection*	IV started	10	1	11	0.260
3	Hypertension	Stage 2 hypertension [...]; medical intervention indicated; [...]	10	1	11	0.254
4	Mucositis oral*	Severe pain; interfering with oral intake	11	0	11	0.225
5	Nausea	Inadequate oral caloric or fluid intake, TPN	9	0	9	0.207
6	Pulmonary oedema	diuretics indicated	3	0	3	0.239
7	Sinus tachycardia	Urgent medical intervention indicated	3	0	3	0.258
8	Allergic reaction	Prolonged [...] and/or brief interruption of infusion	3	0	3	0.272
9	Pain	Severe pain; limiting self-care ADL	3	0	3	0.200
10	Paroxysmal atrial tachycardia	IV medication indicated	1	2	3	0.290
11	Hypotension	Medical intervention or hospitalization indicated	2	0	2	0.253
12	Dyspnoea	Shortness of breath at rest; limiting self-care ADL	2	0	2	0.234
13	Diarrhoea	Increase of ≥ 7 stools per day over baseline	2	0	2	0.254
14	Syncope	Fainting; orthostatic collapse	2	0	2	0.209
15	Periorbital oedema	Diuretics indicated	1	0	1	0.206
16	Oral pain	Severe pain; limiting self-care ADL	1	0	1	0.244
17	Colitis*	Severe abdominal pain [...]; medical intervention indicated; peritoneal signs	1	0	1	0.207
18	Hypokalaemia	$< 3.0 - 2.5$ mmol/L; hospitalization indicated	1	0	1	0.214
19	Immune system disorders - Other	Severe or medically significant but not immediately life-threatening; hospitalisation [...]	1	0	1	0.114
20	Cholecystitis*	Severe symptoms; radiologic, endoscopic or elective operative intervention indicated	1	0	1	0.155
21	Catheter related infection*	IV antibiotic, antifungal, antiviral, radiologic or operative intervention indicated;	1	0	1	0.202
22	Hepatobiliary disorders - Other	Severe or medically significant but not immediately life-threatening; hospitalization [...]	1	0	1	0.241
23	GGT increased	$> 5.0 - 20.0$ x ULN	0	1	1	0.226

ADL: Activities of Daily Living; IV: Intravenous; TPN: Total Parenteral Nutrition; BP: Blood Pressure; GGT: Gamma-Glutamyl transferase; ULN: Upper Limit of Normal; *Grouped as 'infectious SCC'; [...] left out for visualisation

Table 2

SCC-Scores of 'hours for testing' containing 'regular' hours and 'non-regular' hours are reported. These hours were previously unseen by the AI model. Differences in mean SCC-Scores in the respective cohorts (SCC_{IC} , SCC_{OC} , SCC_{Total}) between 'regular' hours and 'non-regular' hours/days and respective p-values (two-sided t-test) are reported. To account for multiple testing, Bonferroni correction was applied and the significance level set to $0.05/12=0.0042$. Performance indicators (at Youden Index) of the SCC-Score were calculated for detection of SCCs, and infectious SCCs in patients in the IC, OC, and Total separated for hours/days. In addition, specificity is reported at a sensitivity of approximately 95% to ensure a high ratio of SCC detection. AUROC of the SCC-Scores are given in the last column (with 95% confidence interval).

Interval	Type	Model	'regular' hours / days [n]	'non-regular' hours / days [n]	SCC-Score 'regular' hours [mean±SD]	SCC-Score 'non-regular' hours [mean±SD]	P-Value	Sensitivity / Specificity [%]	~95% Sensitivity / Specificity [%]	AUROC (95% CI)
Hourly	SCC	IC	1,234	10,729	0.196±0.060	0.246±0.061	<0.0001	69.0 / 63.5	95.0 / 23.9	0.72 (0.71- 0.74)
		OC	526	1,701	0.176±0.056	0.223±0.073	<0.0001	57.0 / 69.4	95.2 / 19.4	0.68 (0.66- 0.71)
		Total	1,760	12,430	0.190±0.066	0.245±0.069	<0.0001	60.3 / 71.0	95.0 / 24.4	0.72 (0.71- 0.73)
	infectious SCC	IC	1,474	8,251	0.171±0.056	0.222±0.060	<0.0001	67.1 / 67.4	95.0 / 23.0	0.74 (0.73- 0.75)
		OC	541	1,553	0.123±0.042	0.162±0.059	<0.0001	63.2 / 70.6	95.1 / 19.2	0.71 (0.68- 0.73)
		Total	2,015	9,804	0.155±0.050	0.199±0.055	<0.0001	72.9 / 60.4	95.0 / 25.4	0.73 (0.72- 0.74)
Daily	SCC	IC	578	604	0.197±0.052	0.245±0.036	<0.0001	84.8 / 66.6	95.0 / 51.7	0.80 (0.78- 0.83)
		OC	245	88	0.176±0.045	0.221±0.044	<0.0001	93.2 / 46.9	95.5 / 41.6	0.75 (0.68- 0.81)
		Total	823	692	0.191±0.057	0.244±0.043	<0.0001	83.1 / 64.0	95.1 / 46.3	0.78 (0.76- 0.81)
	infectious SCC	IC	682	470	0.169±0.047	0.222±0.037	<0.0001	85.2 / 67.3	95.1 / 50.9	0.83 (0.80- 0.85)
		OC	245	80	0.122±0.035	0.161±0.033	<0.0001	92.5 / 58.8	95.0 / 51.8	0.82 (0.76- 0.88)
		Total	927	550	0.154±0.042	0.200±0.034	<0.0001	85.1 / 69.2	95.3 / 51.1	0.83 (0.80- 0.85)

Figure 1

Development of an AI model for calculation of a SCC-Score (serious clinical complication). (A) Time series of vital signs and physical activity recorded by a medical wearable. (B) Clinical documentation, such as patient charts or laboratory results, that were reviewed for identifying SCC events. (C) According to the clinical documentation, the hours without evidence of SCCs were annotated as 'regular' hours, the remaining hours were regarded as 'non-regular'. (D) 'Regular' hours for each individual patient were randomly split into two datasets: 90% for training and 10% for testing and generating a null-distribution. For cross-validation, the splitting was repeated ten times. For training the AI model, the 'regular' hours were presented to a deep neural network as part of a self-supervised contrastive learning objective. A SCC-Score based on the similarity between a test hour and the closest 'regular' hour from the training set was calculated. (E) A null-distribution of SCC-Scores from 'regular' hours not used in training was established. (F) For a given hour a statistical test under the null-distribution was applied to detect SCCs, with significance level selected on clinical requirements.

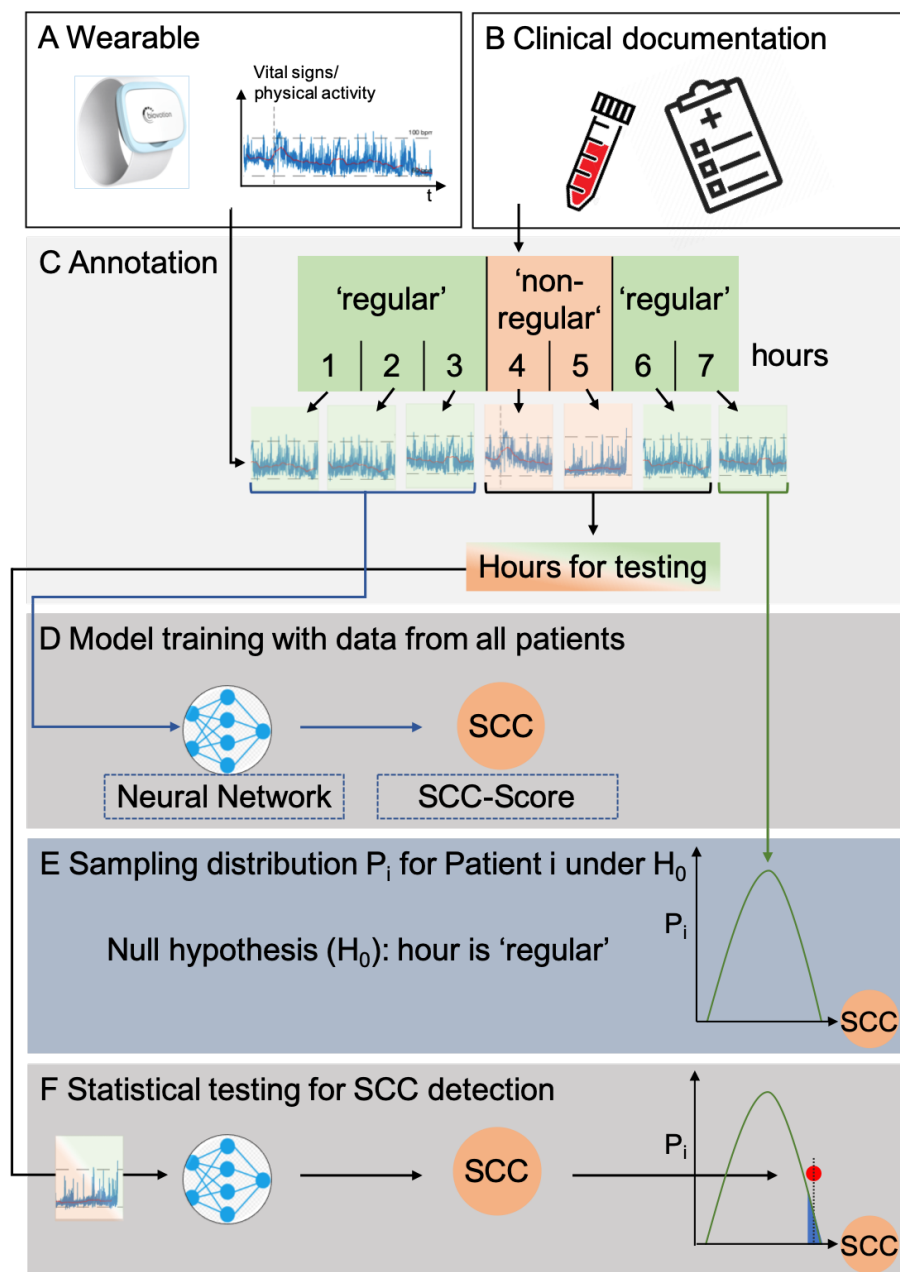


Figure 2

Area under the receiver operating characteristic curves (AUROC) for the SCC-Score for the three cohorts (SCC_{IC} (blue line), SCC_{OC} (green line) and SCC_{Total} (red line)) separately for SCC Hourly and Daily. The same analysis was performed for infectious SCC. The dots mark the cut-point that optimises the detection of true positive results and false-positive results (Youden index).

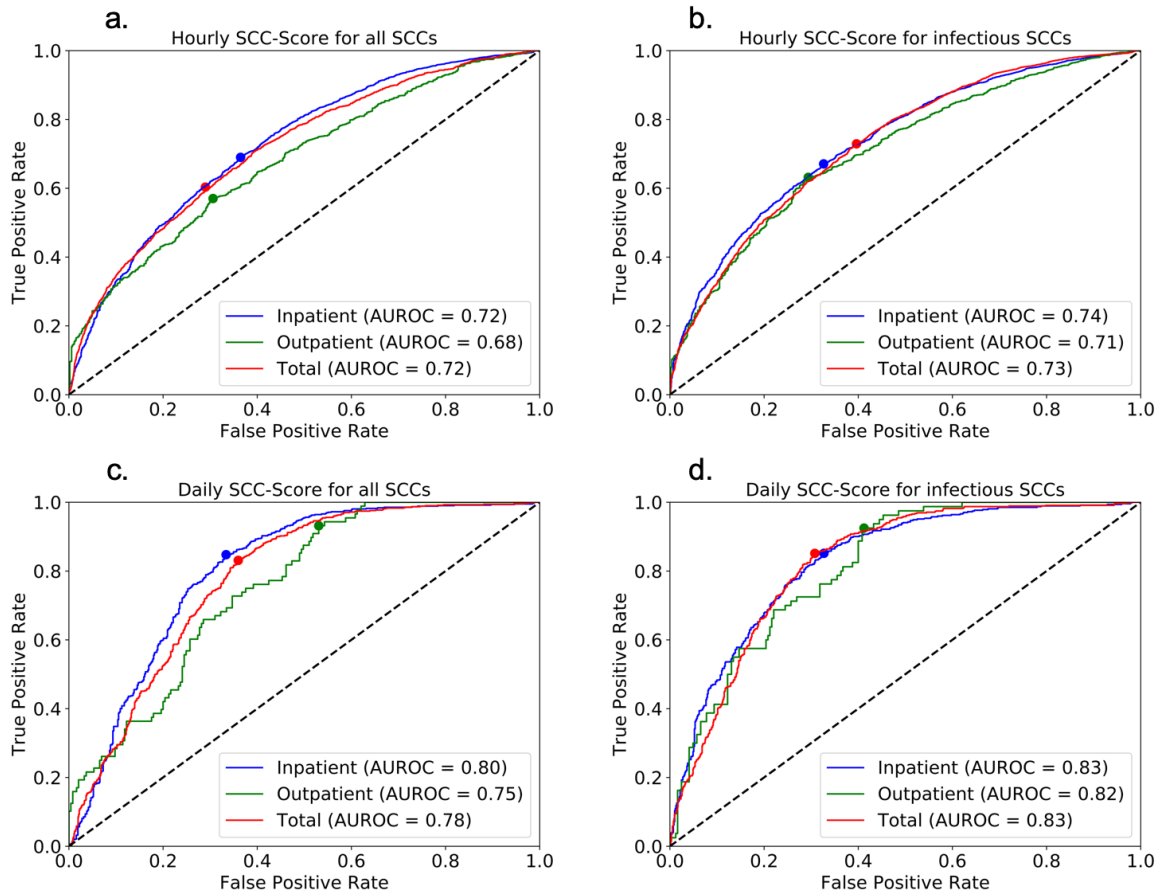
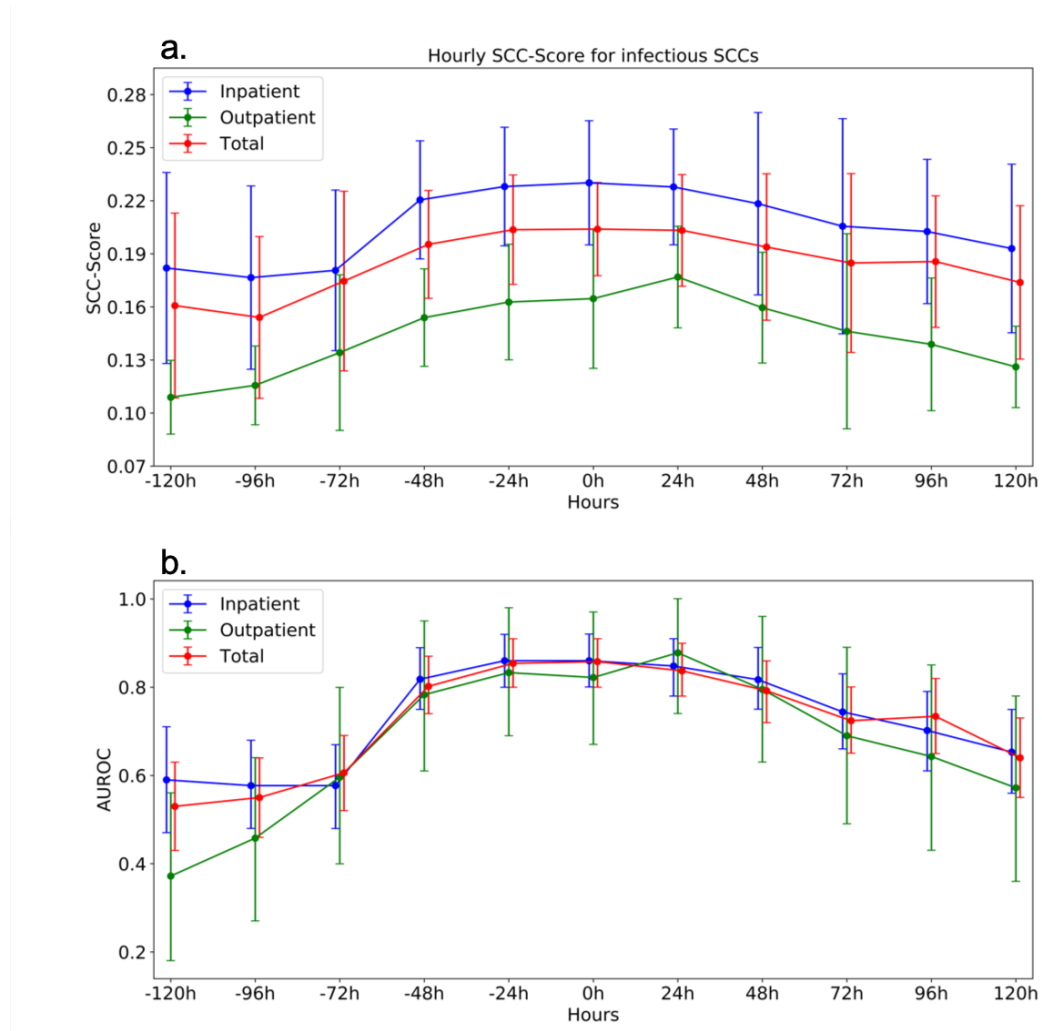


Figure 3

Time dependence of the SCC-Score for infectious SCC. (a.) Score values prior and post diagnosis at time point $t=0$ hour for the patients in the Inpatient cohort, Outpatient cohort, and Total cohort. Bars indicate standard deviation Score values. (b.) Prediction performance (AUROC) of hours containing infectious SCCs based on SCC-Score. Bars indicate 95% confidence intervals of AUROC values.



References (up to 30)

1. Jairam V, Lee V, Park HS, et al. Treatment-Related Complications of Systemic Therapy and Radiotherapy. *JAMA Oncol* 2019; **5**(7): 1028-35.
2. Sahin U, Toprak SK, Atilla PA, Atilla E, Demirer T. An overview of infectious complications after allogeneic hematopoietic stem cell transplantation. *J Infect Chemother* 2016; **22**(8): 505-14.
3. Zimmer AJ, Freifeld AG. Optimal Management of Neutropenic Fever in Patients With Cancer. *J Oncol Pract* 2019; **15**(1): 19-24.
4. Saxena A, Rubens M, Ramamoorthy V, et al. Hospitalization rates for complications due to systemic therapy in the United States. *Sci Rep* 2021; **11**(1): 7385.
5. Halpern MT, Yabroff KR. Prevalence of outpatient cancer treatment in the United States: estimates from the Medical Panel Expenditures Survey (MEPS). *Cancer Invest* 2008; **26**(6): 647-51.
6. Low CA. Harnessing consumer smartphone and wearable sensors for clinical cancer research. *NPJ Digit Med* 2020; **3**: 140.
7. Panattoni L, Fedorenko C, Greenwood-Hickman MA, et al. Characterizing Potentially Preventable Cancer- and Chronic Disease-Related Emergency Department Use in the Year After Treatment Initiation: A Regional Study. *J Oncol Pract* 2018; **14**(3): e176-e85.
8. Goodman LM, Estfan B, Montero A, et al. Improving the Management of Patients With Low-Risk Neutropenic Fever at the Cleveland Clinic Taussig Cancer Institute. *J Oncol Pract* 2017; **13**(3): e259-e65.
9. Jacobsen M DT, Kobbe G, Gaidzik PW, and Heinemann L. Noninvasive Continuous Monitoring of Vital Signs With Wearables: Fit for Medical Use? *Journal of Diabetes Science and Technology* 2020.
10. Beauchamp UL, Pappot H, Hollander-Mieritz C. The Use of Wearables in Clinical Trials During Cancer Treatment: Systematic Review. *JMIR Mhealth Uhealth* 2020; **8**(11): e22006.
11. Goldsack JA, A.; Coravos, A.; Economos, C.; Lyden K.;. The role of digital clinical measures in improving cancer care and research. *Journal of Clinical Oncology* 2021: e13584-e.
12. Wright AA, Raman N, Staples P, et al. The HOPE Pilot Study: Harnessing Patient-Reported Outcomes and Biometric Data to Enhance Cancer Care. *JCO Clin Cancer Inform* 2018; **2**: 1-12.
13. Steinhubl SR, Muse ED, Topol EJ. The emerging field of mobile health. *Sci Transl Med* 2015; **7**(283): 283rv3.
14. Jacobsen M, Rottmann P, Dembek TA, et al. Feasibility of Wearable-Based Remote Monitoring in Patients During Intensive Treatment for Aggressive Hematologic Malignancies. *JCO Clin Cancer Inform* 2022; **6**: e2100126.
15. Common Terminology Criteria for Adverse Events. Version 4.03. US Department of Health and Human Services; Published June, 2010.
16. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. International conference on machine learning; 2020: PMLR; 2020. p. 1597-607.
17. Hong S, Xu Y, Khare A, et al. Holmes: health online model ensemble serving for deep learning models in intensive care units. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2020; 2020. p. 1614-24.
18. Cortes C, Mohri M. Confidence intervals for the area under the ROC curve. *Advances in neural information processing systems* 2004; **17**.
19. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019; **17**(1): 230.
20. Stehlik J, Schmalfuss C, Bozkurt B, et al. Continuous Wearable Monitoring Analytics Predict Heart Failure Hospitalization: The LINK-HF Multicenter Study. *Circ Heart Fail* 2020; **13**(3): e006513.

21. Perez M, Pericchi L. Changing statistical significance with the amount of information: The adaptive α significance level. *Statistics & Probability Letters* 2014; **Volume 85**: Pages 20-4.
22. Beg MS, Gupta A, Stewart T, Rethorst CD. Promise of Wearable Physical Activity Monitors in Oncology Practice. *J Oncol Pract* 2017; **13**(2): 82-9.
23. Dunn J, Kidzinski L, Runge R, et al. Wearable sensors enable personalized predictions of clinical laboratory measurements. *Nat Med* 2021; **27**(6): 1105-12.

Supplementary Material

Table supplementary 1

Patient characteristics, malignancies, comorbidities, and concomitant medication of the patients in the Inpatient and Outpatient cohort.

Patient characteristics	Inpatient cohort (n=54) No. (%)	Outpatient cohort (n=25) No. (%)	Total No. (%)
Male	30 (55.6)	14 (56.0)	44 (55.7)
Female	24 (44.4)	11 (44.0)	35 (44.3)
Age [years] (median (IQR))	56 (49-62)	54 (50-62)	55 (50-62)
BMI [kg/m ²]	25.0 (21.7-28.5)	24.9 (23.3-27.1)	25.0 (22.1-28.0)
Haematological malignancy			
Acute leukaemia (ALL, AML)	21 (38.9)	12 (48.0)	33 (41.8)
MDS, MPN (PMF, CML)	20 (40.7)	0 (0.0)	21 (26.6)
Other (CLL, HL, NHL, Myeloma, Germ cell)	13 (20.4)	13 (52.0)	25 (31.6)
Comorbidities			
Arterial hypertension	13 (24.1)	14 (56.0)	27 (34.2)
Diabetes mellitus (Type 1 and 2)	2 (3.7)	0 (0.0)	2 (2.5)
Macrovascular event (e.g. Stroke)	1 (1.9)	1 (4.0)	2 (2.5)
Heart failure with reduced ejection fraction	0 (0.0)	0 (0.0)	0 (0.0)
Arrhythmias	2 (3.7)	2 (8.0)	4 (5.0)
Concomitant medication			
Antiplatelet drugs	3 (5.6)	7 (28.0)	10 (12.7)
Beta blocker	5 (9.3)	5 (20.0)	10 (12.7)
Calcium channel blocker	5 (9.3)	11 (44.0)	16 (20.0)
Renin-angiotensin system inhibitors	9 (16.7)	7 (28.0)	16 (20.0)
Other antihypertensives	4 (7.4)	0 (0.0)	4 (5.0)

BMI: Body Mass Index, ALL: Acute lymphocytic leukaemia, AML: Acute myeloid leukaemia, MDS: Myelodysplastic syndrome, MPN: Myeloproliferative neoplasms, PMF: primary myelofibrosis, CML: Chronic myeloid leukaemia, CLL: Chronic lymphocytic leukaemia, HL: Hodgkin lymphoma, NHL: Non-Hodgkin lymphoma

Table supplementary 2

Treatment regimen for Inpatient (a.) and Outpatient cohort (b.).

a.	Inpatient cohort	No. (%)
	Allogenic stem cell transplantation	48 (88·9)
	Autologous stem cell transplantation	6 (11·1)
	Conditioning protocol	
	Alkylator based (Treosulfan, Busulfan, Melphalan)	29 (53·7)
	FLAMSA-based	17 (31·5)
	TBI-based	7 (13·0)
	Other	1 (1·9)
	Intensity of conditioning regimens	
	Reduced intensity regimens	30 (55·6)
	Myeloablative conditioning	24 (44·4)
	GvHD prophylaxis	
	Antithymocyte globulin (ATG)	33 (61·1)
b.	Outpatient cohort	No. (%)
	High-/Intermediate-dose Cytarabine with or without Mitoxantron	14 (56·0)
	High-dose Cyclophosphamid	9 (36·0)
	Others (e.g. Ifosfamide, Carboplatin, Etoposide)	2 (8·0)

Table supplementary 3

Vital signs and physical activity parameters provided by the medical wearable according to the instruction for use.

Parameter	Unit	Quality Index
Heart rate	30 – 240 beats per minute	X
Oxygen saturation	65 – 100%	X
Perfusion index	0 – 255 (arbitrary)	
Activity classification	Categorical	X
Activity	0 – 255 (arbitrary)	
Steps	0 – 65,535 per day	
Blood pressure wave	0 – 5·1 (arbitrary)	
Heart rate variability	0 – 255 ms (RMSSD)	X
Respiration rate	6 – 30 per minute	X
Energy expenditure	0 – 65,535 kcal per day	X
Temperature	0 – 60°C	
Inter-beat interval	1 – 4,095 ms	X
Electrodermal activity	0 – 21·8 kOhm	

RMSSD – root mean square of successive differences between normal heartbeats

Table supplementary 4

Hours assessed by the respective model to estimate risk for infectious SCCs in periods -72, -48 and -24 hours before a diagnosis was documented (0 hour) in patients in the IC, OC and for both cohorts. Mean and standard deviation with respective models were calculated. Differences between regular and non-regular hours were tested using a two-sided t-test. To account for multiple testing, Bonferroni correction was applied and the significance level set to $0.05/9=0.0056$. Significant p-values are marked with *. Sensitivity and Specificity at the Youden index are reported.

Hours prior to SCC-diagnosis	Model	'regular' hours [n]	'non-regular' hours [n]	SCC-Score 'regular' hours [mean±SD]	SCC-Score 'non-regular' hours [mean±SD]	P-Value	Sensitivity / Specificity [%]	AUROC (95% CI)
-24	IC	682	60	0.169±0.047	0.228±0.034	<0.0001*	86.7 / 76.4	0.86 (0.80- 0.92)
	OC	245	12	0.122±0.035	0.163±0.033	0.002*	83.3 / 74.7	0.84 (0.70- 0.98)
	Total	927	72	0.154±0.042	0.204±0.031	<0.0001*	86.1 / 77.1	0.86 (0.81- 0.91)
-48	IC	679	57	0.169±0.047	0.220±0.033	<0.0001*	82.5 / 72.9	0.82 (0.75- 0.89)
	OC	245	10	0.122±0.035	0.154±0.028	0.005*	80.0 / 74.7	0.79 (0.62- 0.96)
	Total	924	67	0.154±0.042	0.195±0.030	<0.0001*	88.1 / 64.4	0.81 (0.74- 0.87)
-72	IC	648	41	0.169±0.047	0.181±0.045	0.095	61.0 / 51.7	0.56 (0.47- 0.66)
	OC	237	9	0.122±0.034	0.134±0.044	0.373	55.6 / 75.5	0.62 (0.43- 0.82)
	Total	885	50	0.153±0.041	0.175±0.051	0.011	72.0 / 48.5	0.62 (0.53- 0.70)

Figure supplementary 1

Study design: Procedures applied to enable continuous wearable-based remote monitoring of vital signs and physical activity. Data were recorded in inpatients and outpatients allocated to two different cohorts. Patients attended the baseline visit prior to starting oncological treatment. At this visit, two wearables (#1 and #2) were assigned to each patient. At Visit 2, wearable #1 was swapped to #2 and data of #1 were downloaded. Frequency of subsequent visits (including swapping wearables) was determined by the limited data storage capacity of the wearable (no web application for data download was used due to regulatory requirements) to every 3rd day. The investigators retrospectively identified serious clinical complications from the clinical documentation.

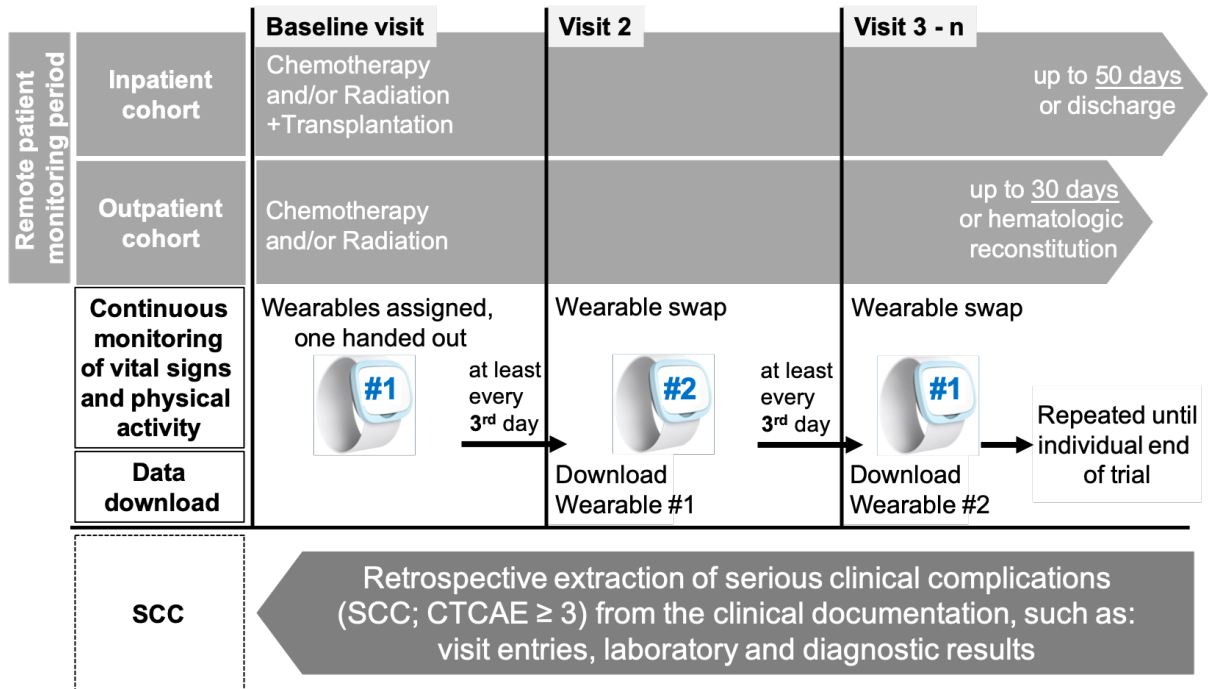


Figure supplementary 2

Acquisition of input data in the Inpatients cohort (IC) and Outpatient cohort (OC). For all patients only, hours were included if they had $\geq 3,000$ data points per hour (of maximal 3,600). Data annotation refers to the separation of hours without serious clinical complications (SCC) (= 'regular' hours) and 'non-regular' hours. Data of two patients without recording hours annotated as 'regular' hours were excluded. Allocation of data for training data set and test dataset: For training, the 'regular' hours dataset was randomly divided into two datasets (90% training data set, 10% test data set). This split was equivalently applied for the datasets of each individual patient.

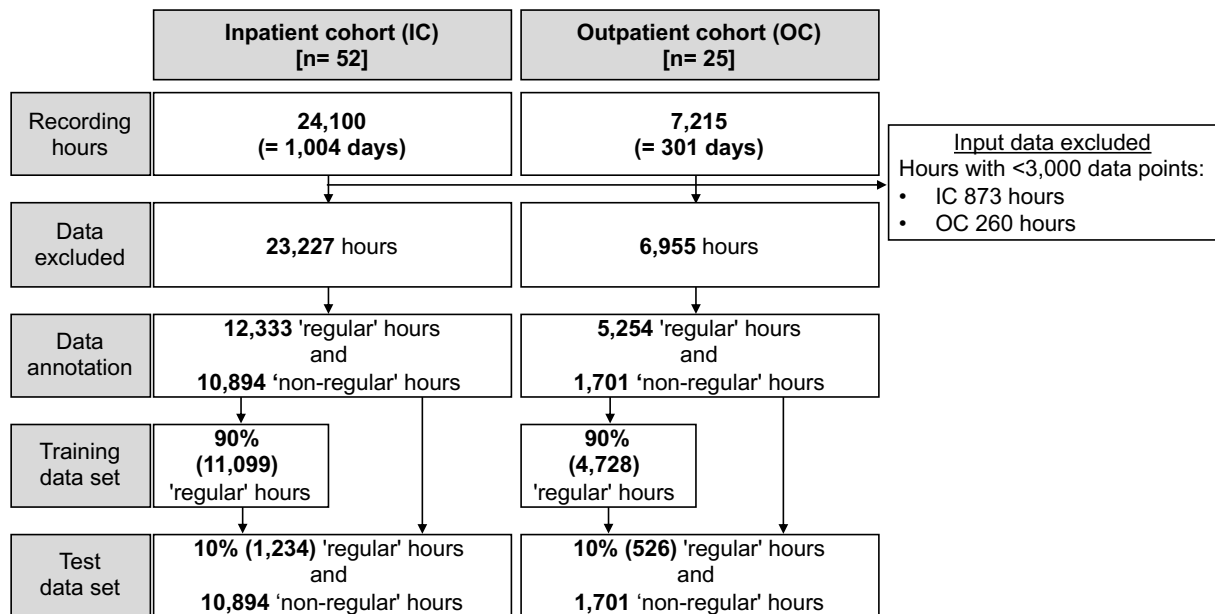


Figure supplementary 3

Distribution of SCC-Scores calculated by the AI model for ‘regular’ hours (green) and ‘non-regular’ hours (red) for (a.) Inpatient cohort (SCC_{Ic}), (b.) Outpatient cohort (SCC_{Oc}), and (c.) Total (SCC_{Total}) for ten different data splits (cross-validation). The dot indicates the mean value of each data set. The bars indicate standard deviation. Statistical significance was tested by an ANOVA between ‘regular’ and ‘non-regular’ hours and is indicated by asterisks (***) $p < 0.001$

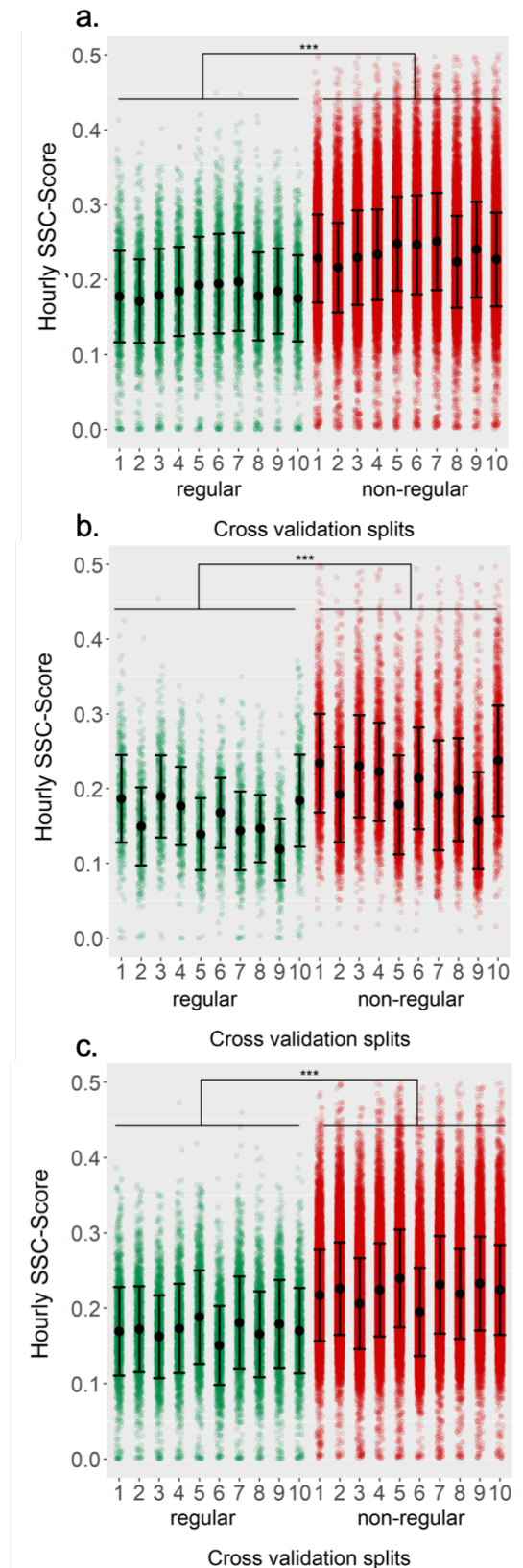


Figure supplementary 4

Occurrence of serious clinical complications in days after initiation of oncological treatment in the Inpatient cohort (blue bars) and Outpatient cohort (green bars).

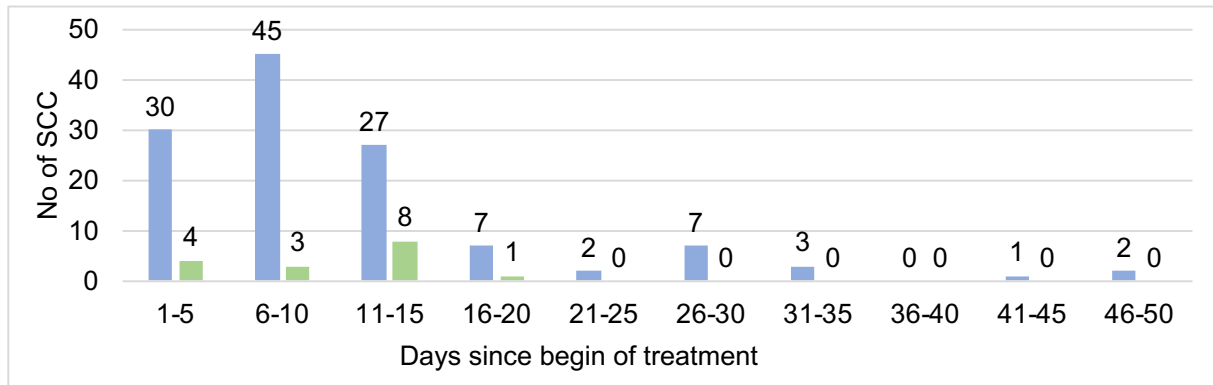
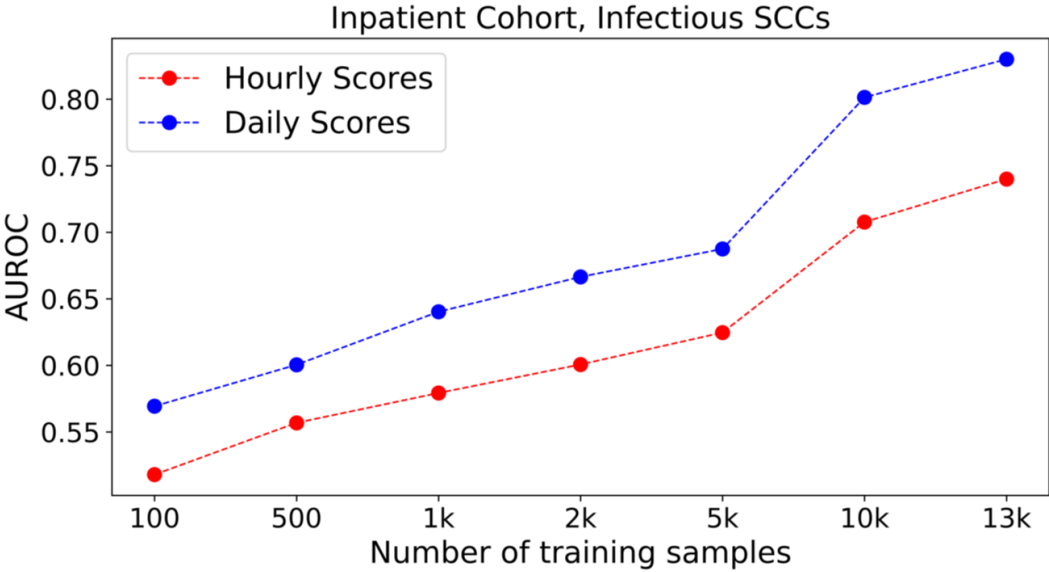


Figure supplementary 5

SCC-Score performance calculated by AUROC in relation to the number of hours used for training. For infectious SCCs hourly scores (red line) and daily score (blue line).



Chapter 8

Anomaly Detection by Negative Sampling

Detecting whether an example belongs to a given in-distribution or is Out-Of-Distribution (OOD) requires identifying features specific to the in-distribution. In the absence of labels, these features can be learned by self-supervised techniques under the generic assumption that the most abstract features are those which are statistically most over-represented in comparison to other distributions from the same domain. We show that self-distillation of the in-distribution training set together with contrasting against negative examples derived from shifting transformation of auxiliary data strongly improves OOD detection. We find that this improvement depends on how the negative samples are generated. In particular, we observe that by leveraging negative samples, which keep the statistics of low-level features while changing the high-level semantics, higher average detection performance is obtained. Furthermore, good negative sampling strategies can be identified from the sensitivity of the OOD detection score. In our proposed model, a decision boundary is formed during training unlike other contrastive methods such as *CSI* [30] which requires multiplying feature vectors by their norm during the evaluation phase to improve the discrimination of in-distribution and OOD samples. The efficiency of our approach is demonstrated across a diverse range of OOD detection problems, setting new benchmarks for unsupervised OOD detection in the visual domain in both natural and medical images.

8.1 Self-Supervised Anomaly Detection by Self-Distillation and Negative Sampling

Nima Rafiee, Rahil Gholamipoor, Nikolas Adaloglou, Simon Jaxy, Julius Ramakers, Markus Kollmann. *ICANN*, 2022.

Status: Published.

Contributions: The author contributed with evaluation, visualization and writing under the supervision of Prof. Dr. Markus Kollmann.

Self-Supervised Anomaly Detection by Self-Distillation and Negative Sampling

Nima Rafiee¹[0000-0002-3193-9534], Rahil Gholamipoor¹[0000-0001-8207-7295], Nikolas Adaloglou¹[0000-0003-4938-6322], Simon Jaxy¹[0000-0002-7076-4108], Julius Ramakers¹[0000-0002-2925-152X], and Markus Kollmann^{1,2}[0000-0002-5317-5408]

¹ Department of Computer Science, Heinrich Heine University, Düsseldorf, Germany

² Department of Biology, Heinrich Heine University, Düsseldorf, Germany

{rafiee,rahil.gholamipoorfard,nikolaos.adaloglou,simon.jaxy,ramakers,kollmann}@hhu.de

Abstract. Detecting whether examples belong to a given in-distribution or are out-of-distribution (OOD) requires identifying features that are specific to the in-distribution. In the absence of labels, these features can be learned by self-supervised representation learning techniques under the generic assumption that the most abstract features are those which are statistically most over-represented in comparison to other distributions from the same domain. This work shows that self-distillation of the in-distribution training set together with contrasting against negative examples derived from shifting transformation of auxiliary data strongly improves OOD detection. We find that this improvement depends on how the negative samples are generated, with the general observation that negative samples that keep the statistics of lower level features but change the global semantics result in higher detection accuracy on average. For the first time, we introduce a sensitivity score using which we can optimise negative sampling in a systematic way in an unsupervised setting. We demonstrate the efficiency of our approach across a diverse range of OOD detection problems, setting new benchmarks for unsupervised OOD detection in the visual domain.

Keywords: Anomaly Detection · Self-Supervised Learning · Self-Distillation · Negative Sampling.

1 Introduction

OOD detection or anomaly detection is the problem of deciding whether a given test sample is drawn from the same in-distribution as a given training set or belongs to an alternative distribution. Many real-world applications require highly accurate OOD detection for secure deployment, such as in medical diagnosis. Despite the advances in deep learning, neural network estimators can generate systematic errors for test examples that are far from the training set [25]. For example, it has been shown that Deep Neural Networks (DNNs) with ReLU activation functions can make false predictions for OOD samples with arbitrarily high confidence [12].

A major challenge in OOD detection is the case where the features of outlier examples are statistically close to the features of in-distribution examples, which is frequently the case for natural images. In particular, it has been shown that deep density estimators like Variational Autoencoders (VAEs) [16], PixelCNNs [33], and

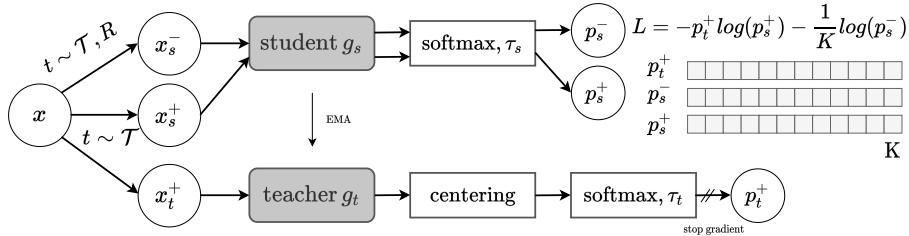


Fig. 1. An overview of the proposed contrastive self-distillation framework, consisting of student and teacher networks, g_s and g_t , that map two random transformations of the same image, $x_s^+ \sim \mathcal{T}(x)$ and $x_t^+ \sim \mathcal{T}(x)$ to the same class. Negative views, x^- , arise from first applying a shifting transformation R , such as random rotation, followed by \mathcal{T} to either an in-distribution image x or an auxiliary image x_{aux} .

normalising flow models [28] can on average assign higher likelihood to OOD examples than to examples from the in-distribution [22]. This surprising finding can be partially attributed to an inductive bias from upweighting local pixel correlations as a consequence of using convolutional neural networks.

A challenging scenario of anomaly detection is near OOD detection [35], where the OOD distribution samples are statistically very similar to the in-distribution. A particular challenging OOD detection task is given by CIFAR10 [18] as in-distribution and CIFAR100 [18] as OOD, where the larger number of classes in CIFAR100 makes it harder to identify features that are specific to the in-distribution. For the cases where the in and out distributions are not closely related, we refer to as far OOD.

State-Of-The-Art (SOTA) performance has been obtained for the CIFAR10/CIFAR100 near OOD detection task, using pretrained classification models on ImageNet-21K [8]. However, as CIFAR100 and CIFAR10 share many of their classes with ImageNet but the classes among themselves are mutually exclusive, the pretrained model effectively solves the OOD detection problem for this special case. The advantage of using pretrained models as OOD detectors drops if there is no class overlap with the OOD test set, such as for SVHN [8]. To overcome these limitations, a plethora of self-supervised pretext tasks have been proposed that provide a richer learning signal that enables abstract feature learning [4, 2, 11]. These advancements in self-supervised learning have shown remarkable results on unsupervised anomaly detection [31, 30, 35] by solely relying on the in-distribution data.

More recently, it has been suggested to include dataset-specific augmentations that shift the in-distribution – so-called negative samples. The core idea behind using shifting transformations is to concentrate the learned representation in feature space. This can result in a more conservative decision boundary for the in-distribution [14]. However, in-distribution shifting requires dataset-specific prior knowledge [21]. Therefore, a bad choice of augmentations may result in rejecting the in-distribution test samples, which reduces the OOD detection performance.

In this paper, we propose an improved version of the DINO [3] framework together with a sensitivity score for the problem of OOD detection. The main contributions of this work are summarized as follows:

- We propose a self-supervised self-distillation method that leverages unlabelled data for OOD detection which aims at drawing a tight, not necessary simply connected, decision boundary between the in-distribution data and an auxiliary negative distribution.
- We introduce an auxiliary loss that encourages unlabeled negative samples to be uniformly assigned to the existing in-distribution soft-classes.
- To the best of our knowledge, for the first time we introduce a sensitivity score defined by the AUROC value between the in-distribution training set and the in-distribution test set. Using sensitivity score, we can intuitively compare the effect of different negative auxiliary sets and to find optimal values by grid search for training hyperparameters without the access to OOD validation set.
- Finally, we show that the proposed framework does not only improve OOD detection performance but also improves representation learning for the in-distribution, as measured by the K-Nearest Neighbour (K-NN) accuracy.

2 Related Works

Supervised OOD detection methods. In-distribution classification accuracy is highly correlated with OOD performance [8]. This has motivated supervised OOD detection approaches to learn representations from classification networks [13, 19]. This can be achieved by directly training a classifier on the in-distribution or by pre-training on a larger dataset .

Fine-tuning pretrained transformers [34] has shown promising OOD scores. In computer vision, Koner et al. [17] leveraged the contextualization capabilities of pretrained Vision Transformer (ViT) [7] by exploiting the global image context. Such models heavily rely on the classes of the pretrained dataset, which often include classes from both the in and out distribution. Although, supervised pretraining can form a good boundary for OOD detection, it has two limitations, firstly the pretraining dataset should share labels with both in and out distributions, and secondly impeded OOD performance is observed when the distributions have overlapping classes.

Mohseni et al. [21] recently presented a 2-step method that initially learns how to weight the in-distribution transformations based on a supervised objective. Then, the selected shifting transformations are applied in a self-supervised setup for OOD detection. Human-level supervision is still required to learn the best shifting transformations for each training dataset. In *Geometric* [15], Hendrycks et al. defined a self-supervised task to predict geometric transformations to improve the robustness and uncertainty of deep learning models. They further improve their self-supervised technique with supervision through outlier exposure.

Unsupervised OOD detection methods. Existing label-free OOD detection approaches can be separated in density-based [27, 23], reconstruction-based [26, 38], and self-supervised learning [9, 15] methods. Density-based methods aim to fit a probability distribution (e.g. Gaussian) to the training data and then use it for OOD detection. Reconstruction-based methods assume the network would generalize less for unseen

OOD samples. Meanwhile, recent studies [22] revealed that probabilistic generative models can fail to distinguish between training data and OOD inputs.

Self-supervised methods have recently shown that adopting pretext tasks results in learning general data representations [1] for OOD detection. Choi et al. [5] used blurred data as adversarial examples to discriminate the training data from their blurred versions. In *CSI* [31], Tack et al. leverage shifting data transformations in contrastive learning for OOD detection, combined with an auxiliary task that predicts which shifting transformation was applied to a given input. In *SSD* [30], the authors further improved contrastive self-supervised training by developing a cluster-conditioned OOD detection method in the feature space.

Outlier Exposure (OE). OE leverages auxiliary data that are utterly disjoint from the OOD data [14]. Furthermore, OE assumes that the provided auxiliary samples are always OOD. To guarantee this, one needs human supervision to remove the overlap between auxiliary and in-distribution. OE has been successfully applied to training classifiers, by enforcing the auxiliary samples to be equally distributed among the in-distribution classes. In contrast to [14], we attempt to teach the network better representations for OOD detection by incorporating auxiliary data into a self-distillation soft-labeling framework.

Finally, since the proposed method does not require labels, there is no information whether the in-distribution data are meaningfully similar to the auxiliary ones. In this aspect, this work is different from OE, as it only requires the in-distribution to be sufficiently statistically underrepresented. To ensure the latter an additional transformation is applied on the auxiliary data (e.g. rotation).

3 Proposed Method

3.1 The vanilla DINO framework

The DINO framework uses two identical networks $g_s = g(x|\theta_s)$ and $g_t = g(x|\theta_t)$ called student and teacher, which differ by their sets of parameters θ_s and θ_t , respectively. For each transformed input image x , both networks produce K -dimensional output vectors, where K is the number of soft-classes. Both outputs enter a temperature-scaled softmax functions $p_t = \text{softmax}(g_t, \tau_t)$ and $p_s = \text{softmax}(g_s, \tau_s)$ defined by:

$$p^i(x) = \frac{\exp(g^i(x)/\tau)}{\sum_{k=1}^K \exp(g^k(x)/\tau)}, \quad (1)$$

where $p^i(x)$ is the probability of x falling in soft-class i and τ_s, τ_t are the student and teacher temperatures. In contrast to knowledge distillation methods, the teacher is built from previous training iterations of the student network. To do so, the gradients are back-propagated only through the student network and the teacher parameters are updated with the Exponential Moving Average (EMA) of the student parameters

$$\theta_t \leftarrow m\theta_t + (1 - m)\theta_s, \quad (2)$$

where $0 \leq m \leq 1$ is a momentum parameter. For $\tau_t < \tau_s$, the training objective is given by the cross entropy loss for two non-identical transformations x'', x' of an image x drawn from the in-distribution training set \mathcal{D}_{train}^{in}

Table 1. AUROC scores for OOD detection without label supervision.

\mathcal{D}_{train}^{in}	\mathcal{D}_{test}^{out}	Geometric* [15]	SSD[30]	CSI[31]	MTL [†] [21]	Ours	
						Rot.ImgN	Combined
CIFAR10	CIFAR100	91.91	90.63	89.20	93.24	92.51	94.20
	SVHN	97.96	99.62	99.80	99.92	99.69	99.92
	ImageNet30	–	90.20	87.92	–	94.16	93.40
	TinyImageNet	92.06	92.25	92.44	92.99	96.28	95.02
	LSUN	93.57	96.51	91.60	95.03	98.08	97.52
	STL10	–	70.28	64.25	–	77.29	74.34
	Places365	92.57	95.21	90.18	93.72	97.14	96.01
	Texture	96.25	97.61	98.96	–	99.16	98.69
CIFAR100	CIFAR10	74.73	69.60	58.87	79.25	69.96	67.63
	SVHN	83.62	94.90	96.44	87.11	96.00	97.17
	ImageNet30	–	75.53	71.82	–	84.82	75.36
	TinyImagenet	77.56	79.52	79.28	80.66	81.41	79.75
	LSUN	71.86	79.50	61.83	74.32	85.03	74.55
	STL10	–	72.76	64.26	–	79.96	71.70
	Places365	74.57	79.60	65.48	77.87	81.67	72.79
	Texture	82.39	82.90	87.47	–	80.65	77.33

* Requires labels for the supervised training loss. Results reported from [21].

† Requires labels to select the optimal transformations.

$$\mathcal{L}_{pos} = - \sum_{x'' \in G} \sum_{\substack{x' \in V \\ x' \neq x''}} p_t(x'') \log(p_s(x')). \quad (3)$$

Additionally, DINO uses the multi-crop strategy [2], wherein M global views $G = \{x_1^g, \dots, x_M^g\}$ and N local views, $L = \{x_1^l, \dots, x_N^l\}$, are generated based on a set of transformations \mathcal{T} , e.g. crop and resize, horizontal flip, Gaussian blur, and color jitter. Global views are crops that occupy a larger region of the image (e.g. $\geq 40\%$) while local views cover small parts of the image (e.g. $\leq 40\%$). All $V = G \cup L$ views are passed through the student network, while the teacher has only access to the global views such that local-to-global correspondences are enforced. The trained teacher network is used for evaluation.

3.2 Negative samples

The learning objective (Eq. 3) assigns two transformed views of an image to the same soft-class. The applied transformations \mathcal{T} are chosen to be sufficiently strong and diverse, such that the generated images generalise well over the training set but keep the semantics of the image they were derived from. The transformations are designed to learn higher-level features such as labels that represent semantic information and avoid learning lower-level features, such as edges or the color statistics over pixels [4]. The quality of the learned representation can be quantified by evaluating the K-NN accuracy for an in-distribution test set \mathcal{D}_{test}^{in} , using as higher-level feature vector an activity map of the network near the last layer. For OOD detection, the feature vector

representation should be enriched by in-distribution-specific features and depleted by features that frequently appear in other distributions from the same domain. This can be achieved by designing a negative distribution D_{neg} that keeps most of the low-level features of the in-distribution but changes the high-level semantics.

For example, a negative distribution for natural images can be realised by additionally rotating in-distribution images or images from a related auxiliary distribution by $r \sim R = \mathcal{U}(\{90^\circ, 180^\circ, 270^\circ\})$, where \mathcal{U} is the uniform distribution. It has been shown that using rotation as an additional positive transformation degrades the performance in the contrastive learning setup, where the objective is to maximize the mutual information between positive examples [4]. Motivated by this, authors in [31] report a performance gain for OOD detection by using rotation to generate negative examples.

3.3 Auxiliary objective

In addition to the self-distillation objective Eq. 3 we define an auxiliary task to encourage the student to have a uniform softmax response for negative examples. This task can be realised by a similar objective as Eq. 3 but with changed temperature $\tau_t \rightarrow \infty$ and transformations \mathcal{T} applied to examples x from the negative set \mathcal{D}_{neg} , defined as:

$$\mathcal{L}_{neg} = -\frac{1}{K} \sum_{x' \in V} \log p_s(x'). \quad (4)$$

The total loss of our proposed method is defined by a linear combination of the two objectives

$$\mathcal{L}_{total} = \mathcal{L}_{pos} + \lambda \mathcal{L}_{neg}, \quad (5)$$

where $\lambda > 0$ is a balancing hyperparameter.

3.4 Sensitivity Score

Intuitively the sensitivity score is the degree of rejection of in-distribution examples which gives us a measure about the sensitivity of the OOD score to examples that have very similar features statistics to \mathcal{D}_{train}^{in} . To calculate the sensitivity score we randomly extract B samples from \mathcal{D}_{train}^{in} without replacement as $\mathcal{D}_{train}^{sens}$ and denote the remaining train samples as $\mathcal{D}_{train}^{ref}$. We define the sensitivity score as the AUROC value between $\mathcal{D}_{train}^{sens}$ and \mathcal{D}_{test}^{in} , where $\mathcal{D}_{train}^{ref}$ is used as new train data during the evaluation.

4 Experiments

The proposed method is based on the vanilla DINO [3] implementation³. Unless otherwise specified, we use ViT-Small (ViT-S) with a patch size of 16. We use $N = 8$ local views for both positives and negatives, but two global positive views and one global negative view. Global views are resized to 256×256 while local views to 128×128 . The temperatures are set to $\tau_t = 0.01$ and $\tau_s = 0.1$. In each epoch, we

³ <https://github.com/facebookresearch/dino>

linearly decrease τ_t starting from 0.055 for CIFAR10 and from 0.050 for CIFAR100 to 0.01 during training. Sensitivity score is used to find optimal $\lambda = 1$. we set $K = 4096$ for all experiments. We use the Adamw optimizer [20] with an effective batch size of 256. The learning rate lr follows the linear scaling rule of $lr = lr_{\text{base}} \times \text{batchsize} / 256$, where $lr_{\text{base}} = 0.004$. All models are trained for 500 epochs. Experiments were conducted using 4 NVIDIA-A100 GPUs with 40GB of memory. The image augmentation pipeline \mathcal{T} is based on [10, 3]. Finally, weight decay and learning rate are scaled with a cosine scheduler.

Table 2. AUROC scores for OOD Detection with CIFAR10 as $\mathcal{D}_{\text{train}}^{\text{in}}$ and different \mathcal{D}_{neg} . ImgN denotes ImageNet samples.

Negative Sampling:	None	Auxiliary								In-Dist
$\mathcal{D}_{\text{test}}^{\text{out}}$	DINO $\lambda = 0$	ImgN	Rot.	Rot.	DTI	Perm-	Perm-	Rot.	Pix.	Rot.
			ImgN	360		16	4	DTI	Perm.	In-Dist.
CIFAR100	90.29	90.46	92.51	88.62	93.77	88.32	89.57	93.77	87.67	93.96
SVHN	99.38	99.50	99.69	99.42	99.86	99.59	99.13	99.86	99.62	99.92
ImageNet30	88.81	89.96	94.16	88.95	93.39	89.17	84.71	96.04	87.46	91.69
TinyImageNet	91.07	94.14	96.28	91.60	94.53	89.72	91.27	95.64	89.39	94.27
LSUN	92.20	93.41	98.08	93.24	98.56	94.58	89.32	99.12	93.33	94.93
STL10	66.50	77.65	77.29	72.41	72.01	69.22	68.81	81.49	68.55	69.11
Places365	91.28	93.12	97.14	92.58	97.03	92.77	87.63	98.12	91.89	93.53
Texture	96.21	95.01	99.16	93.93	97.55	93.38	89.86	95.11	93.08	98.29
Average	89.47	91.66	94.29	90.09	93.34	89.59	87.54	94.89	88.87	91.96

4.1 Datasets and negative sample variants

We evaluate our method on CIFAR10 and CIFAR100 as in-distribution data. For auxiliary datasets, we use ImageNet [29] and Debiased 300K Tiny Images (DTI) [14]. The latter is a subset with 300K images from [32], where images belong to CIFAR10, CIFAR100, Places365 [37], and LSUN [36] classes are removed. To avoid shortcut learning (due to different image resolutions), we resize the auxiliary data to the size of the in-distribution data before applying any augmentation. For OOD detection, we consider common benchmark datasets, such as SVHN [24], Places365, Texture [6] and STL10. The following cases are considered for generating negative samples:

- DINO: no negatives are included ($\lambda = 0$).
- ImgN: samples from ImageNet.
- DTI: samples from Debiased Tiny Images.
- Rot.: samples are randomly rotated by $r \sim R = \mathcal{U}(\{90^\circ, 180^\circ, 270^\circ\})$.
- Rot.360: samples are rotated by an angle randomly sampled from range $(0^\circ, 360^\circ)$.
- Perm- N : randomly permutes each part of the evenly partitioned image in N patches.
- Pix. Perm: randomly shuffles all the pixels in the image.
- Rot. In-Dist: a random rotation $r \sim R$ is applied to the in-distribution data.
- Combined: sample from both Rot. In-Dist and Rot. ImageNet are used.

4.2 Evaluation protocol for OOD detection

The DINO network structure $g(x)$ used in this work consists of a ViT-S as backbone, which maps the input x to a d -dimensional feature vector $f \in \mathbb{R}^d$, and two fully connected layers as head, which converts the features vector f to a K -dimensional output vector that enters the softmax layer. We define an anomaly detection score, \mathcal{S} , for the OOD test data \mathcal{D}_{test}^{out} by computing the cosine similarity between the feature vector for a test image f_{test} and all features vectors f_m of the in-distribution training set. Instead of taking the maximum cosine similarity as a OOD score, we opt for a temperature weighted non-linear score,

$$\mathcal{S}(x) = -\frac{1}{M} \sum_{m=1}^M \exp\left(\frac{1}{\tau} \cdot \frac{f_{test}^T f_m}{\|f_{test}\| \|f_m\|}\right), \quad (6)$$

with $\tau = 0.04$ which is found by maximizing the sensitivity score. The value $\tau = 0.04$ is the average over optimal values for different datasets that typically lie in the range $[0.02, 0.08]$. The score is used to evaluate OOD performance by reporting the Area Under the Receiver Operating characteristic Curve (AUROC) between a given OOD test set and the in-distribution test set.

4.3 Experimental results

In Table 1, quantitative results are reported for CIFAR10 and CIFAR100 as in-distribution. We report results with Rot. ImgN as well as combining them with in-distribution rotated samples (Combined). When using CIFAR10 as \mathcal{D}_{train}^{in} , the proposed method shows superior performance in 6 out of 8 (75%) OOD datasets compared to current SOTA self-supervised methods. Surprisingly, we surpass hybrid methods, where self-supervised training is combined with human-labelled images. By further leveraging in-distribution negatives, we are able to surpass all other methods by 3.57% and 0.96% against self-supervised and supervised methods, respectively. Our results are roughly consistent for CIFAR100 as \mathcal{D}_{train}^{in} . We report superior performance in 6 out of 8 (75%) OOD datasets. Far OOD datasets have a substantial benefit, such as LSUN where we report a 5.53% gain against the best self-supervised method. Our results on near OOD, CIFAR10, are on par with self-supervised methods [31], while lacking behind supervised methods.

In Table 2, we investigate several ways to generate negative samples, as detailed in Section 4.1. It can be observed that by rotating both ImageNet and DTI with R , both distributions demonstrate an average performance gain of 2.63% and 1.55% respectively compared to no additional transformation.

It is worth noting that we abstain from reporting the performance of DTI in Table 1, since labels were used to form this subset of 300K images. Finally, we report an inferior (or on par) average AUROC score when employing Pix. Perm, Perm-4, and Perm-16 against the vanilla DINO method using ImageNet as the auxiliary dataset.

5 Discussion

Do negative samples lead to more condensed in-distribution representations? To understand the impact of the introduced negative sampling methods, we

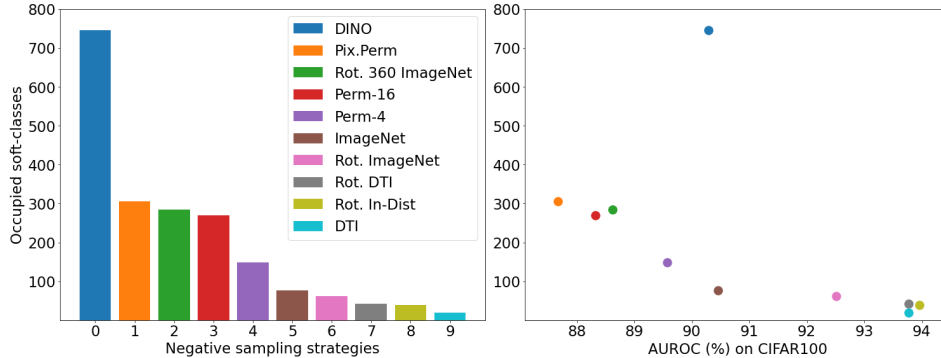


Fig. 2. We define a soft-class as “occupied” if the probability assigned to that soft-class is greater than the average probability of all K soft-classes. Colors indicate multiple \mathcal{D}_{neg} and are shared within the two plots. The teacher network g_t is used to generate p_t from \mathcal{D}_{test}^{in} . Training is performed on CIFAR10. **Left:** \mathcal{D}_{test}^{in} occupy less soft-classes with negative sampling compared to the DINO baseline. **Right:** relationship of occupied soft-classes with respect to AUROC score in CIFAR100.

investigate how many of the $K = 4096$ soft-classes are “occupied” by the \mathcal{D}_{test}^{in} after training on CIFAR10. A soft-class is considered occupied if the probability assigned to the intended soft-class computed from test data is greater than the average soft-class probability. As depicted in Fig. 2 (left), negative sampling reduces the occupied classes compared to the DINO baseline. This observation is independent of how \mathcal{D}_{neg} is created. More specifically, Rot. ImageNet, Rot. DTI, and Rot. In-Dist use roughly the same number of soft-classes and achieve SOTA AUROC scores on CIFAR100. By combining the aforementioned qualitative evaluations with Table 2, we claim that by contrasting \mathcal{D}_{train}^{in} against \mathcal{D}_{neg} a more condensed representation can be learnt.

Is OOD detection related to in-distribution classification? To answer this question, we investigate if there is a relationship between the OOD detection performance and the K-NN accuracy, determined from human-generated labels. To do so, we use CIFAR10 as \mathcal{D}_{train}^{in} and CIFAR100 and Texture as \mathcal{D}_{test}^{out} , as representative cases of near OOD and far OOD, respectively. We find that the OOD AUROC score is positively correlated with K-NN accuracy for both near and far OOD detection (Fig. 3, top row).

Is the performance gain from use of transformers or auxiliary loss function? The performance gain stems from a more compact representation of high-level features for the in-distribution. This can be seen from the high K-NN values, that can be partially attributed to the DINO self-distillation framework (CIFAR10 K-NN accuracy of 93.2% for vanilla DINO vs 87.1% for CSI) and in part due to the negative loss (4.82% AUROC improvement with Rot.ImgN compared to vanilla DINO on CIFAR10). We highlight that K-NN correlates positively with OOD AUROC values (Fig. 3, top row).

Can an arbitrary auxiliary dataset be detrimental? Auxiliary negative datasets can be detrimental if they are semantically too close to the in-distribution, which

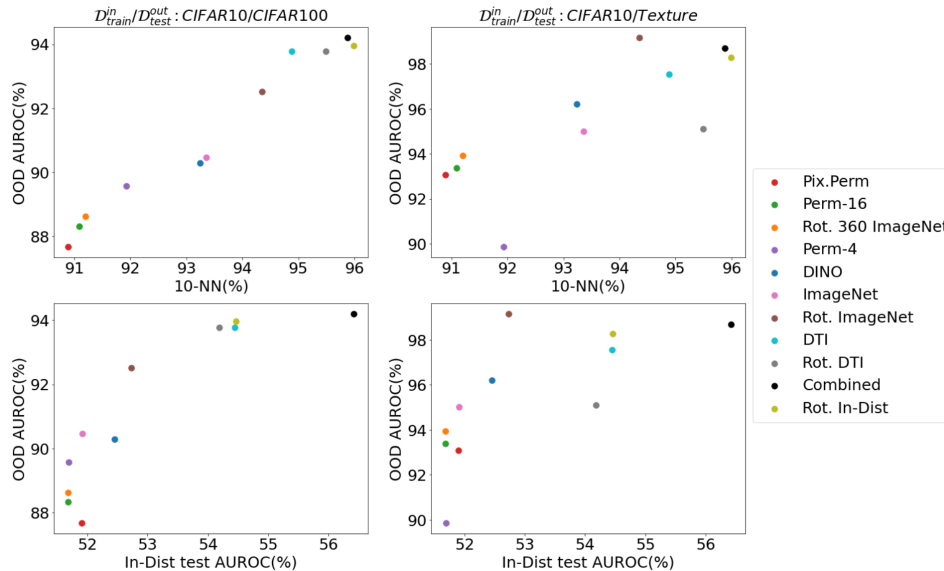


Fig. 3. We evaluate different models trained on CIFAR10 for two OOD datasets, CIFAR100 (left column) and Texture (right column). In each plot, points indicate different negative sampling strategies (colors are shared). **Top row:** correlation between OOD detection AUROC and K-NN accuracy on \mathcal{D}_{test}^{in} . **Bottom row:** correlation between OOD detection AUROC and AUROC score of \mathcal{D}_{train}^{in} vs. \mathcal{D}_{test}^{in} . We observe models with higher sensitivity to detect \mathcal{D}_{test}^{in} as outliers have higher OOD detection performance.

explains why non-rotated ImgN gives worse AUROC than Rot. ImgN for CIFAR10, despite the former being closer to the in-distribution. However, this effect can be detected by our sensitivity score, which is higher for Rot. ImgN (Fig. 3, bottom row). **How to choose good negative examples?** We use the sensitivity score to select \mathcal{D}_{neg} (dataset + augmentation). Sensitivity values significantly higher than 0.5 indicate that negative examples are close enough to induce a difference between \mathcal{D}_{train}^{in} and \mathcal{D}_{test}^{in} , but far enough to avoid a significant overlap of \mathcal{D}_{train}^{in} with \mathcal{D}_{neg} (see sensitivities of ImgN vs. Rot. ImgN, Fig. 3, bottom row).

6 Conclusion

In this work, we presented a new general method for self-supervised OOD detection. We demonstrated how self-distillation can be extended to account for positive and negative examples by introducing an auxiliary objective. The proposed objective introduces a form of contrastive learning, which pushes negative samples to be uniformly distributed among the existing in-distribution soft-classes. Additionally, we introduced a sensitivity analysis technique with which we can compare negative datasets and find the optimal values for the negative loss weight and the evaluation temperature without accessing the OOD validation set. The proposed method outperforms current SOTA for self-supervised OOD detection methods in the majority

of OOD benchmark datasets for both CIFAR10 and CIFAR100 as \mathcal{D}_{train}^{in} . We hope that the provided insights of our analysis will shed light on how to choose negative samples in more challenging vision domains.

References

1. Alexey, D., Fischer, P., Tobias, J., Springenberg, M.R., Brox, T.: Discriminative, unsupervised feature learning with exemplar convolutional, neural networks. *IEEE TPAMI* **38**(9), 1734–1747 (2016)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882* (2020)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294* (2021)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *International conference on machine learning*. pp. 1597–1607. PMLR (2020)
5. Choi, S., Chung, S.Y.: Novelty detection via blurring. *arXiv preprint arXiv:1911.11943* (2019)
6. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3606–3613 (2014)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
8. Fort, S., Ren, J., Lakshminarayanan, B.: Exploring the limits of out-of-distribution detection. *arXiv preprint arXiv:2106.03004* (2021)
9. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. *Advances in neural information processing systems* **31** (2018)
10. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent—a new approach to self-supervised learning. *Advances in Neural Information Processing Systems* **33**, 21271–21284 (2020)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9729–9738 (2020)
12. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem (2019)
13. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* (2016)
14. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606* (2018)
15. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems* **32** (2019)
16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013)
17. Koner, R., Sinhamahapatra, P., Roscher, K., Günnemann, S., Tresp, V.: Oodformer: Out-of-distribution detection transformer. *arXiv preprint arXiv:2107.08976* (2021)

18. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
19. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv preprint arXiv:1706.02690 (2017)
20. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)
21. Mohseni, S., Vahdat, A., Yadawa, J.: Shifting transformation learning for out-of-distribution detection (2021)
22. Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? arXiv preprint arXiv:1810.09136 (2018)
23. Nalisnick, E.T., Matsukawa, A., Teh, Y.W., Lakshminarayanan, B.: Detecting out-of-distribution inputs to deep generative models using a test for typicality. (2019)
24. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.: Reading digits in natural images with unsupervised feature learning (2011)
25. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 427–436 (2015)
26. Pidhorskyi, S., Almohsen, R., Doretto, G.: Generative probabilistic novelty detection with adversarial autoencoders. *Advances in neural information processing systems* **31** (2018)
27. Ren, J., Liu, P.J., Fertig, E., Snoek, J., Poplin, R., Deprieto, M., Dillon, J., Lakshminarayanan, B.: Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems* **32** (2019)
28. Rezende, D.J., Mohamed, S.: Variational inference with normalizing flows (2016)
29. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *International journal of computer vision* **115**(3), 211–252 (2015)
30. Sehwal, V., Chiang, M., Mittal, P.: Ssd: A unified framework for self-supervised outlier detection. In: International Conference on Learning Representations (2021), <https://openreview.net/forum?id=v5gjXpmR8J>
31. Tack, J., Mo, S., Jeong, J., Shin, J.: Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems* **33**, 11839–11852 (2020)
32. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence* **30**(11), 1958–1970 (2008)
33. Van Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International conference on machine learning. pp. 1747–1756. PMLR (2016)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: *Advances in neural information processing systems*. pp. 5998–6008 (2017)
35. Winkens, J., Bunel, R., Roy, A.G., Stanforth, R., Natarajan, V., Ledsam, J.R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., et al.: Contrastive training for improved out-of-distribution detection. arXiv preprint arXiv:2007.05566 (2020)
36. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
37. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* **40**(6), 1452–1464 (2017)
38. Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: International conference on learning representations (2018)

8.2 Abnormality Detection for Medical Images using Self-Supervision and Negative Samples

Nima Rafiee, Rahil Gholamipoor, Markus Kollmann. 2022.

Status: Submitted to *MICCAI*.

Contributions: The research and preparation of this manuscript were done jointly by Nima Rafiee and Rahil Gholamipoor under the supervision of Prof. Dr. Markus Kollmann.

Abnormality Detection for Medical Images Using Self-Supervision and Negative Samples

Nima Rafiee¹*[0000-0002-3193-9534], Rahil Gholamipoor¹*[0000-0001-8207-7295],
and Markus Kollmann^{1,2}[0000-0002-5317-5408]

¹ Department of Computer Science, Heinrich Heine University, Düsseldorf, Germany

² Department of Biology, Heinrich Heine University, Düsseldorf, Germany
{nima.rafiee,rahil.gholamipoorfard,markus.kollmann}@hhu.de

Abstract. Recent progress in computer-aided technologies has considerable impact on helping experts with a reliable and fast diagnosis of abnormal samples. In particular, self-supervised and self-distillation techniques have advanced automated out-of-distribution (OOD) detection in the image domain. Further improvements for OOD detection have been observed by including negative samples derived from shifting transformations of real images. In this work, we study different ways of creating negative samples for medical images and how effective they are when leveraging them in a self-supervised self-distillation framework. We investigate the impact of various types of negative examples by applying different shifting transformations on samples when they are derived from in-distribution training data, from an auxiliary dataset, or a combination of both. For the case of the auxiliary dataset, we compare the OOD detection performance when auxiliary samples are extracted from an in-domain or an out-domain. Our approach uses only data belonging to healthy people during the training procedure and does not require any additional information from labels. We demonstrate the efficiency of our technique by comparing abnormality detection performance on diverse medical datasets, setting new benchmarks for pneumonia, polyp, and glaucoma detection from X-ray, colonoscopy, and ophthalmology images.

Keywords: Abnormality detection · Self-supervised learning · Medical imaging.

1 Introduction

In recent years, computer-aided diagnosis in medical image screening has gained increased attention. In particular, detecting whether a sample includes some abnormality can help medical experts with faster and more reliable decision making. Diagnosis problems can be frequently assigned to the problem of out-of-distribution (OOD) detection in machine learning and statistical inference. OOD detection or anomaly detection refers to the problem of detecting if a test

* Equal contribution

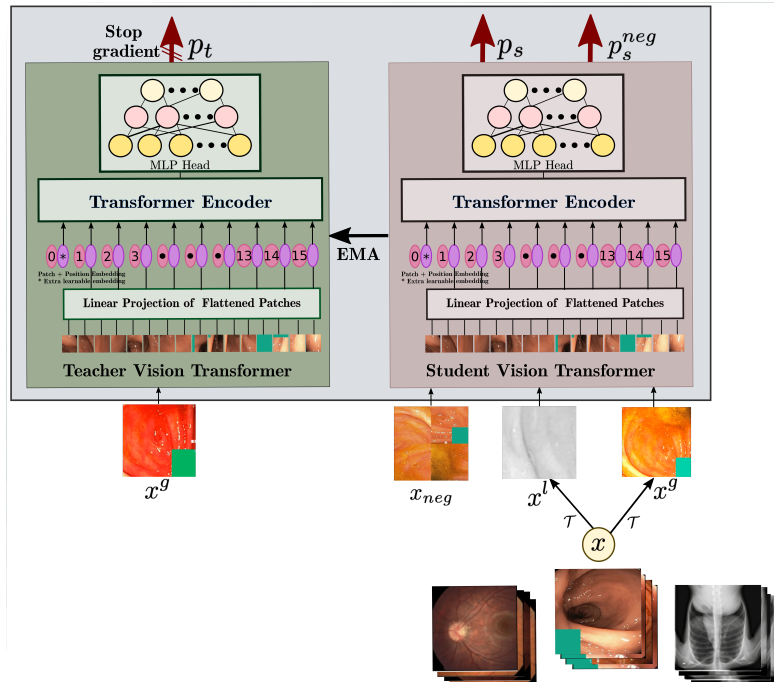


Fig. 1: Overview of the proposed self-supervised framework, comprising student network (right) and teacher network (left). Student and teacher map two randomly augmented views of the same image to the same class. x^g and x^l are global and local views of image x where $x^g \sim \mathcal{T}(x)$ and $x^l \sim \mathcal{T}(x)$. Negative sample, x_{neg} , is generated by applying first a shifting transformation, such as random rotation, followed by \mathcal{T} to either an in-dist image x or an auxiliary image x_{aux} .

sample has the same distribution as training data or is drawn from a different distribution. Diverse techniques developed for computer vision problems have been successfully applied to abnormality detection in the medical field. In [25, 29] deep supervised methods are used to classify X-ray and colonoscopy images. Despite the promising results, these approaches rely on annotated samples for abnormalities that are not available or only available to very limited number. Typically, the number of healthy samples outnumbers abnormal ones, which results in a challenging unbalanced classification problem. To overcome these problems, many studies have investigated the use of unsupervised or semi-supervised methods [7, 26, 19]. These methods aim at detecting abnormalities by learning the distribution of healthy/normal data. A well-studied category of unsupervised methods named Variational Autoencoders (VAEs) [13] uses reconstruction error. The assumption is that abnormal samples can not be reconstructed equally accurately as training data (lower likelihood) where the model only uses normal images during the training. However, it has been shown that in practice, these models can

be prone to reconstruct the abnormal samples fairly well, which lowers the detection performance [25, 2]. Furthermore, it is shown that these density estimation based methods can assign a higher likelihood to OOD samples compared to in-distribution (in-dist) test data [18]. Recently the effectiveness of self-supervised learning has received considerable attention in different domains, such as the visual domain [5], which enables learning robust representations from unlabeled data. Due to their efficiency, self-supervised pretext tasks such as predicting geometric transformations [10] or contrastive learning [27, 30, 9, 22, 24] have been designed for OOD detection in both natural and medical images. In [24] negative samples, drawn from shifting transformation of train data, are incorporated into a contrastive method to further tighten the decision boundary around normal samples resulting in an improved OOD detection score. This approach is also supported by Ref. [11] where supervised and density estimator models are exposed to some auxiliary datasets and negative samples. In [20] a self-distillation approach similar to DINO [4] is used with negative samples in order to compensate for limitations of contrastive based methods. Despite the numerous studies of leveraging negative samples in natural images, we believe it has remained untapped in the field of medical image processing. In this work, we study different ways of creating negative examples by applying shifting transformations on in-dist train data, samples from an auxiliary dataset or a combination of both. We show how these different negative samples can affect the performance of abnormality detection when leveraging them into a self-distillation self-supervised method. With the general assumption that effective transformations are the ones that change the high-level semantic while keeping the low-level statistics, we can achieve state-of-the-art (SOTA) results on abnormality detection for three different medical datasets including detecting pneumonia, polyp, and glaucoma from chest X-ray, colonoscopy, and from ophthalmology images with only access to normal samples. Additionally, we compare two evaluation metrics, cosine similarity and Mahalanobis distance, for OOD detection.

2 Method

In this section, we describe our proposed approach, Fig. 1. Similar to [4], our framework use teacher and student networks that have the same architecture, Vision Transformer [8] (ViT), and use distillation during training. Student and teacher are parametrized by two identical networks $g_s = g(x|\theta_s)$ and $g_t = g(x|\theta_t)$ which have different set of parameters. For an augmented input image x , both student and teacher output K -dimensional vectors including soft-classes. The probability of x falling in soft-class k is computed using temperature-scaled softmax function defined as

$$p_s^k(x) = \frac{\exp(g_s^k(x)/\tau_s)}{\sum_{i=1}^K \exp(g_s^i(x)/\tau_s)}, \quad (1)$$

where $\tau_s > 0$ is student temperature. The same formula, Eq. 1, holds for teacher with temperature τ_t . The student parameters are updated by back-propagating

the gradients through the student network while the teacher parameters are updated with the Exponential Moving Average (EMA) of the student parameters

$$\theta_t \leftarrow m\theta_t + (1 - m)\theta_s, \quad (2)$$

where $0 \leq m \leq 1$ is a momentum parameter. For $\tau_t < \tau_s$, the training objective is given by the cross entropy (CE) loss for two non-identical transformations x'', x' of an image x drawn from the training set, \mathcal{D}_{train}

$$\mathcal{L} = - \sum_{x'' \in G} \sum_{\substack{x' \in G \cup L \\ x' \neq x''}} p_t(x'') \log p_s(x'). \quad (3)$$

We additionally use the multi-crop strategy [3], wherein M global views $G = \{x_1^g, \dots, x_M^g\}$ and N local views, $L = \{x_1^l, \dots, x_N^l\}$, are generated based on a set of transformations \mathcal{T} . Global views usually cover a larger region of the original image while local views cover smaller as they are results of a stronger cropping. All global and local views are passed through the student network, while the teacher has only access to the global views encouraging local-to-global correspondence. The CE loss, Eq. 3, is minimized such that two transformed views of an input image are assigned to the same soft-class. The applied transformations \mathcal{T} are chosen to be strong and diverse enough such that the generated images generalise well over the training data. The transformations are designed in order to learn higher-level features, semantic information, and avoid learning lower-level features.

2.1 Auxiliary objective for OOD detection

For OOD detection, the representations should be enriched by in-dist specific features and deprived of features that frequently appear in other distributions from the same domain. This can be achieved by designing a negative distribution \mathcal{D}_{neg} that keeps most of the low-level features of the in-dist data but changes the high-level semantics.

In addition to the self-distillation objective, Eq. 3, we define an auxiliary task to encourage the student to have a uniform softmax response for negative examples. This can be done by a similar objective as Eq. 3 when temperature $\tau_t \rightarrow \infty$

$$\mathcal{L}_{neg} = -\frac{1}{K} \sum_{x_{neg} \in \mathcal{D}_{neg}} \log p_s(x_{neg}). \quad (4)$$

The total loss of our proposed method is defined by a linear combination of the two objectives

$$\mathcal{L}_{total} = \mathcal{L} + \lambda \mathcal{L}_{neg}, \quad (5)$$

where $\lambda > 0$ is a balancing hyperparameter.

2.2 Negative samples

A negative distribution \mathcal{D}_{neg} can be realised by additionally applying shifting transformations to samples from \mathcal{D}_{train} or from an auxiliary set augmented by \mathcal{T} . We consider the following shifting transformations to shape \mathcal{D}_{neg} .

- NoNeg: no negative samples are included ($\lambda = 0$).
- Rot: samples are randomly rotated by $r \sim \mathcal{U}(\{90^\circ, 180^\circ, 270^\circ\})$.
- Rot-360: rotation by an angle randomly sampled from range $(0^\circ, 360^\circ)$.
- Perm- n : random permutation of image patches where the image is sliced in n square patches.
- Pixel-Shuffle: randomly shuffles all pixels in the image.

2.3 Evaluation protocol for OOD detection

Different studies have shown the advantage of using Mahalanobis distance and cosine similarity as two metrics for OOD detection [24, 22, 9]. We compare the effectiveness of these two metrics for different medical datasets in section 4. To calculate scores, we drop the fully connected head and use normalised ViT backbone output as feature vector f for calculating evaluation scores. For each given test sample x , we calculate Mahalanobis distance based anomaly score, $\mathcal{S}_{md}(x)$, as

$$\mathcal{S}_{md}(x) := (f_{test} - \mu_m)^T \Sigma_m^{-1} (f_{test} - \mu_m) \quad (6)$$

where μ_m and Σ_m are the mean and covariance of the all feature vectors f_m from the training data, \mathcal{D}_{train} . We calculate the cosine similarity based anomaly score $\mathcal{S}_{cs}(x)$ for test sample x

$$\mathcal{S}_{cs}(x) := - \max_m \exp \left(\frac{f_{test}^T f_m}{\|f_{test}\| \|f_m\|} \right) \quad (7)$$

Detection is assessed with Area Under the Receiver Operating Characteristic curve (AUROC).

3 Experimental Setup

3.1 Dataset

We assess our model performance on three different health screening medical imaging benchmarks, chest X-ray images, colonoscopy images and fundus images for glaucoma detection.

RSNA. The Radiological Society of North America (RSNA) Pneumonia Detection Challenge dataset [23] is a publicly available dataset of frontal view chest radiographs. Each image was labeled as ‘‘Normal’’, ‘‘No Opacity/Not Normal’’ or ‘‘Opacity’’. The Opacity group consists of images with opacities suspicious for pneumonia, and images labeled ‘‘No Opacity/Not Normal’’ may have lung opacity but no opacity suspicious for pneumonia. The RSNA dataset contains 26, 684

X-rays with 8,851 normal, 11,821 no lung opacity/not normal and 6,012 lung opacity.

Hyper-Kvasir. The Hyper-Kvasir dataset is the largest public gastrointestinal dataset [1]. The data were collected during real examinations and partially labeled by experienced endoscopists. The dataset contains 110,079 images from patients, with 10,662 labelled images. Following [27] we take 2,100 images from “cecum”, “ileum” and “bbps-2-3” cases as normal and 1000 abnormal images from “polyp” as abnormal. We take 1,600 images for training set and 500 images for test set.

LAG. The LAG dataset is a large scale image dataset for glaucoma detection [14], containing 4,854 images with 1,711 positive glaucoma (abnormal) and 3,143 negative glaucoma (normal) scans. For consistent comparison, following [27], we take 2,343 normal images for training and 800 normal images and 1,711 abnormal images for testing.

Table 1: AUROC of OOD detection method trained on **RSNA** dataset

Method	$\mathcal{D}_{ood} : \text{Opacity}$		$\mathcal{D}_{ood} : \text{No Opacity}$	
<i>Unsupervised methods trained on normal samples</i>				
UAE [16]	0.89		0.78	
Deep AD [17]	0.838		0.704	
[9]	0.940		0.828	
Score	\mathcal{S}_{md}	\mathcal{S}_{cs}	\mathcal{S}_{md}	\mathcal{S}_{cs}
Ours	0.941	0.764	0.841	0.714

3.2 Auxiliary Dataset

For auxiliary dataset, we compare use of samples from ImageNet or from a in-domain one if any available. For RSNA dataset of X-ray images we use CheXpert [12] and for Hyper-Kvasir dataset of colonoscopy images all unlabeled Hyper-Kvasir images are taken as in-domain. For LAG dataset, we only use ImageNet due to unavailability of any in-domain dataset. We highlight that we do not use any label information to shape negative samples.

3.3 Training

Our proposed method has the same structure as DINO implementation. we use ViT-Small (ViT-S) backbone for all different training data. A patch size of 16 and $N=8$ local views for both positives and negatives, but two global positive views and one global negative view are used. All global views are resized to 256×256 while local views to 96×96 . The temperatures for teacher and student network are set to $\tau_t = 0.01$ and $\tau_s = 0.1$. During training, τ_t is linearly decreased from 0.04 to 0.01 in each epoch. λ and K are set to 1 and 4096 respectively for all our experiments. We use the Adamw optimizer [15] with an effective batch size of 256. For the base learning rate lr_{base} , we use 0.001 for Hyper-Kvasir and LAG datasets and 0.002 for RSNA. For each dataset, we trained the model for 700

epochs. We conducted our experiments using 4 NVIDIA-A100 GPUs with 40 GB of memory. The image augmentation pipeline \mathcal{T} is based on DINO except that for Hyper-Kvasir dataset we rotate all positive views with same randomly chosen angle to avoid information leak from position of existing green boxes in images. Finally, weight decay and learning rate are scaled with a cosine scheduler.

Table 2: AUROC results on **Hyper-Kvasir** and **LAG** datasets

Method	$\mathcal{D}_{train} : \text{Hyper-Kvasir}, \mathcal{D}_{ood} : \text{Polyp}$		$\mathcal{D}_{train} : \text{LAG}, \mathcal{D}_{ood} : \text{Glaucoma}$	
<i>Unsupervised methods trained on normal samples</i>				
CAVGA-R _u [28]	0.928		0.819	
IGD [6]	0.937		0.857	
CCD [27]	0.972		0.874	
Score	\mathcal{S}_{md}	\mathcal{S}_{cs}	\mathcal{S}_{md}	\mathcal{S}_{cs}
Ours	0.996	0.994	0.849	0.879

4 Experimental Results

We compare the proposed method with unsupervised methods trained on only healthy images. We report our results for both \mathcal{S}_{md} and \mathcal{S}_{cs} scores. In Table 1, on RSNA dataset, our method outperforms the contrastive self-supervised SOTA method [9] when taking \mathcal{S}_{md} as the anomaly score. In Table 2, we inspect the anomaly detection performance on the Hyper-Kvasir dataset for polyp detection and on the LAG dataset for glaucoma detection. Our method can surpass the recently proposed self-supervised anomaly detection method, CCD [27] on both polyp and glaucoma detection where we take \mathcal{S}_{md} and \mathcal{S}_{cs} respectively. In Table 3, the impact of different shifting transformations, as explained in section 2.2 is explored. We found out that transformations such as Rot have better performance than excluding negative samples or using non-effective transformations such as Pixel-Shuffle. This result supports our general assumption about a good negative transformation that changes high-level semantics and keeps low-level statistics. Note that for Hyper-Kvasir, positive views are rotated by the same angle randomly selected from $\mathcal{U}(\{90^\circ, 180^\circ, 270^\circ\})$ thus, we skip applying Rot-360 as a shifting transformation. In Fig. 2 [Left], we examine the effect of creating negative samples by applying shifting transformation on samples from each in-dist training, auxiliary dataset, or a combination of both. For RSNA and LAG dataset, as it is shown, the AUROC score increases where a combination of both is used, while for Hyper-Kvasir, we see no difference. Moreover, the use of only auxiliary datasets shows slightly better performance for RSNA compared to only taking in-dist negative samples on the other hand for LAG in-dist negative samples have higher score. The reason can be that for RSNA the in-domain auxiliary datasets are from a broader distribution compared to in-dist train data with a higher chance of resembling OOD samples but for LAG even though the ImageNet dataset has a broader distribution, in-dist negatives are harder negative samples which can be more advantageous [21]. The evaluation on taking in-domain or out-domain auxiliary datasets is shown in Fig. 2

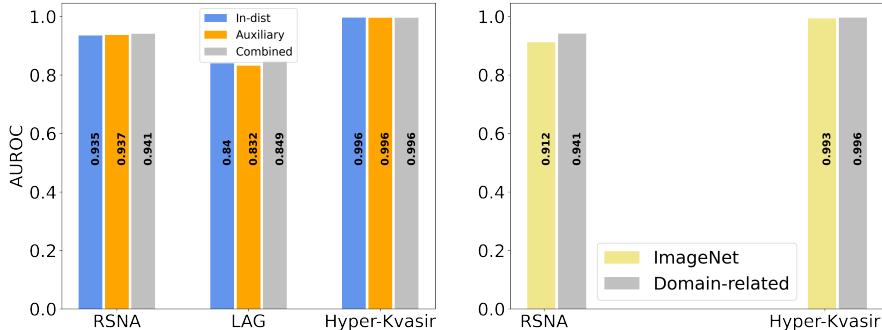


Fig. 2: **Left.** AUROC results based on \mathcal{S}_{md} for different negative sets where generated from in-dist train data, an auxiliary dataset or a combination of both. **Right.** AUROC results across different auxiliary datasets where we take images from an in-domain medical dataset or out-domain.

[Right]. For RSNA X-ray images, OOD detection performance is improved by a large margin when negative samples are from an in-domain auxiliary set. However, for Hyper-Kvasir, the out-domain auxiliary has approximately the same performance as the in-domain.

Table 3: The impact of different shifting transformations on AUROC results. Reported scores are for \mathcal{S}_{md} (\mathcal{S}_{cs}).

In-dist Dataset	NoNeg	Shifting transformations				
		Rot	Rot-360	Perm-4	Perm-16	Pixel-Shuffle
RSNA	0.925(0.888)	0.941(0.764)	0.933(0.766)	0.924(0.634)	0.908(0.887)	0.925(0.733)
LAG	0.799(0.862)	0.849(0.879)	0.831(0.866)	0.807(0.873)	0.814(0.881)	0.797(0.860)
Hyper-Kvasir	0.974(0.875)	0.989(0.915)	–	0.996(0.994)	0.985(0.960)	0.994(0.985)

5 Conclusion

In this study, we present a self-supervised method which leverages self-distillation and negative samples for the task of abnormality detection without accessing label information. We study different ways of creating negative samples by applying shifting transformations on in-dist training data, an auxiliary dataset, or a combination of both. Additionally, we compare the impact of having auxiliary samples from domain-related distribution or from a different domain such as ImageNet. Moreover, we compare the abnormality detection performance using two different evaluation metrics including cosine similarity and Mahalanobis distance. A major motivation behind this work is that we take only normal samples during training which makes our method suitable for yet unknown abnormalities. In anomaly detection, our method outperforms SOTA methods on the RSNA, Hyper-Kvasir and LAG datasets.

References

1. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data* **7**(1), 1–14 (2020)
2. Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K.G., Murphy, K.: Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis* p. 102125 (2021)
3. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882* (2020)
4. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294* (2021)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *International conference on machine learning*. pp. 1597–1607. PMLR (2020)
6. Chen, Y., Tian, Y., Pang, G., Carneiro, G.: Unsupervised anomaly detection and localisation with multi-scale interpolated gaussian descriptors. *arXiv e-prints* pp. arXiv–2101 (2021)
7. Davletshina, D., Melnychuk, V., Tran, V., Singla, H., Berrendorf, M., Faerman, E., Fromm, M., Schubert, M.: Unsupervised anomaly detection for x-ray images (2020)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations* (2021)
9. Gholamipour, R., Rafiee, N., Kollmann, M.: Pneumonia detection with semantic similarity scores (2021)
10. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. *Advances in neural information processing systems* **31** (2018)
11. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606* (2018)
12. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison (2019)
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013)
14. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: a large-scale database and cnn model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10571–10580 (2019)
15. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)
16. Mao, Y., Xue, F.F., Wang, R., Zhang, J., Zheng, W.S., Liu, H.: Abnormality detection in chest x-ray images using uncertainty prediction autoencoders. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 529–538. Springer (2020)
17. Nakao, T., Hanaoka, S., Nomura, Y., Murata, M., Takenaga, T., Miki, S., Watadani, T., Yoshikawa, T., Hayashi, N., Abe, O.: Unsupervised deep anomaly detection in chest radiographs. *Journal of Digital Imaging* pp. 1–10 (2021)

18. Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? (2019)
19. Perera, P., Nallapati, R., Xiang, B.: Ocgan: One-class novelty detection using gans with constrained latent representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2898–2906 (2019)
20. Rafiee, N., Gholamipoorfarid, R., Adaloglou, N., Jaxy, S., Ramakers, J., Kollmann, M.: Self-supervised anomaly detection by self-distillation and negative sampling. arXiv preprint arXiv:2201.06378 (2022)
21. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. arXiv preprint arXiv:2010.04592 (2020)
22. Sehwal, V., Chiang, M., Mittal, P.: Ssd: A unified framework for self-supervised outlier detection. arXiv preprint arXiv:2103.12051 (2021)
23. Shih, G., Wu, C.C., Halabi, S., Kohli, M., Prevedello, L., Cook, T., Sharma, A., Amorosa, J., Arteaga, V., Galperin-Aizenberg, M., Gill, R., Godoy, M., Hobbs, S., Jeudy, J., Laroia, A., Shah, P., Vummidi, D., Yaddanapudi, K., Stein, A.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology. Artificial intelligence* **1** 1, e180041 (2019)
24. Tack, J., Mo, S., Jeong, J., Shin, J.: Csi: Novelty detection via contrastive learning on distributionally shifted instances. In: 34th Conference on Neural Information Processing Systems (NeurIPS) 2020. vol. 33, pp. 11839–11852 (2020)
25. Tang, Y.X., Tang, Y.B., Peng, Y., Yan, K., Bagheri, M., Redd, B.A., Brandon, C.J., Lu, Z., Han, M., Xiao, J., Summers, R.M.: Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digital Medicine* **3**(1), 1–8 (2020)
26. Tang, Y., Tang, Y., Han, M., Xiao, J., Summers, R.M.: Abnormal chest x-ray identification with generative adversarial one-class classifier (2019)
27. Tian, Y., Pang, G., Liu, F., Chen, Y., Shin, S.H., Verjans, J.W., Singh, R., Carneiro, G.: Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 128–140. Springer (2021)
28. Venkataramanan, S., Peng, K.C., Singh, R.V., Mahalanobis, A.: Attention guided anomaly localization in images. In: European Conference on Computer Vision. pp. 485–503. Springer (2020)
29. Wang, Y., Feng, Z., Song, L., Liu, X., Liu, S.: Multiclassification of endoscopic colonoscopy images based on deep transfer learning. *Computational and Mathematical Methods in Medicine* **2021**
30. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text (2020)

Chapter 9

Conclusion and Future Work

In this dissertation, we have mainly focused on the problem of Out-of-Distribution (OOD) detection using self-supervised representation learning in different domains. OOD detection is the problem of deciding whether a given test sample is drawn from the same in-distribution as a given training set or belongs to an alternative distribution. Many real-world applications require highly accurate OOD detection for secure deployment. Recently many studies have been conducted for OOD detection. Despite the promising results, these approaches usually rely on annotated abnormal samples that are either not available or only a limited number are available. OOD detection requires identifying features specific to the in-distribution. In the absence of labels, these features can be learned by self-supervised techniques under the generic assumption that the most abstract features are those which are statistically most over-represented in comparison to other distributions from the same domain. The main contribution of this work is our proposed framework for anomaly detection which outperforms supervised and unsupervised methods on challenging OOD detection tasks. The proposed method does not require any label information and can be widely applied to OOD detection tasks, including visual and time series data.

More precisely, in chapter 5, we addressed the Atrial Fibrillation (AF) detection problem which is the most common arrhythmia; however, detection of asymptomatic AF is a challenging task. We aim to evaluate the sensitivity and specificity of non-invasive AF detection by a medical wearable. We applied different algorithms to five-minute periods of inter-beat intervals (IBI) for the AF detection. A Deep Neural Network (DNN) is trained unsupervised to extract relevant features for AF detection. The training objective is given by maximising the Mutual Information (MI) between IBI values that are separated by a randomly chosen time point within the five-minute period. Unsupervised feature extraction followed by an unsupervised classification results in higher sensitivity and specificity compared with

normalised root mean square of the successive difference (nRMSSD) an established metric for the AF detection.

In chapter 6, we proposed a self-supervised contrastive learning framework for anomaly detection in chest X-ray images. X-ray images have been widely used for medical diagnoses. The anomaly detector is trained only on the normal (i.e., healthy) images to make our method future-ready for yet unknown anomalies. Given unlabeled training data, a feature extractor is trained using the self-supervised contrastive loss to pull together different augmented views of the same image and push augmented views of different images away. The learnt self-supervised representations are highly effective for the task of pneumonia detection in our framework. We defined an anomaly detection score based on Mahalanobis distance applicable for detecting anomalies. We found that our approach outperforms all previous unsupervised methods on a pneumonia detection challenge dataset. Future work concerns deeper analysis of taken data augmentations and different evaluation metrics suitable for anomaly detection in X-ray images.

Chapter 7 shows how we can make use of self-supervised contrastive learning combined with cosine similarity as a score function to detect serious clinical complications (SCC) in patients receiving oncological treatment for their hematologic malignancies. In fact, nearly every patient on such treatment protocols experiences at least one SCC requiring treatment. Early diagnosis of SCC is not only of high clinical relevance for the safety and well-being of the patients as it enables a more rapid treatment of SCC, but it would also potentially help reduce the number of hospitalisations. Continuous monitoring of vital signs by means of medical wearables will potentially lead to an earlier diagnosis and better treatment. We aim to evaluate whether wearable-based monitoring enables detection of SCC with sufficient reliability. To do so, we take an OOD detection method to identify whether defined episodes of vital signs are "regular" (= absence of SCC) or not in order to detect clinical complications. To learn statistical relevant features for SCC detection, self-supervised contrastive learning is used by ensuring that in the representation space embeddings of similar inputs are pulled closer while simultaneously embeddings from dissimilar inputs are pushed apart. In the case of object classification as computer vision task, the generalizing features are typically defined as those that are invariant under image transformations that keep the semantics of the shown object, such as horizontal flip, cropping, and slight changes in colouring. For the time series data used in this work, we define generalizing features as the information shared between two time series samples. Alternative ways of defining generalizing features could be more explored in future work. We show that wearable-based remote patient monitoring combined with a DNN model enables calculation of a SCC Score that allows for detection and prediction of SCCs in patients receiving intensive treatment for haematological malignancies.

In chapter 8, we showed that self-distillation of the in-distribution training set together with contrasting against negative examples derived from shifting transformation of auxiliary data strongly improves OOD detection. We found that this improvement depends on how the negative samples are generated. In particular, we observed that by leveraging negative samples, which keep the statistics of low-level features while changing the high-level semantics, higher average detection performance is obtained. The novelty of our work is the use of sensitivity scores to find optimized negative sampling strategies and hyperparameters in absence of any OOD validation set. Our approach aims to draw a tight, not necessary simply connected, decision boundary between the in-distribution and an auxiliary negative distribution. In our proposed model, a decision boundary is formed during training unlike other contrastive methods which require some modifications during the evaluation phase to improve the discrimination of in-distribution and OOD samples. The efficiency of our approach is demonstrated across a diverse range of OOD detection problems, setting new benchmarks for unsupervised OOD detection in the visual domain in both natural and medical images. Despite different experiments to create the negative examples, many different adaptations and tests have been left for the future.

We also leveraged self-distillation and negative samples for anomaly detection in medical images. We conducted different experiments to demonstrate the selection of negative sample strategy and the evaluation metrics on several disease diagnosis tasks. A systematic way to find the best evaluation metric could be more explored for the future. The results show that our proposed framework outperforms state-of-the-art unsupervised methods.

Chapter 10

Publications

2020

1. Jacobsen M, Dembek TA, Ziakos AP, Gholamipoor R, Kobbe G, Kollmann M, Blum C, Müller-Wieland D, Napp A, Heinemann L, Deubner N, Marx N, Isenmann S, Seyfarth M. Reliable Detection of Atrial Fibrillation with a Medical Wearable during Inpatient Conditions. *Sensors*, 2020.

Contributions: The author contributed with designing and implementation of deep neural network, training, evaluation, and visualization. The author contributed with writing parts related to DNN-based algorithm under the supervision of Prof. Dr. Markus Kollmann. **Status:** Published.

2022

2. Malte Jacobsen, Pauline Rottmann, Till A. Dembek, Anna L. Gerke, Rahil Gholamipoor, Christopher Blum, Niels-Ulrik Hartmann, Marlo Verket, Jennifer Kaivers, Paul Jäger, Ben-Niklas Baermann, Lutz Heinemann, Nikolaus Marx, Dirk Müller-Wieland, Markus Kollmann, Melchior Seyfarth, and Guido Kobbe. Feasibility of Wearable-Based Remote Monitoring in Patients During Intensive Treatment for Aggressive Hematologic Malignancies. *JCO Clinical Cancer Informatics*, 2022.

Contributions: The author contributed with collection and assembly of data and data analysis and interpretation. **Status:** Published.

3. R. Gholamipoor, N. Rafiee, M. Kollmann. Pneumonia Detection with Semantic Similarity Scores. *ISBI*, 2022.

Contributions: The research and preparation of this manuscript were done jointly by R. Gholamipoor and N. Rafiee under the supervision of Prof. Dr. Markus Kollmann. **Status:** Published.

4. Malte Jacobsen, Rahil Gholamipoor, Till A. Dembek, Pauline Rottmann, Marlo Verket, Julia Brandts, Paul Jäger, Ben Niklas Baermann, Lutz Heinemann, Anna L. Gerke, Nikolaus Marx, Dirk Müller-Wieland, Kathrin Moellenhoff, Markus Kollmann, Melchior Seyfarth, Guido Kobbe. 2022.

Contributions: The author contributed with the methodology, implementation of the anomaly detection algorithm, evaluation, visualization, and describing the algorithm under the supervision of Prof. Dr. Markus Kollmann. **Status:** Submitted to *Lancet Digital Health*.

5. Nima Rafiee, Rahil Gholamipoor, Nikolas Adaloglou, Simon Jaxy, Julius Ramakers, Markus Kollmann. Self-Supervised Anomaly Detection by Self-Distillation and Negative Sampling. *ICANN*, 2022.

Contributions: The author contributed with evaluation, visualization, and writing under the supervision of Prof. Dr. Markus Kollmann. **Status:** Published.

6. Nima Rafiee, Rahil Gholamipoor, Markus Kollmann. Abnormality detection for medical images using self-supervision and negative samples. 2022.

Contributions: The research and preparation of this manuscript were done jointly by Nima Rafiee and Rahil Gholamipoor under the supervision of Prof. Dr. Markus Kollmann. **Status:** Submitted to *MICCAI*.

Appendix A

Appendix of Chapter 5

A.1 Alternative Derivation of I_{NCE}

Here we present a simpler derivation of I_{NCE} than [17]. Let us take a general form of contrastive loss as follows

$$\mathcal{L}_{NCE} = -\mathbb{E}_{(x,x')\sim p(x,x')} \left[\log \frac{Q(x,x')}{\mathbb{E}_{x_{neg}\sim p(x_{neg})} [Q(x,x_{neg})]} \right] \quad (\text{A.1})$$

where $p(x,x')$ is the joint probability and $p(x)$ is the marginal of $p(x,x')$. \mathcal{L}_{NCE} is lower bounded by the negative of the mutual information $I_{NCE}(X; X')$. We first prove that the minimum loss can be achieved for $Q^*(x,x') = \frac{p(x,x')}{p(x)p(x')}$.

By using variational calculus [2], let $Q(x,x') = Q^*(x,x') + \alpha\eta(x,x')$, where $Q^*(x,x')$ is the optimal function $Q(x,x')$, α is a scalar value, and $\eta(x,x')$ is any arbitrary function that depends on x and x' . By definition, \mathcal{L}_{NCE} has a minimum in $\alpha = 0$. Thus, if $\mathcal{L}_{NCE}[Q]$ is differentiable, its derivative with respect to α vanishes in $\alpha = 0$. Substituting Q in Eq. A.1

$$\begin{aligned} \mathcal{L}_{NCE} &= -\int_x \int_{x'} p(x,x') \left[\log \frac{Q^*(x,x') + \alpha\eta(x,x')}{\int p(x_{neg}) [Q^*(x,x_{neg}) + \alpha\eta(x,x_{neg})] dx_{neg}} \right] dx' dx \\ \frac{\partial \mathcal{L}_{NCE}}{\partial \alpha} \Big|_{\alpha=0} &= -\int_x \int_{x'} p(x,x') \left[\frac{\eta(x,x')}{Q^*(x,x')} - \frac{\int_{x_{neg}} p(x_{neg})\eta(x,x_{neg}) dx_{neg}}{\int_{x_{neg}} p(x_{neg})Q^*(x,x_{neg}) dx_{neg}} \right] dx' dx \\ &= -\int_x \int_{x'} \int_{x_{neg}} p(x,x') \left[\frac{\eta(x,x')}{Q^*(x,x')} - \frac{p(x_{neg})\eta(x,x_{neg})}{\int p(x_{neg})Q^*(x,x_{neg}) dx_{neg}} \right] dx_{neg} dx' dx \\ &= -\int_x \int_{x'} \int_{x_{neg}} p(x,x')\eta(x,x_{neg}) \left[\frac{\delta(x' - x_{neg})}{Q^*(x,x')} - \frac{p(x_{neg})}{\int p(x_{neg})Q^*(x,x_{neg}) dx_{neg}} \right] dx_{neg} dx' dx \end{aligned}$$

where $\delta(x)$ is the Dirac delta function, and is defined as follows

$$\psi(0) = \int_{-\infty}^{\infty} \delta(x)\psi(x)dx$$

where ψ is a smooth function and has as many derivatives as required [80]. The delta function, $\delta(x)$, is constrained to satisfy

$$\int_{-\infty}^{\infty} \delta(x)dx = 1$$

then $\frac{\partial \mathcal{L}_{NCE}}{\partial \alpha}|_{\alpha=0}$ vanishes for all functions η . Because η is arbitrary, this can occur only if

$$\frac{\partial \mathcal{L}_{NCE}}{\partial \alpha}|_{\alpha=0} = 0 \iff \int_{x'} p(x, x') \left[\frac{\delta(x' - x_{neg})}{Q^*(x, x')} - \frac{p(x_{neg})}{\int p(x_{neg})Q^*(x, x_{neg})dx_{neg}} \right] = 0$$

$$\frac{p(x, x_{neg})}{p(x)p(x_{neg})} = \frac{Q^*(x, x_{neg})}{\int p(x_{neg})Q^*(x, x_{neg})dx_{neg}} \quad (\text{A.2})$$

where $Q^*(x, x') = \frac{p(x, x')}{p(x)p(x')}$. By substituting Q^* in Eq. A.1, we have

$$\mathcal{L}_{NCE} = - \int_x \int_{x'} p(x, x') \log \frac{p(x, x')}{p(x)p(x')} dx' dx = -I(X; X') \quad (\text{A.3})$$

Bibliography

- [1] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, “Self-supervised learning: Generative or contrastive,” *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [4] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” in *6th International Conference on Learning Representations, ICLR*, 2018.
- [5] C. Doersch, A. Gupta, and A. A. Efros, “Unsupervised visual representation learning by context prediction,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.
- [6] A. Coates, A. Ng, and H. Lee, “An analysis of single-layer networks in unsupervised feature learning,” in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 215–223.
- [7] D. Alexey, P. Fischer, J. Tobias, M. R. Springenberg, and T. Brox, “Discriminative, unsupervised feature learning with exemplar convolutional, neural networks,” *IEEE TPAMI*, vol. 38, no. 9, pp. 1734–1747, 2016.
- [8] M. Norouzi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in *European conference on computer vision*. Springer, 2016, pp. 69–84.
- [9] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in *European conference on computer vision*. Springer, 2016, pp. 649–666.
- [10] T. Wang and P. Isola, “Understanding contrastive representation learning through alignment and uniformity on the hypersphere,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 9929–9939.
- [11] A. Krizhevsky, G. Hinton *et al.*, “Learning multiple layers of features from tiny images,” 2009.
- [12] A. Goyal and Y. Bengio, “Inductive biases for deep learning of higher-level cognition,” *arXiv preprint arXiv:2011.15091*, 2020.
- [13] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Pearson, 2020.
- [14] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*. Cambridge University Press, 2020.

-
- [15] Y. Bengio, A. C. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [16] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, “Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness,” *arXiv preprint arXiv:1811.12231*, 2018.
- [17] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [18] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, “Learning deep representations by mutual information estimation and maximization,” in *7th International Conference on Learning Representations, ICLR*, 2019.
- [19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.
- [20] R. Linsker, “Self-organization in a perceptual network.” *IEEE Computer*, vol. 21, pp. 105–117, 03 1988.
- [21] J. B. Kinney and G. S. Atwal, “Equitability, mutual information, and the maximal information coefficient,” *Proceedings of the National Academy of Sciences*, vol. 111, no. 9, pp. 3354–3359, 2014.
- [22] L. Paninski, “Estimation of entropy and mutual information,” *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003.
- [23] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, “Mutual information neural estimation,” in *International conference on machine learning*. PMLR, 2018, pp. 531–540.
- [24] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker, “On variational bounds of mutual information,” in *Proceedings of the 36th International Conference on Machine Learning, ICML*, 2019.
- [25] A. Kolesnikov, X. Zhai, and L. Beyer, “Revisiting self-supervised visual representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1920–1929.
- [26] J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque, “Input complexity and out-of-distribution detection with likelihood-based generative models,” in *8th International Conference on Learning Representations, ICLR*, 2020.
- [27] E. T. Nalisnick, A. Matsukawa, Y. W. Teh, and B. Lakshminarayanan, “Detecting out-of-distribution inputs to deep generative models using a test for typicality.” 2019.
- [28] E. T. Nalisnick, A. Matsukawa, Y. W. Teh, D. Görür, and B. Lakshminarayanan, “Do deep generative models know what they don’t know?” in *International Conference on Learning Representations*, 2019.
- [29] W. Falcon and K. Cho, “A framework for contrastive self-supervised learning and designing a new approach,” *arXiv preprint arXiv:2009.00104*, 2020.

- [30] J. Tack, S. Mo, J. Jeong, and J. Shin, “Csi: Novelty detection via contrastive learning on distributionally shifted instances,” *Advances in neural information processing systems*, vol. 33, pp. 11 839–11 852, 2020.
- [31] V. Sehwag, M. Chiang, and P. Mittal, “SSD: A unified framework for self-supervised outlier detection,” in *9th International Conference on Learning Representations, ICLR*, 2021.
- [32] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *nature*, vol. 542, no. 7639, pp. 115–118, 2017.
- [33] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu, “Attention based glaucoma detection: a large-scale database and cnn model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 571–10 580.
- [34] Y. Tian, G. Pang, F. Liu, Y. Chen, S. H. Shin, J. W. Verjans, R. Singh, and G. Carneiro, “Constrained contrastive distribution learning for unsupervised anomaly detection and localisation in medical images,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 128–140.
- [35] Z. Li, K. Kamnitsas, and B. Glocker, “Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 402–410.
- [36] Y. Tian, G. Maicas, L. Z. C. T. Pu, R. Singh, J. W. Verjans, and G. Carneiro, “Few-shot anomaly detection for polyp frames from colonoscopy,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 274–284.
- [37] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, “Deep one-class classification,” in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.
- [38] Y. Chen, Y. Tian, G. Pang, and G. Carneiro, “Unsupervised anomaly detection and localisation with multi-scale interpolated gaussian descriptors,” *arXiv e-prints*, pp. arXiv–2101, 2021.
- [39] I. Golan and R. El-Yaniv, “Deep anomaly detection using geometric transformations,” *Advances in neural information processing systems*, vol. 31, 2018.
- [40] T. M. Cover, *Elements of information theory*. Wiley, 1999.
- [41] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [42] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An introduction to variational methods for graphical models,” *Machine learning*, vol. 37, no. 2, pp. 183–233, 1999.
- [43] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [44] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” *arXiv preprint physics/0004057*, 2000.
- [45] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain.” *Psychological review*, vol. 65, no. 6, p. 386, 1958.

-
- [46] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.
- [47] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [48] A. Y. Ng, “Feature selection, l_1 vs. l_2 regularization, and rotational invariance,” in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 78.
- [49] R. Heckel and F. F. Yilmaz, “Early stopping in deep networks: Double descent and how to eliminate it,” in *9th International Conference on Learning Representations, ICLR*, 2021.
- [50] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “RandAugment: Practical automated data augmentation with a reduced search space,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 702–703.
- [51] S. Edunov, M. Ott, M. Auli, and D. Grangier, “Understanding back-translation at scale,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.
- [52] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [53] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2020.
- [54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [55] H.-Y. Zhou, C. Lu, S. Yang, and Y. Yu, “Convnets vs. transformers: Whose visual representations are more transferable?” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2230–2238.
- [56] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *International Conference on Machine Learning*, 2021, pp. 10 347–10 357.
- [57] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, “How to train your vit? data, augmentation, and regularization in vision transformers,” *arXiv preprint arXiv:2106.10270*, 2021.
- [58] G. Larsson, M. Maire, and G. Shakhnarovich, “Learning representations for automatic colorization,” in *European conference on computer vision*. Springer, 2016, pp. 577–593.
- [59] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, vol. 2. IEEE, 2006, pp. 1735–1742.
- [60] M. Gutmann and A. Hyvärinen, “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models,” in *Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings*, 2010, pp. 297–304.

- [61] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning, ICML*, 2020.
- [62] Y. Tian, D. Krishnan, and P. Isola, “Contrastive multiview coding,” in *ECCV*, 2020.
- [63] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, “What makes for good views for contrastive learning?” in *Annual Conference on Neural Information Processing Systems 2020, NeurIPS*, 2020.
- [64] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, “Unsupervised embedding learning via invariant and spreading instance feature,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6210–6219.
- [65] O. Henaff, “Data-efficient image recognition with contrastive predictive coding,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 4182–4192.
- [66] P. Bachman, R. D. Hjelm, and W. Buchwalter, “Learning representations by maximizing mutual information across views,” in *Advances in Neural Information Processing Systems*, 2019, pp. 15 535–15 545.
- [67] M. Tschannen, J. Djolonga, P. K. Rubenstein, S. Gelly, and M. Lucic, “On mutual information maximization for representation learning,” in *8th International Conference on Learning Representations, ICLR*, 2020.
- [68] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *Advances in neural information processing systems*, vol. 26, 2013.
- [69] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, “Normface: L2 hypersphere embedding for face verification,” in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1041–1049.
- [70] F. Wang and H. Liu, “Understanding the behaviour of contrastive loss,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2495–2504.
- [71] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Supervised contrastive learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.
- [72] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.
- [73] J. D. Robinson, C. Chuang, S. Sra, and S. Jegelka, “Contrastive learning with hard negative samples,” in *9th International Conference on Learning Representations, ICLR*, 2021.
- [74] R. Zhu, B. Zhao, J. Liu, Z. Sun, and C. W. Chen, “Improving contrastive learning by visualizing feature transformation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 306–10 315.
- [75] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.

- [76] M. Gutmann and A. Hyvärinen, “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models,” in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.
- [77] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2014.
- [78] S. J. Reddi, S. Kale, and S. Kumar, “On the convergence of adam and beyond,” in *6th International Conference on Learning Representations, ICLR*, 2018.
- [79] S. Hong, Y. Xu, A. Khare, S. Priambada, K. Maher, A. Aljiffry, J. Sun, and A. Tumanov, “Holmes: health online model ensemble serving for deep learning models in intensive care units,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1614–1624.
- [80] R. S. Strichartz, *A guide to distribution theory and Fourier transforms*. World Scientific Publishing Company, 2003.