Data-driven Disease Assessment from time-resolved Fluorescence Optical Imaging

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften (Dr. rer. nat.)

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

vorgelegt von

Marc A. Osterland

Berlin, 2019

Erstgutachter: Prof. Dr. Christof Schütte
Zweitgutachter: Prof. Dr.-Ing. Anja Hennemuth

Tag der Disputation: 13. Juli 2021

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe. Ich erkläre weiterhin, dass ich die vorliegende Arbeit oder deren Inhalt nicht in einem früheren Promotionsverfahren eingereicht habe.

Marc A. Osterland
Berlin, 10 December 2019

Acknowledgements

This thesis and all accompanying research have been conducted in the Visual Data Analysis department at the Zuse Institute Berlin (ZIB) and the Biocomputing group at the Freie Universität Berlin (FU). Many people supported me during the last five years, and without them you would not be reading this.

I would like to thank Christof Schütte for giving me the opportunity to work at FU and ZIB to pursue my research. His broad experience and knowledge inspired me and guided me through this thesis. Furthermore, I owe a huge thanks to Steffen Prohaska and Tim Conrad for including me in their research groups and supporting me in-depth with all my technical and methodological problems. Their critical questions and remarks provided important input and challenged my scientific thinking.

Equally important for the development of this thesis is the social support, that I experienced in the past years. The discussions with my colleagues were both, helpful and uplifting. The people from the Rotunda terrace helped me releasing my frustration and cheered me up when necessary. Very creative brainstorming as well as scientific and non-scientific exchange was enabled by many people from the annual Winterseminar. Furthermore, I want to namely thank Marcus Weber and Stefan Klus for supporting me with my questions about linear algebra.

Outside of the Dahlem community I was supported emotionally by my family, friends, flat mates, and Daphne. They helped me relaxing, listened to my concerns, and were there for me when I needed them.

Thank you!

Contents

List of Figures

Abstract

Fluorescence Optical Imaging (FOI) is a new method to assess Rheumatoid Arthritis, Psoriasis, and other inflammatory diseases. It can reveal inflammatory tissues and microcirculatory disorders with high spatial and temporal resolution. However, the analysis of the image data is currently performed manually with no or weak consideration of the time component. To date, there is no automatic image analysis pipeline for inflammatory diseases based on FOI data. Furthermore, the distinct phenotypes of Rheumatoid Arthritis, Psoriatic Arthritis, and comparable diseases are not fully described in FOI, yet. This thesis proposes a new unsupervised, data-driven approach, that enables disease assessment of inflammatory diseases of the hands under the unfavorable conditions (e.g., low data-availability) of medical imaging. Data-driven methods such as deep neural networks often require extensive and well-annotated data sets, which are rare and expensive in clinical research. The here presented approach uses a Variational Autoencoder and reduces the complexity of the problem by learning a low-dimensional latent space. This latent space enables further analyses such as data exploration, subgroup classification, and analysis of the underlying dynamics under low data-availability and low quality of clinical labels. For data exploration, subgroups can be summarized in latent space and then be decoded back into an image. This feature-wise average results in superior images compared to the pixel-wise average. Furthermore, the latent space allows for cluster identification by employing two-dimensional projections such as UMAP. The latent space representations enable classification tasks under low data-availability with a two-step approach. Therefore, extensions of the proposed model using Neural Networks and Random Forests are evaluated and compared. The approach can distinguish between Psoriasis Vulgaris and Psoriatic Arthritis with accuracies over 70%. On synthetical data, accuracies of up to 97% are achieved. In combination with the Koopman Operator Theory the underlying dynamics can be approximated linearly. This approach decomposes the temporal effects within the data, and it enables subgroup comparisons and outlier detection. This thesis investigates the application of the proposed pipeline with respect to the quality of the underlying data and discusses the necessary conditions to learn a generalizing model. The dependency of high-quality labels for supervised approaches is demonstrated with synthetical and clinical datasets.

Zusammenfassung

Optische Fluoreszenz Bildgebung (Fluorescence Optical Imaging, FOI) ist ein neues Verfahren, mit dessen Hilfe entzündliche Gelenkkrankheiten wie rheumatische Arthritis und Psoriasis bewertet werden können. Entzündetes Gewebe und Störungen der Mikrozirkulation können mit hoher räumlicher und zeitlicher Auflösung dargestellt werden. Die Analyse dieser Bilddaten erfolgt zurzeit jedoch hauptsächlich manuell ohne Berücksichtigung der Zeitkomponente. Derzeit existiert kein automatischer Bildanalyseprozess für entzündliche Gelenkerkrankungen auf Basis von FOI-Daten. Darüber hinaus sind die charakteristischen Phänotypen der einzelnen Krankheiten wie rheumatischer Arthritis oder Psoriasis noch nicht vollständig für FOI beschrieben. Diese Arbeit präsentiert einen neuen datengetriebenen Ansatz mit Hilfe des unüberwachten Lernens, der eine Krankheitsbewertung bei entzündlichen Gelenkerkrankungen der Hände auch unter den ungünstigen Bedingungen (z.B. geringe Datenverfügbarkeit) der medizinischen Bildgebung ermöglicht. Datengetriebene Methoden wie die Tiefen Neuronalen Netzwerke erfordern häufig eine große Menge an gut annotierten Daten, die in der klinischen Forschung selten und teuer sind. Der hier vorgestellte Ansatz benutzt einen Variational Autoencoder, um einen niedrigdimensionalen latenten Raum zu lernen, der die Komplexität des ursprünglichen Problems drastisch reduziert. Dieser latente Raum ermöglichte somit weitere Auswertungen, darunter die Datenexploration, Klassifikation von Teilgruppen sowie die Analyse der zugrundeliegenden Dynamiken, auch wenn die Datenverfügbarkeit und die Qualität der Zielvariablen gering sind. Zur Datenexploration können Teilgruppen im latenten Raum zusammengefasst und wieder in ein Bild übersetzt werden. Diese Bilder auf Basis eines Durchschnitts der Merkmale sind Bildern eines pixelbasierten Durchschnitts überlegen. Darüber hinaus können Cluster von ähnlichen Patienten in einer zwei-dimensionalen Projektion mittels UMAP identifiziert werden. In einem Zwei-Schritt-Verfahren ermöglicht die Darstellungen im latenten Raum die Klassifikation von Teilgruppen, auch wenn die Verfügbarkeit von Zielvariablen eingeschränkt ist. Dafür wird das vorgeschlagene Modell um ein weiteres Neuronales Netzwerk oder einen Random Forest erweitert und evaluiert. Dieser Ansatz kann zwischen Psoriasis Vulgaris und Psoriatischer Arthritis mit einer Genauigkeit von über 70% unterscheiden. Auf synthetischen FOI-Daten werden bis zu 97% erreicht. In Anlehnung an die Theorie der Koopman Operatoren können die zugrundeliegenden Dynamiken linear approximiert werden. Dieser Ansatz zerlegt die unterschiedlichen zeitlichen Effekte innerhalb der Daten und ermöglicht so Vergleiche von Teilgruppen, sowie die Detektion von Ausreißern. Diese Arbeit untersucht die Anwendung des vorgeschlagenen Ansatzes in Bezug

auf die Qualität der zugrundeliegenden Daten und diskutiert die notwendigen Bedingungen, um ein verallgemeinerndes Modell zu lernen. Die Abhängigkeit von hochwertigen Zielvariablen für überwachte Ansätze wird an synthetischen und klinischen Datensätzen veranschaulicht.

Abbreviations

- CNN = Convolutional Neural Network
- DAS28 = Disease Activity Score 28
- FOI = Fluorescence Optical Imaging
- GAP = Global Average Pooling
- MCP = metacarpophalangeal joints
- PsV, PsA = Psoriasis Vulgaris, Psoriatic Arthritis
- PIP = proximal interphalangeal joints
- RA = Rheumatoid Arthritis
- TJC28 = Tender Joint Count 28
- VAE = Variational Autoencoder

1 Introduction

1.1 Medical Imaging

Medical Imaging has become indispensable to modern medicine and is an essential tool for diagnostics [1]. A variety of different imaging techniques enables physicians a non-invasive view into the human body. Depending on the physical principle behind the technique, it is possible to reveal anatomical and functional features. Computer tomography (CT) is based on electromagnetic radiation (X-rays) and has its strength in highlighting different tissue densities. It is thus ideal for revealing fractures of bones and internal bleedings. In oncology, it is often combined with contrast fluids and Positron-emission tomography (PET) to detect tumors and metastases. Magnetic resonance imaging (MRI) is a radiation-free method that enables volumetric imaging of soft tissues. It is primarily used to detect tumors, inflammatory tissue, as well as lesions of cartilage and tendons. For conditions like bone fractures, medical imaging can be the primary tool for the diagnosis. However, for many conditions and diseases, additional diagnostic procedures play an equally important role, such as histopathological, molecular biological, and physical examinations.

With the costs decreasing, the popularity of these methods is increasing for clinical diagnosis, as well as for research [2] [3]. This means that for every patient more and more data can be acquired. Thus, Medical Imaging potentially enables physicians to make holistic, differential, and personalized diagnoses. However, this also means that physicians have to analyze, assess, and consider more data per patient. Thus, computer-aided diagnostics (CAD) has become increasingly important to modern medicine as well. However, the algorithmic analysis of medical imaging data is not trivial, and can erroneous results can occur unnoticed [4].

This thesis aims to underline the importance and the potential, as well as the problems and challenges of a new type of medical imaging - fluorescence optical imaging - and its analysis in the context of inflammatory diseases of the joints.

1.2 Inflammatory Autoimmune Diseases of the Joints

In numerous conditions, the immune system is attacking normal, healthy tissues of the human body. These so-called autoimmune diseases are highly prevalent in the population [5]. The affected tissues, severity, and the impact on the individual patients vary strongly between the diseases. Diseases like Multiple sclerosis affect mainly neurons and nerval tissues and result in sensory and motoric defects, whereas the Coeliac disease affects the tissues of the small intestine, which results in gastrointestinal problems.

This thesis focuses on conditions and disorders that affect joints and their surrounding tissues, such as cartilage, joint capsule, and bone. The joints of the human body are essential to its mobility and flexibility. Joint-affecting disorders can thus profoundly impair the mobility and quality of life. Many of these diseases share the underlying mechanism. During a process called citrullination, which is a normal physiological post-transcriptional process in dying cells, peptidylarginine deiminase (PAD) converts arginine to citrulline. Citrullinated proteins can be a target for the immune system. However, due to genetic mutations and environmental factors, citrullination can happen far more frequently in healthy, non-dying cells making them a permanent target for the immune system.

1.2.1 The Human Hand and its Joints

The human hand is made up of 27 bones connected by several types of joints, which are essential to its nimbleness. The joints of the human hand and an illustration of a joint under healthy and inflammatory conditions are shown in Figure 1. The most relevant joints for this thesis are the metacarpophalangeal joints (MCP) between the fingers and the palm, and the proximal and distal interphalangeal joints (PIP, DIP) between the finger segments ("phalanges"). As well as the joints of the carpal bones of the wrist (intercarpal joints), they belong to the synovial joints. Synovial joints sit in the synovial cavity and are defined by the synovial capsule, which is filled with the synovial fluid. This is the region that is affected by most inflammatory diseases.

1.2.2 Rheumatoid Arthritis

Rheumatoid Arthritis (RA) is one of the most common autoimmune diseases, mainly affecting the joints such as the synovial joints in the hand. It is estimated to have a 0.5 - 1.0% worldwide prevalence [6] [7] [8]. Arthritis patients often suffer from chronic pain, limited mobility, and severe joint deformations. This is

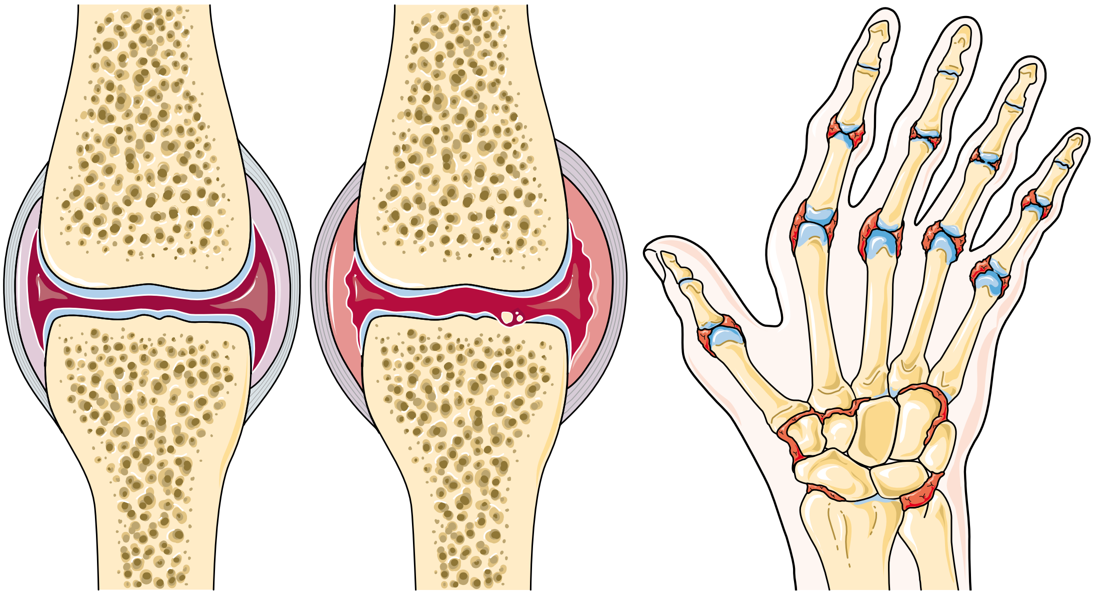


Figure 1: Illustration of the joints in the human hand

(left) Schematic structure of a healthy synovial joint. (middle) An inflammatory joint. (right) Schematic illustration of the bones and joints in the right human hand with Rheumatoid Arthritis. (Source: Servier Medical Art)

usually due to immune system-specific inflammatory reactions targeted at joint tissue, such as cartilage [9] [10]. Genome-wide studies have identified several genetic factors that favor RA, including some alleles of the well-known HLA-DRB1 gene [11].

Current diagnosis approaches include molecular biological markers (e.g., rheumatoid factors like immunoglobulins) and image-based techniques, such as X-ray, CT, and MRI [12].

Current treatment strategies focus on relief of symptoms, slowing down the progression of the disease, or surgically removing parts of the joints in a synovectomy. Thus, the effectiveness of the treatment is highly dependent on early diagnosis and good monitoring of the progression [6] [13].

1.2.3 Psoriatic Arthritis

Besides the common Rheumatoid Arthritis, there is also the Psoriatic Arthritis (PsA). In contrast to a regular Psoriasis Vulgaris (PsV), it usually involves affected joints as seen in RA, but characterizes itself with an abundance of Rheumatoid factor [14]. PsA shows similar symptoms as RA with additional dermatological manifestations. It affects the patient's quality of life similarly as RA [15]. The estimated prevalence of PsA varies from 0.05 – 0.25% to 0.3 - 1.0% of the population [16] [17].

Many genetic risk factors are known to favor PsA. Most of these factors are mutations or variants of genes, which all play critical roles in the immune system, including HLA-B, IL12B, TRAF3IP2, TNIP1, and TYK2 [18].

A differential diagnosis of PsA is difficult. As in RA, a combination of physical examinations of the joints, radiological imaging, and blood markers are necessary [19]. In addition, careful investigation of the nails and the skin is required to identify known symptoms of Psoriasis.

As in RA, there is no cure for PsA, and treatments only slow down the progression. According to current guidelines inhibitors against tumor necrosis factor (TNF), like etanercept, infliximab, and adalimumab, are recommended for severe PsA [20]. Recently, antibodies against interleukin (IL) 12/23 and 17A have been approved and show good efficacy [21] [22].

1.3 Fluorescence Optical Imaging

Recently, Indocyanine green-enhanced fluorescence optical imaging (FOI) has been introduced and proven to be competitive to MRI [23] [24] [25]. Indocyanine green (ICG) is a fluorescent dye that binds to plasma proteins such as globulins. In the diagnostics of inflammatory diseases, ICG is applied intravenously while the examined body parts (e.g., hands) are being imaged continuously. The image acquisition starts prior to the injection and ends after a few minutes when the dye has been completely removed from the circulation by the liver. This enables to visualize the vascular circulation and microcirculation and detect alterations or anomalies. Because of its property to bind globulins, inflammatory regions will show up with high signal intensities in the image. Recent studies suggest that FOI can reveal subclinical inflammatory activity in asymptomatic patients with early Rheumatoid Arthritis and negative MRI [24]. Figure 2 illustrates the FOI acquisition procedure.

Currently, the main application for FOI is the disease assessment and clinical research of inflammatory diseases, such as RA and PsA. Furthermore, FOI is actively used in pre-clinical cancer research for non-invasive imaging of cancer-related drug delivery and lymph node growth monitoring [26] [27] [28] [29] [30].

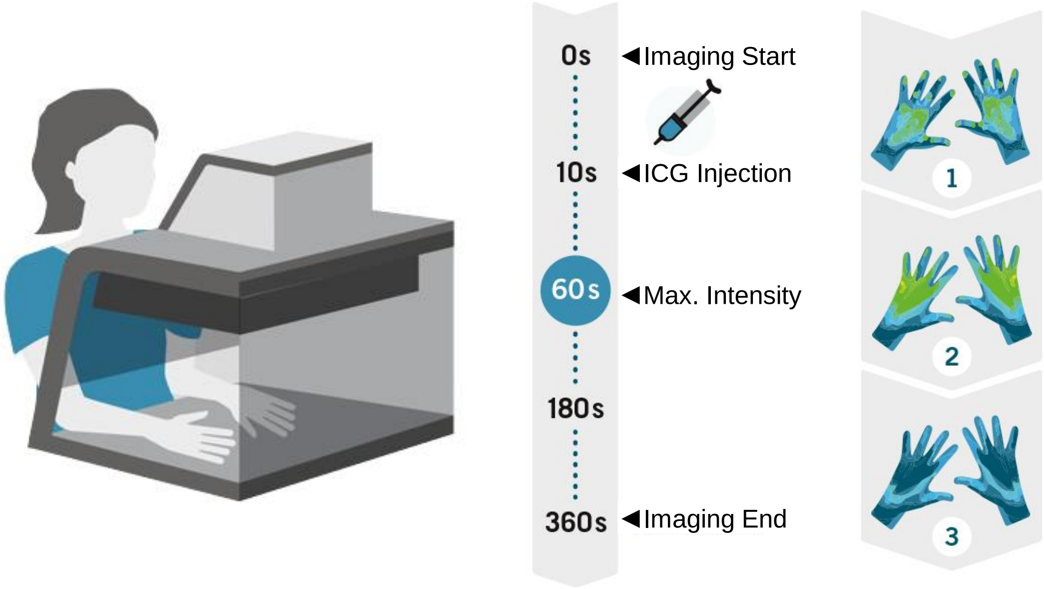


Figure 2: Illustration of the FOI procedure

(left) The hands of the patient are placed in the imaging chamber of the FOI device. The light source and the image sensor are situated in the top part of the device. (right) The three phases of signal distribution. (Source: nanoPET)

1.3.1 Image Analysis of FOI Data

The analysis of the images is currently performed by experienced FOI observers, who manually mark joints with higher intensities. Approaches from the analysis of MRI data cannot easily be transferred. Most approaches for assessing inflammatory joint diseases based on MRI are focusing on measurements of the joint space width in finger and wrist joints [31] [32] [33], which cannot be assessed with FOI. The lack of automatic analyses for MRI data is tried to cope with precise guidelines on manual measurements for RA and PsA [34] [35]. Since analyzing all frames individually is impractical, the frames are summed up within three phases, which are then analyzed separately. Phase 1 (early enhancement phase) begins with the image acquisition and ends with the beginning of the fingertip illumination. The following phase 2 (intermediate phase) ends once the signal intensities in the fingertips begin to decrease. All frames after the second phase until the end of the image acquisition are defined as phase 3 (late phase).

The standard procedure for FOI data analysis is, according to recent publications [23] [24], very subjective. At first, the signal intensities of the image need to be scaled in relation to the fingertip signal. All joints are then examined and checked for activity. A joint is counted as active if it shows higher signal intensities relative to its surrounding tissue or its counterpart on the other hand. An activity score is assigned to each examined joint between 0 and 3 depending on the size of the affected area of the joint. A score of 0 is assigned if no activity is present. The other three scores are given for area sizes of up to 25%, up to 50%, and above 50%, respectively. The summation of the scores of all 30 joints defines the fluorescence optical imaging activity score (FOIAS).

This procedure allows for bias at the phase definition, image scaling, and activity determination of each joint. Although the analysis is usually performed by an experienced analyst and without additional knowledge about the patient ("blinded"), the underlying image data allows for further indications besides inflammatory activity, that may bias the analyst. Especially in RA, hands can be severely deformed, which is then clearly visible in the FOI data. This could influence the rater subconsciously and favor a "diseased" classification. Figure 3 shows two example cases of PsV and PsA. The PsA patient shows clear inflammations of the PIP joints in both index and middle fingers.

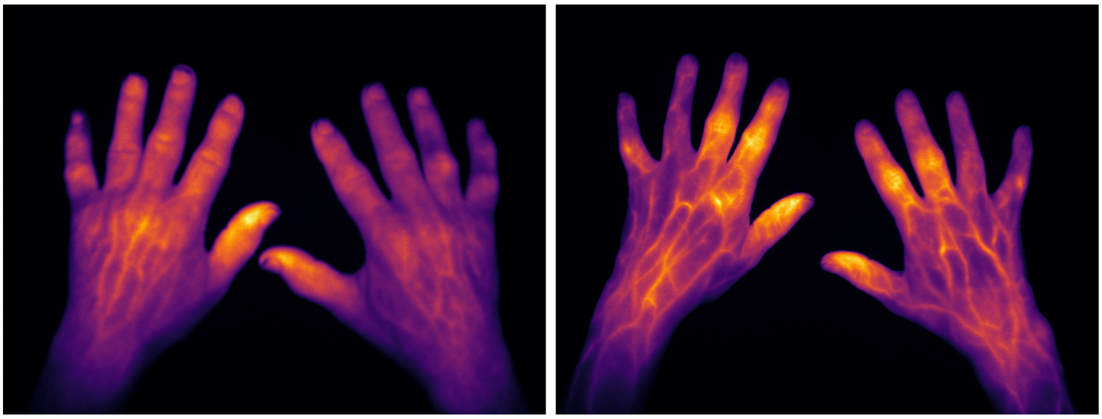


Figure 3: Example FOI images from Psoriasis patients
(left) A patient with Psoriasis Vulgaris (PsV). Only the thumbs show increased signal intensities. (right) A patient with Psoriatic Arthritis (PsA). The proximal interphalangeal joints of index and middle finger show higher signal intensities on both hands. This indicates increased inflammatory activity in these joints.

1.4 Clinical Data

The data in this study was provided by nanoPET Pharma GmbH within the BMBF-funded project "Fluoromath". Experienced physicians acquired the images in a multicenter trial on Xiralite devices. The imaging started ten seconds prior to ICG injection and was performed every second for 6 minutes, resulting in 360 frames per patient.

Figure 4 gives an overview of all diseases that are present in this study. The data set contains 2383 patients, including 185 with PsV and 383 with PsA. 75.3%

of the patients are female. 475 patients are diagnosed with Rheumatoid Arthritis (RA). Besides the assigned diseases, other clinical parameters have been recorded for the patients. The median age of all patients at examination is 58 years (25% – 75% IQR: 50 – 68 years). The median body weight is 75kg (IQR: 65 – 88kg) over all patients, and 87 kg (IQR: 78 – 96 kg) and 71 kg (IQR: 62 – 83 kg) in male and female patients, respectively. Additionally, the Disease Activity Score (DAS28) and the Tender Joint Count (TJC28) are available for all patients.

The DAS28 is a validated scoring system for the assessment of RA, established by the European League Against Rheumatism [36]. It is based on the number of tender joints, number of swollen joints, Erythrocyte Sedimentation Rate (ESR), and the subjective disease assessment (SDA) of the patient. The DAS28 ranges between 0 - 10 and can be discretized into three categories. Values until 3.2 are considered “inactive”, above 3.2 and until 5.1 “moderate active”, and values above 5.1 “very active”. It is calculated using the formula:

$$\begin{aligned} DAS28 = & 0.56 * \sqrt{\#tender\ joints} + \\ & 0.28 * \sqrt{\#swollen\ joints} + \\ & 0.7 * \ln ESR + 0.014 * SDA \end{aligned}$$

The TJC28 is the number of joints with tenderness upon touching. The score includes in total 28 joints. These are the MCP and PIP joints of both hands, as well as the wrists, elbows, shoulders, and knees. There is no defined procedure available that corrects these joint-based scores for patients that have undergone an amputation of a relevant joint or limb. Thus, the only present case in this study has been excluded from these analyses.

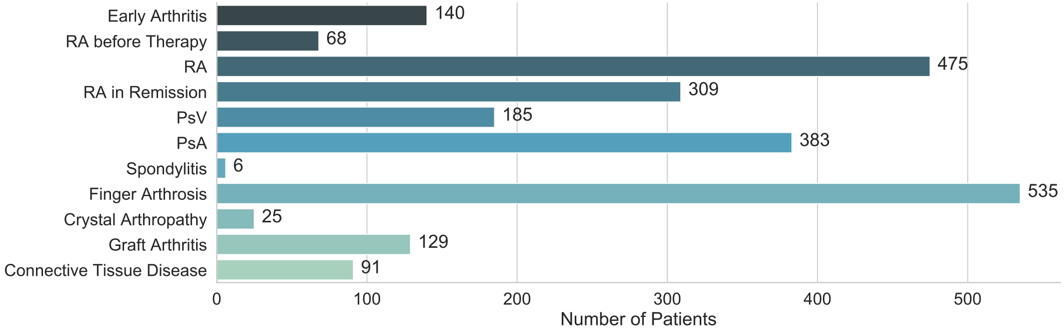


Figure 4: Cohort Statistics

Distribution of assigned diagnoses, that have been assigned to each patient prior to enrollment.

1.5 Problem, Motivation & Approach

Inflammatory diseases are highly prevalent. But many of them are not fully understood yet. The individual diagnosis and assessment of severity currently require the judgment of experienced physicians. For many years, magnetic resonance imaging (MRI) has been the de-facto standard for radiation-free, image-based disease assessment of inflammatory diseases like rheumatic Arthritis. Recently, fluorescence optical imaging (FOI) has been established as a new method. In contrast to MRI, FOI acquires images at a higher temporal resolution without compromising the spatial resolution, which enables the identification of disturbances in microcirculation. However, the quantitative assessment still involves manual interaction of an experienced physician or technician, and the temporal resolution is not used to the full extent [23] [37] [38]. This introduces a bias and is hardly comparable or reproducible. But an unbiased, comparable, and reproducible quantitative disease assessment is crucial in order to fully understand diseases, make reliable diagnoses, and monitor the progression.

Based on time-resolved FOI data, this thesis presents a novel approach for analyzing inflammatory disorders. The presented data-driven approach aims to highlight differences between diseases in general, assess the severity of individual patients, and monitor their progression over the course of a therapy. This is achieved by automatically extracting features from the images, summarizing them in a low-dimensional representation, and thus enabling a variety of mathematical methods for further analyses such as classification and approximation of the underlying dynamics.

This chapter has described the general motivation by highlighting the medical importance and current approaches to analyzing inflammatory diseases in time-resolved fluorescence imaging. The second chapter elaborates on the problems of Machine Learning approaches for medical image analysis and introduces a new data-driven approach and its prerequisites. The approach is based on a Variational Autoencoder and generates a low-dimensional representation of the original data. This low-dimensional representation aims to enable and facilitate further analyses. The third chapter extends the approach for classification and describes how characteristic problems of labeled data can be dealt with. In a two-step approach, the classifier is built using the low-dimensional representations for subgroup classification. The fourth chapter extends the approach with an analysis of the time component of medical image data and how to discover their underlying dynamics. This is realized by employing two copies of the model and joining them with a linear operator in the low-dimensional space. This procedure is derived from the Koopman Operator Theory. The fifth and final chapter summarizes the thesis and discusses the findings.

2

A new unsupervised Approach to enable Disease Analysis

2.1 Problem Formulation

The latest approaches in the field of Image Analysis are using Machine Learning, more specifically Deep Neural Networks, to solve problems of segmentation, classification, or recognition. Some approaches are able to outperform humans in specific tasks [39] [40] [41]. However, they perform only well under conditions, which are rarely fulfilled in medical imaging problems. A common requirement for the success of data-driven methods, like Convolutional Neural Networks (CNNs), is the availability of large, well-annotated data sets. While there is no exact formula to give a lower bound on the necessary amount of data, this theoretical lower bound will increase with the dimensionality and complexity of the problem, the number of classes and the inter-class similarity [42] [43]. So-called Transfer Learning approaches can achieve competitive results under limited data-availability [44]. However, these approaches still require unambiguous and well-annotated data.

This thesis presents a solution for scenarios in which classical supervised approaches with CNNs usually fail. These scenarios are also typical for medical image analysis or share at least most of the characteristics. The first characteristic of these scenarios is that the given data for the problem is usually high-dimensional. In general, this is true for many image-based problems, since each pixel in the image represents a mathematical dimension. But especially in medical image analysis, the resolution matters [45] [46]. While many general image analysis tasks, like classification, can be performed on low-resolution images or down-sampled images, dealing with high resolutions and the resulting ability to detect very fine structures or subliminal anomalies is the unique selling point for Machine Learning in medical applications. Another problem is the lack of sufficiently sized data sets [47]. This results directly from the costs of the image acquisition itself. Medical devices, including imaging devices like Computed Tomography (CT), Magnetic resonance imaging (MRI), are expensive to procure and maintain. Medically trained personal is usually necessary to perform the image acquisition, and experienced physicians are crucial for the diagnosis. This is one of the main reasons, besides ethical questions, why large medical image data sets are rarely publicly available. However, a third characteristic of medical image analysis is often the limited domain. For example, the task of detecting tumors in X-ray images of lungs usually starts with hundreds of X-ray images of thoraxes in an upright orientation. This is reducing the actual degrees of freedom, as the individual pixels are strongly correlated in certain areas of the image. Furthermore, working in a limited domain allows to include domain knowledge into the model. Domain knowledge can be included in the design of the architecture of a Neural Network, or solely in the form of constraints [48] [49].

Further problems are not limited to medical imaging. They have been described for many medical and biological data analyses. The success of supervised Machine Learning approaches not only depends on a suitable input data set. The availability and quality of the respective target values, e.g., labels or classes, are equally important. And although the target values are usually lesser complex structures, they are as valuable as the input data. While the image acquisition itself can be performed by specially trained personnel or even the patient itself, the target value or label usually requires a differential diagnosis by an experienced physician or several physicians. As this circumstance directly correlates with costs, it results in reduced availability, since the data is either not acquired or not published. Moreover, several factors can impair the quality of the target values. For example, in some cases, there is no clear boundary between healthy and diseased patients. An ill-posed problem formulation can also lead to ambiguous results. The third chapter of this thesis is dealing with classification problems and presents the difficulties with labeled data in more depth.

The presented approach in this chapter aims to reduce the dimensionality of the data by learning a low-dimensional space, in which each dimension will aim to represent meaningful features. This unsupervised approach is independent of good labels for each data point, but makes use of those, if they are available for subgroups. The reduced space enables the exploration of the data set and its subgroups. Thus, the proposed approach aims to solve the problems of the unknown phenotypes in FOI and low data-availability. The application for classification of subgroups and the analysis of dynamics are discussed in the subsequent chapters.

2.1.1 Related Work

This thesis focuses on automatic, data-driven apporaches. Thus, methods that belong to the field of typical image analyses are not considered here. Unsupervised Learning can be applied in many fashions on medical imaging data, even beyond traditional clustering methods. The application of Stacked Autoencoders and Deep Belief Networks has been reviewed for feature extraction and image registration in fMRI images [50]. Conditional Variational Autoencoders enable to learn representations from lungs and brains from a healthy population [51]. Imaging data from a diseased patient is then detected as outliers. However, unsupervised learning does not necessarily be implemented with neural networks. Twellmann et al. use a fuzzy clustering by employing vector quantization, a grouping of similar image voxels, to detect lesions in breast MRI [52]. Further approaches are referenced to in the following paragraphs, accompanied by the introduction into their respective methods.

2.2 Machine Learning & Deep Learning

2.2.1 Deep Neural Networks

When analyzing images or data in general, the problem can become too large or complex such that analytical or manual approaches are not feasible. One of the most common reasons is that the wanted effect is unknown or too complex to be formally described. In these scenarios, data-driven methods from the field of Machine Learning become interesting. While many Machine Learning methods have been published within the last decades, the latest developments in data-availability and computational power have led to an increase in performance and popularity.

Deep learning and in particular Convolutional Neural Networks (CNNs) are showing increasing performance and thus popularity for the analysis of extensive, high-dimensional data. CNNs can detect 1000 and more different classes of objects in images with error rates below 5% and are outperforming humans in terms of accuracy and of course, speed [40] [53] [54]. While this is true for general images and photos, the frequency of CNNs in medical imaging is still low, where established image features, such as scale-invariant feature transform (SIFT) and histogram of oriented gradients (HOG), and other manually defined features found the basis of image analyses [55]. One reason for that is access to large and well-annotated image databases. In contrast to big internet companies, which have access to a tremendous amount of photos and images, clinical researchers are limited by privacy policies, ethical restrictions, access to patients, and the costs of image acquisition [56]. Another reason is the stigma of being a non-transparent black box, which is still attached to data-driven methods, in particular, deep learning and CNNs. This inhibits the spreading of these methods in the clinical application, although many approaches have been proposed to shine a light into these black boxes.

The general form of neural networks has not changed drastically since the first proposal of the perceptron in 1958 by Rosenblatt [57]. The idea is inspired by biological neurons. As pictured in Figure 5, the neuron computes a weighted sum of the input signals, comparable to the dendrites in the biological model. Only if this sum exceeds a certain threshold, the neuron outputs a signal, otherwise not. This behavior is adapted from the axon of the biological neuron, which fires a specific signal only if a certain electrophysiological potential is exceeded. In the mathematical model of the neuron, a layer consists of a linear system followed by an activation function (here denoted as g), which is typically non-linear:

$$f\left(x\right) = g\left(Wx + b\right).$$

The shape of the weight matrix W is determined by the dimensionality of the input space and the desired output space. In theory, any function $\mathbb{R}^n \rightarrow \mathbb{R}^m$ can be approximated by scaling the output dimensionality large enough or by chaining enough layers to a network. This universal approximation theorem has first been proven for single layer networks [58], and then for multi-layered networks [59]. Later Lu et al. [60] and Hanin [61] have proven this theorem with the widely used rectified linear unit (ReLU) activation function.

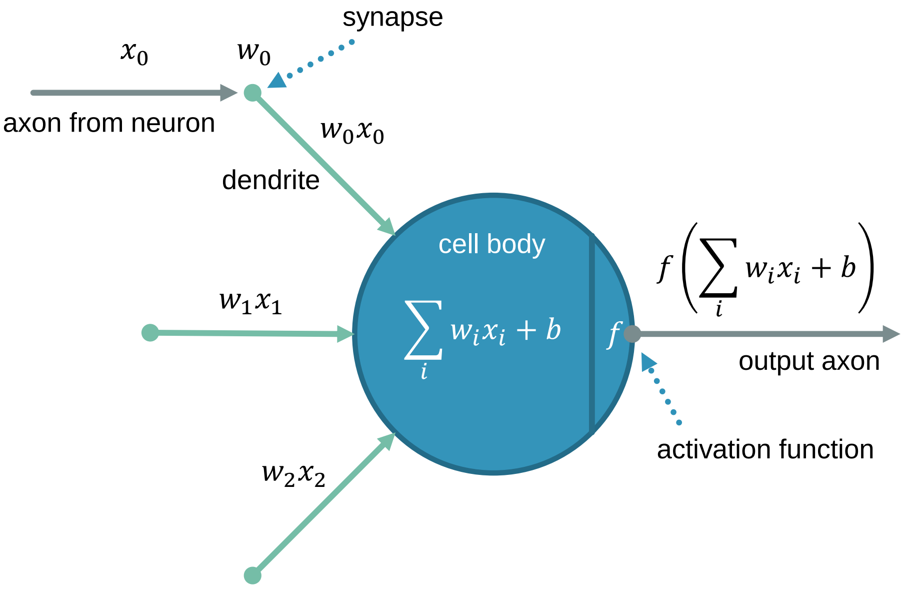


Figure 5: Illustration of the Neural Network Principle

The dendrites of a biological neuron receive the input information of the previous neural layer. Their signals are weighted and summed up in the cell body. Only if the sum exceeds a certain threshold, the neuron outputs a signal. (Source: CS231n Stanford)

2.2.2 Convolutional Neural Networks

When working with image data, classical Neural Networks with fully connected layers become infeasible, since the number of parameters will explode with the number of pixels. Therefore, Convolutional Neural Networks (CNNs) are used for images. CNNs keep the number of parameters low by learning filters, which convolve over the image and extract features (Figure 6). A series of convolutional and pooling layers is often followed by one or more fully connected layers. This two-part layout of feature extractor and classifier is still frequently used in CNNs for classification and regression tasks.

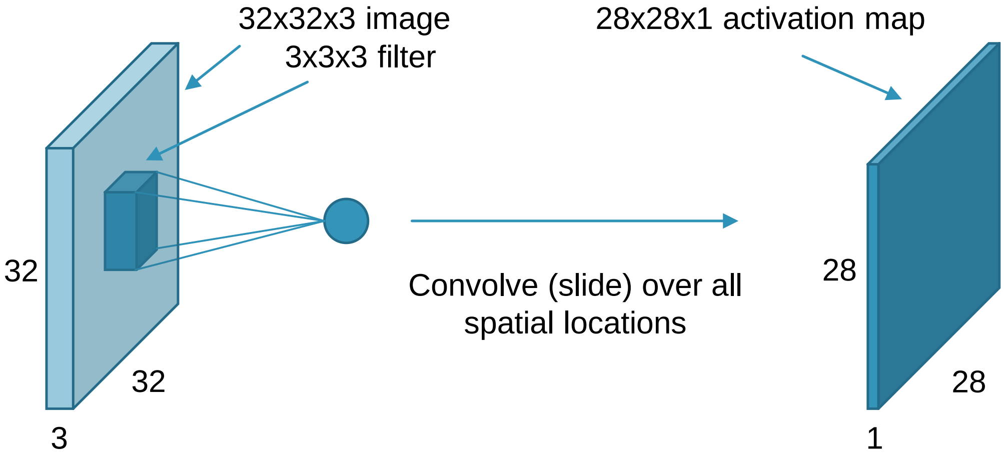


Figure 6: Principle of a Convolutional Layer

Here a 5x5 filter convolves over the image. The product of the filter and each 5x5 patch of the input image results in a smaller activation map. (Source: CS231n Stanford)

Figure 7 (top) shows the first CNN with high accuracy and feasible computational effort. It has been proposed by LeCun et al. [62] [63], and achieves a test error rate well below 1% in the classification task of handwritten digits. In 2012, Krizhevsky et al. [64] succeeded in the "ImageNet Large Scale Visual Recognition Challenge" (ILSVRC-2010), and achieved top-1 and top-5 error rates of 39.7% and 18.9% in classifying 1000 different real-world objects. The inception blocks developed by Lin et al. [65] in 2013 introduced parallel convolutional layers with different filter sizes. This concept has been applied successfully [40] and extended several times in order to achieve achieved top-1 and top-5 error rates of 16.4% and 3.1%. The bottom of Figure 7 illustrates the architecture of the first successful Inception network by Google.

The development of CNNs has been characterized by an increasing number of convolutional layers with the intention to extract more complex information from the images. An illustrative example are the filters from a CNN, which has been trained on a data set containing faces [66]. While the first layers have filters for features like edges, deeper layers have filters, which are activated on high-level features like eyes and noses. The last layers show already face-like structures. Comparable to Eigenfaces, a linear combination of these features should be able to reconstruct any face from the data set [67] [68]. Unfortunately, deeper models are harder to train, while not decreasing the error rates. The introduction of residual blocks helped training models with 152 layers while maintaining comparable error rates [69].

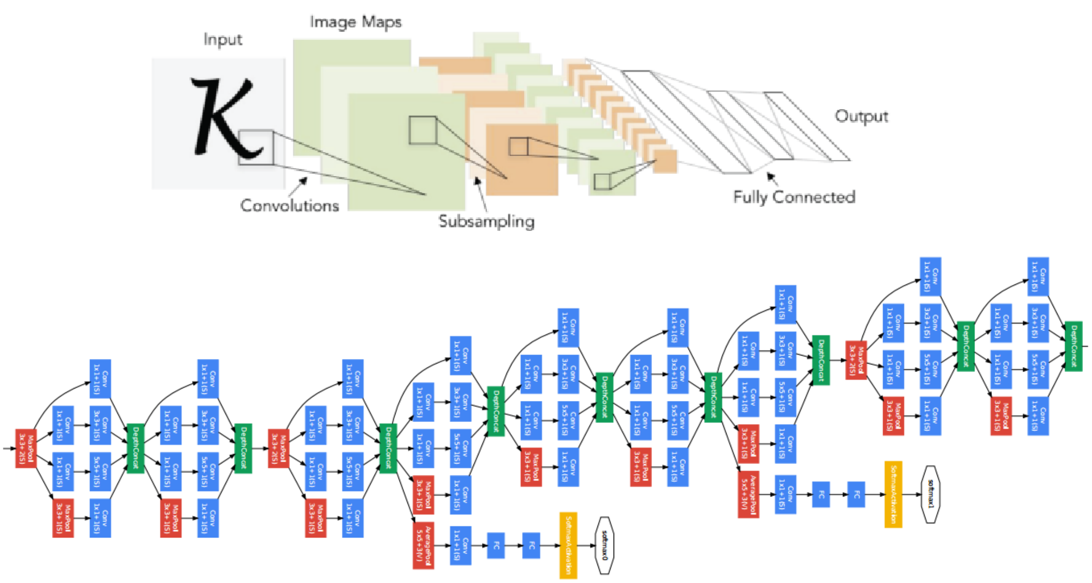


Figure 7: Architecture of the LeNet and the Inception Network
(top) Architecture of the AlexNet. (bottom) The first Inception architecture GoogLeNet (Source: LeCun, Krizhevsky)

2.2.3 Optimizing a Neural Network

2.2.3.1 Optimizers

While the design of a neural network is non-trivial but crucial for a good performance, finding the optimal parameters or "weights" for a given task is the main task. For supervised tasks, the optimal parameters θ for the mapping $f_\theta(x)$ from the input space $x \in X$ to the according output space $y \in Y$, which are usually labels, is desired with low error:

$$\underset{\theta}{argmin} \left\| y - f_\theta \left(x \right) \right\|_2^2$$

In practice, the number of data points and parameters is too large to compute the optimal parameter at once. Thus, the error is minimized iteratively in small steps, on small subsets of the data. This method is called Stochastic Gradient Descent SGD) [70]. Since its introduction in 1951, SGD has been extended many times to improve convergence, but the principle remains the same. The parameters θ at iteration t are updated for every batch of size b of the training data in opposite direction of the gradient ∇_θ of the error function $\mathbb{E}_\theta$, with a restricted step size or learning rate η:

$$\theta_{t+1} = \theta_t - \eta * \nabla_{\theta_t} \mathbb{E}_{\theta_t} (x^{(i:i+b)}; y^{(i:i+b)})$$

A notable, popular extension is the Adaptive Moment Estimation (Adam) [71] optimizer. As most extensions to SGD, Adam computes the first and second moment of the gradient. Simplified, the optimizer memorizes past gradients and tries to maintain the average direction, making it less sensitive to local minima. This is realized by computing the decaying averages of past m_t and past squared gradients v_t:

$$m_t = \beta_1 m_{t-1} + \left(1 - \beta_1 \right) g_t$$

$$v_t = \beta_2 v_{t-1} + \left(1 - \beta_2 \right) g_t^2,$$

where g_t is the gradient at iteration t. The parameters β_1 and β_2 control the exponential decay rate and are usually close to 1. Since the initialization of the moments can introduce bias, the moments are corrected by $\widehat{m}_t = m_t/(1-\beta_1^t)$ and $\widehat{v}_t = v_t/(1-\beta_2^t)$. The parameters are updated using the formula:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\widehat{v}_t} + \epsilon}\widehat{m}_t.$$

It is noteworthy that although neural networks are technically just optimization problems, and the mathematical field of optimization offers a powerful toolbox of fast and efficient global optimizers, in practice, only SGD and its variants are being used for most problems. Since SGD is a local optimization method, it only converges reliably, if the objective function is convex or pseudo-convex. Hence, most efforts in optimizing convergence and performance are methods to avoid local minima and flat regions, and to ensure convergence to the global minimum. Modern, global optimizers are currently not feasible for large networks and large data sets. Only Particle Swarm Optimization techniques have been applied recently [72].

2.2.3.2 Activation Functions

As mentioned, the power of deep neural networks come from chaining layers, which are linear functions, and from non-linear activation functions after the layers. The motivation behind activation functions is again the biological model. In a neuron, all input signals are weighted and summed up. If this weighted sum exceeds a certain threshold, the neuron fires. While early neural networks mimicked exactly this behavior, recent activation functions have been developed with better numerical characteristics. There are many different variants of activation function in use, where Sigmoid, tanh, and ReLU, are one of the most frequently used ones. Furthermore, classification networks usually have a Softmax activation in their last layer. This ensures that the model emits class probabilities, which can be optimized using a categorical Cross-Entropy.

Both, Sigmoid and tanh are inspired by the binary behavior of biological neurons. In contrast to the Sigmoid function, the tanh is zero-centered. However, both suffer from slow convergence due to the vanishing gradient problem. ReLUs

Name	Function
Sigmoid	$f\left(x\right) = \frac{1}{1+e^{-x}}$
Tanh	$f\left(x\right) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$
ReLU	$f\left(x\right) = \max\left(0, x\right)$
Softmax	$f\left(x\right)_k = \frac{e^{x_k}}{\sum_k e^{x_k}}$

Table 1: Equations of Activation Functions

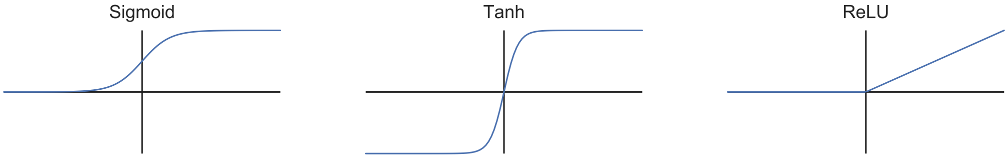


Figure 8: Activation Functions

(left) The Sigmoid function returns either 0, 1, or values in between near $x = 0$. (middle) The Tanh function acts similarly but its values range from -1 to 1. (right) The ReLU function returns the input value, if it is positive, else 0. See Table 1 for the corresponding equations.

were first proposed in 2000 [73] for digital circuits, and later proven to converge faster [74]. In general, rectifying activation functions, such as swish, seem to achieve better results [75], if the vanishing gradients problem is controlled. For classification tasks, the Softmax function is widely used in combination with a cross entropy loss, as it normalizes the output to a probability distribution.

2.3 Variational Autoencoder

2.3.1 Autoencoder

In Unsupervised Learning, autoencoders are one of the most popular methods. In general, autoencoders aim to learn a low-dimensional representation of a given input and subsequently reconstruct the input as accurate as possible. An autoencoder essentially consists of two parts: the encoder ϕ, which transforms the input data into a low-dimensional representation, and the decoder ψ, which transforms the representation back into the original space. Figure 9 illustrates the architecture of an autoencoder with convolutional layers as encoding and decoding transformations. This so-called “informational bottleneck” forces the autoencoder to identify the important information in the data set and to learn a domain-specific representation. While the encoder and decoder can be practically any transformation, they are often implemented as Neural Networks. In general, for given data x the following loss function is minimized:

$$\underset{\theta}{argmin} \, \|x - \psi_\theta(\phi_\theta\,(x))\|_2^2$$

One obvious use case is image compression [76]. If both the sender and the recipient are in possession of the autoencoder, they only need to exchange the low-dimensional representation. In this way, both need to exchange less information. Another popular application is image denoising. Here, the input is presented in a distorted form, e.g., with additional noise or missing image patches. The autoencoder then aims to reconstruct the original, undistorted image [77].

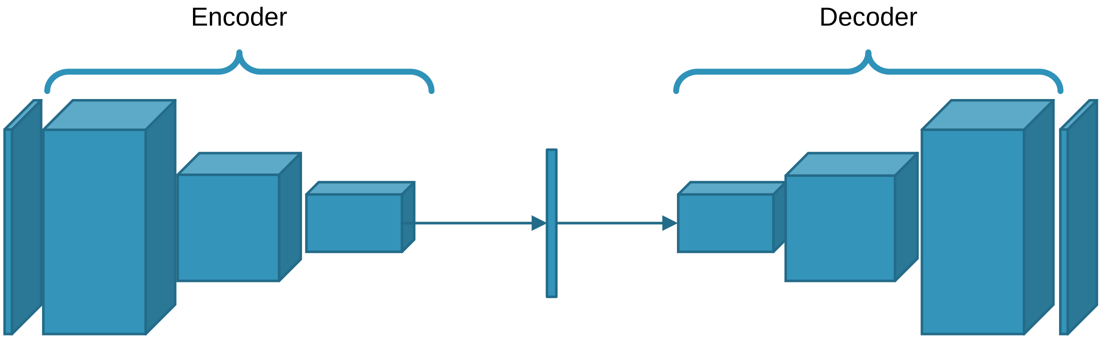


Figure 9: Illustration of a Convolutional Autoencoder

The input image is fed in from the left. After a couple of convolutions, more and more features are extracted, while the spatial resolution is decreased. The middle part represents the low-dimensional representation and thus the informational bottleneck. From there a series of upsampling convolutions is performed to generate the output image.

2.3.2 Variational Autoencoder

Autoencoder can easily overfit and simply “memorize” the data set, without generalizing. As a result, minor deviations of the low-dimensional representation can lead to drastic changes in the decoded information. In the opposite direction, small differences in the input image can result in entirely different representations. In 2013, Kingma and Welling introduced the Variational Autoencoder VAE) [78]. The VAE aims to learn a meaningful, robust latent space, in which the low-dimensional representations live. This is mainly done by learning a multivariate normal distribution $\mathcal{N}(\mu, \sigma)$ instead of fixed representations. The representation is then sampled from this distribution. This forces the encoder and decoder to learn a latent space, in which similar input data points have a small distance. Figure 10 illustrates the architecture of a VAE with convolutional layers as transformations for the encoder and decoder. The central part represents the sampling layer.

This sampling layer introduces a non-differentiable function into the Neural Network. Thus, it is impossible to compute a gradient and perform backpropagation. To circumvent this problem, the so-called “reparameterization trick” is applied. Instead of sampling the representation (here denoted as z) directly from $\mathcal{N}\left(\mu, \sigma\right)$, a vector is sampled from $\mathcal{N}\left(0, 1\right)$ and then scaled with the outputs from the layers, which represent μ and σ:

$$z \sim \mathcal{N}\left(\mu, \sigma\right) \;\rightarrow\; z = \mu + \sigma * \epsilon, \; \epsilon \sim \mathcal{N}(0, 1)$$

The loss function is extended by a regularization term. The Kullback-Leibler divergence is added to the reconstruction loss. This restricts the latent space and forces the VAE to bundle data points with similar features. In practice, the VAE can be seen as a regular autoencoder, but with Gaussian noise added to the low-dimensional representation. The regularization term, which computes the KL-Divergence between the latent space representations and the standard Gaussian distribution, is as follows:

$$Loss_{KL} = -\frac{1}{2}\sum_{k=1}^{K}(1 + \log\sigma_k - \mu_k^2 - \sigma_k^2)$$

As an alternative to the KL-Divergence, the Rényi Divergence and Chi-Divergence are also subject to current research in variational inference [79] [80].

Once a VAE has been trained to a data set, it can be used for several purposes. In this thesis, two characteristics of the VAE will be used. First, all the important information is coded in just a few dimensions. Since the VAE is aiming to reconstruct the input images as good as possible, it can be assumed that all necessary information is preserved and contained in the latent space representation. The problem of having too many dimensions in the input space and too few examples can be solved this way. The second advantage is the informational quality of the latent space. Perceptually similar images should have a small distance in the latent space. Additionally, the dimensions of the latent space should, in theory, correlate with a visual meaning. Assuming this holds true and these dimensions can be identified, it should be possible to define a weighted metric based on the medical relevance of each dimension. This metric would weigh dimensions, which encode shape-related features like the position and size of the hands, less than dimensions that encode the signal distribution.

Besides the mentioned approaches at the beginning of this chapter, recent applications use the trained VAE to compute the likelihood of new data to fit the model to identify abnormalities [81], or the distance to control samples in the latent space to detect pathologies [51]. These approaches compute how well a data point fits to the learned model. In contrast, the proposed approach in this chapter investigates the differences between subgroups within the latent space and aims to describe the prevalent variety of samples fully.

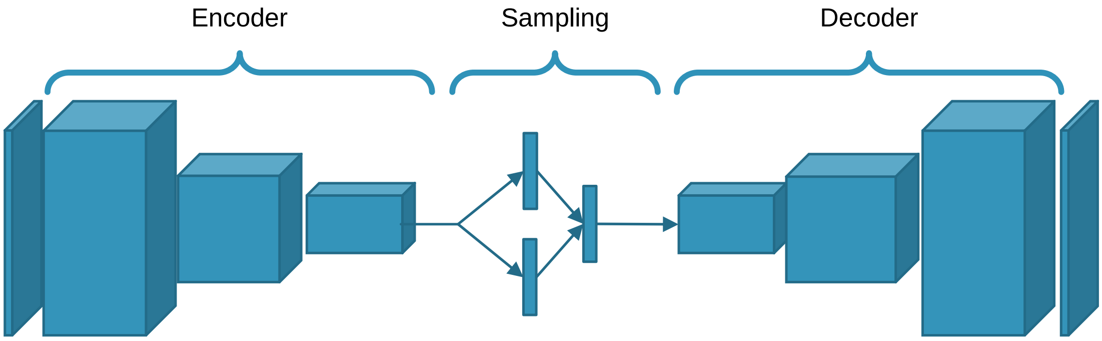


Figure 10: Illustration of a Variational Autoencoder

The input image is fed in from the left. After a couple of convolutions, more and more features are extracted, while the spatial resolution is decreased. The flattened output of the last convolution is used to compute the vectors for μ and σ, from which the low-dimensional representation is sampled. This is then used to generate the output image through a series of upsampling convolutions.

2.4 Methods & Experiments

All experiments are implemented in Python 3.6 using the libraries TensorFlow for the Neural Networks and scikit-image for image handling [82] [84]. scikit-learn is used for the remaining data-processing [83]. The preprocessing of the image data is an important and crucial step prior to the training of the model. Here, the preprocessing consists only of a background segmentation and normalization of the values to a fixed range. Usually, the values of the input data can be normalized to a range between 0 and 1 or standardized to have $\mu = 0$ and $\sigma = 1$. This ensures that the range of gradients is similar across all layers of the network for faster convergence. The background of the image data is set to 0 to remove device-specific background noise and patterns. To distinguish between the foreground, that is the area covered by the two hands, and the background, all pixel values below half of the mean of a time series are defined as background. This threshold has been found empirically and proven to be good enough.

It is good practice to perform the training and compute the performance on two distinct sets of images – the training and the validation set. Here, both sets are sampled from the original data set, and the training set consists of 80% of the original data set, whereas the remaining 20% are used for validation. A third set, the test set, is often used for computing the final performance statistics. It is not applied here since the available data is already very limited. Additionally, all presented model statistics in this thesis are computed from a 5-fold cross-validation.

The model is trained using the Adam optimizer with an initial learning rate of 10^{-4}, exponentially decaying at a rate of 10^{-6}. All models are trained for 50 epochs, that is, the whole training set is presented 50 times in shuffled sequences. The batch size, on which the gradients are computed, is 32. This number is mainly restricted by the computational power of the general-purpose computing graphics processing unit (GPGPU), which was available for this work. Although the batch size should not affect the training outcome drastically other than in terms of speed, it has been shown to have at least a small effect [85]. A small batch size leads to more gradient computations per epoch. A larger batch size enables better distribution estimations. All weights of the model are regularized with small-weighted L_1 and L_2 penalties to favor generalizing weights. To avoid the recognition of device-specific background noise, Gaussian noise is added to the input images with $\sigma = 0.01$.

Hence, the remaining hyperparameters that are needed to be optimized are the latent space dimensionality and the weight of the Kullback-Leibler divergence in the loss term.

2.4.1 Latent Space Size

An obvious question that arises is how large the latent space has to be sized. In general, it should be large enough to encode all the relevant information, but small enough to contain only the necessary information. This chapter will present two approaches to answer the question. The experimental approach iteratively increases the dimensionality of the latent space until the reconstruction quality stops increasing. For comparison, the second approach directly estimates the intrinsic dimensionality of the data set. While there are several approaches to estimate the intrinsic dimensionality of a data set, in this thesis, a Geometry-Aware Maximum Likelihood Estimation (GeoMLE) is used [86]. In contrast to other methods, GeoMLE is claimed to produce robust results also on high-dimensional, non-linear data. It is based on a Maximum Likelihood Estimation MLE) approach [87], which is relying on uniformly distributed data along a linear manifold. Both criteria are usually not met in real-world data sets.

2.4.2 Perceptual Loss Function

Autoencoders aim to reconstruct the input data with a low error. In order to train a neural network based VAE, a differentiable loss function is needed. This is usually the mean squared error (MSE) of the pixel-wise difference between the input image and the reconstructed image. Unfortunately, this leads to slightly blurred or smooth output images [88]. A simplified way to understand this problem is to look at the problem from a probabilistic point of view. The MSE is usually defined as the sum of squared differences divided by the number of samples:

$$MSE = \frac{1}{N} \sum_{i}^{N} \left\| \widehat{x}_i - x_i \right\|_2^2 .$$

The probability function of the Gaussian distribution is defined as follows:

$$p\left(x|\mu,\sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left\|\mu - x\right\|_2^2}{2\sigma^2}\right).$$

Assuming $\mu = \widehat{x}$, $\sigma = 1$, and ignoring the normalization factor $\frac{1}{\sqrt{2\pi\sigma^2}}$, it becomes obvious to see the similarity between the MSE and the log-likelihood of the Gaussian distribution:

$$\log p\,(x|\mu) \propto -\frac{1}{2}\,\|\mu - x\|_2^2\,.$$

Similarity measures like the Structural Similarity Index (SSIM) are not differentiable and can only be used as a metric, but not as an objective function. Ledig et. al [89] proposed a perceptional loss based on Convolutional Layers. Therefore, the input image and reconstruction image are convolved with fixed, pre-trained filters. Depending on the desired reconstruction quality, it is possible to use only the first convolutional layer or many layers. This results in multiple feature maps on which the MSE can be computed. While this workaround does not solve the initial problem, the reconstructed images share more features with their corresponding input images on average. This principle has been applied successfully in several experiments, including the generation of super-resolution images [89] [90] [91] [92] [93].

2.4.3 Visualization Methods

Even though the presented approach aims to reduce the dimensionality of the data significantly, the visualization of data with more than two or three dimensions becomes non-trivial. Principal component analysis (PCA) is an established and well-known method to reduce dimensionality but has many assumptions on the data and other limitations.

t-Distributed Stochastic Neighbor Embedding (t-SNE) [94] is a dimensionality reduction algorithm, that is frequently used for visualization of high-dimensional data. It aims to project the relative distances between data points in a low-dimensional embedding. This way, similar data points remain clustered together, even in 2D. The optimal t-SNE projection is found by solving

$$\min \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}},$$

where p_{ij} is a Gaussian distributed similarity probability in the high-dimensional input space, and q_{ij} a t-distributed similarity probability in the low-dimensional output space.

Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [95] is an extension of t-SNE. It aims to represent the relative distances between clusters of similar data points. This is realized by adding a repellant force into the optimization term of t-SNE:

$$\min \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1 - p_{ij}) \log \frac{1 - p_{ij}}{1 - q_{ij}}.$$

According to McInnes, the first part of the equation “gets the clumps right”, whereas the second part “gets the gaps right”, figuratively speaking.

2.5 Results

This section presents the results of the unsupervised learning approach. Here, a Variational Autoencoder (VAE) is used to learn a low-dimensional latent space of FOI data from hands with various inflammatory diseases. First, the reconstruction properties of the VAE are shown under different training conditions, followed by example cases where the approach fails. Second, the latent space is explored with respect to clinically relevant subgroups. This exploration is done by projecting the representations into 2D using UMAP, and by decoding average latent space representations for subgroups.

2.5.1 VAE generates FOI Images with low error

The VAE model is capable of generating subjectively similar images as the input images. The output images seem to lack details like blood vessels while maintaining essential features like inflammatory joints and the outline of the hands. Figure 11 shows 5 randomly selected example images in the top row and their reconstructions by a trained VAE using the traditional MSE loss and a perceptual loss in the center and bottom row, respectively. The reconstruction quality suffers significantly when the MSE is used as a loss function. The images appear blurred and lack essential details.

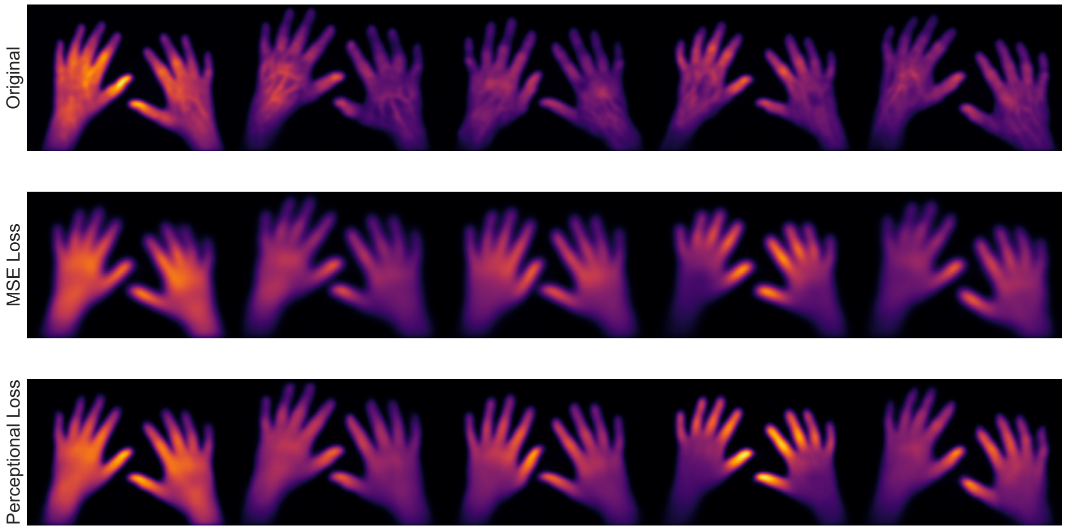


Figure 11: Comparison of Input Images and their Reconstructions by the VAE
(top) Original input images (middle) Reconstructed images with a VAE using MSE loss (bottom) Reconstructed images with a VAE using a perceptual loss

Quantitively, the SSIM between input and output images is on average 0.91 and the pixel-wise Pearson correlation 0.9. Figure 12 shows the quality of reconstruction measured by SSIM and Pearson correlation in dependency of the latent space dimensionality and the applied loss function. Both, SSIM and correlation, are increasing until a latent space dimensionality of 50. A further increment of dimensionality does not lead to significantly better reconstructions. Starting with a dimensionality of 16 the reconstructions become subjectively appealing. Depending on the selected time point of the video data, the estimated intrinsic dimensionality varies between 15 and 49. Figure 12 shows the results from the GeoMLE algorithm for several time points. During the first 50 seconds of the videos, the intrinsic dimension varies between 15 and 25. Subsequently, the intrinsic dimension jumps to 49 and is then slowly decreasing to 20 at the end of the videos. The intrinsic dimension over all patients and frames is estimated at 50. In general, in all cases the employment of the perceptual loss function results in better reconstructions compared to the traditional MSE loss.

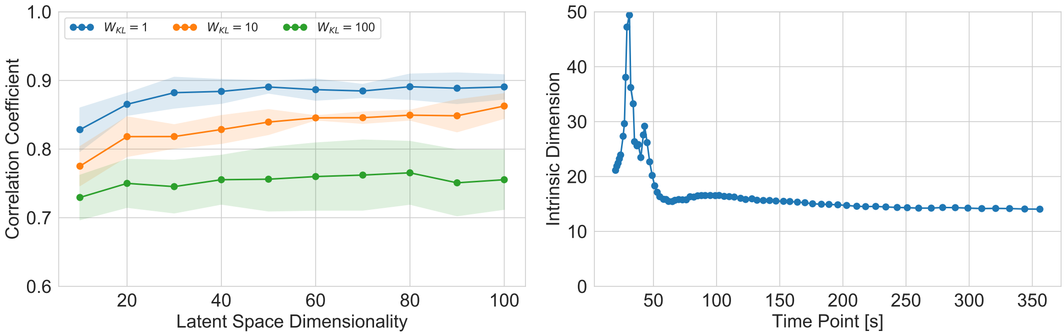


Figure 12: Dimensionality-related Results
(left) Pixel-wise Pearson correlation in dependency of the latent space size and weight of the KL divergence term (right) Estimation of the Intrinsic Dimensionality for each time step of the videos using GeoMLE.

To validate the regularizing effect of the KL-Divergence on the latent space, it was weighted in the optimization term. The weights are chosen from a range between 1 and 100. With the increasing weight of the KL-Divergence term, the reconstruction quality is reduced.

To identify cases, in which the reconstruction fails, the SSIM has been computed for all data points. Figure 13 shows five examples with the lowest SSIM from the training and validation set, respectively. Apparently, cases with severe deformations of the hand cannot be reconstructed accurately in both the training and validation set.

2.5.2 Exploring & Interpreting Latent Space

In order to explore the latent space representations visually, the above-mentioned methods for further reduction of dimensionality can be applied. In Figure 14

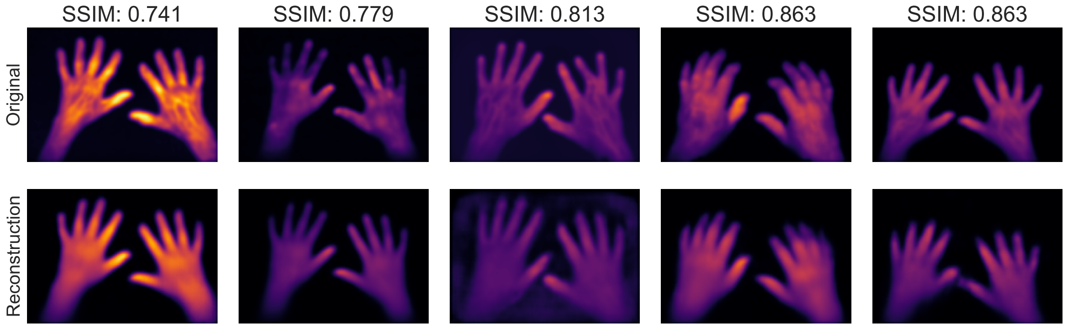


Figure 13: Examples for poor reconstruction results

(top) Original input images (bottom) Reconstructed images with a VAE. The three right-most samples have strongly differing hand shapes, that are underrepresented in the dataset. The reconstructions of the two samples on the left are subjectively good, but lack details of the signal distribution.

the representations for all data points are shown using a UMAP projection into 2D. Different colorizations are applied for the available clinical parameters. The strongest effect on the point distribution has the time. With progressing time, the points are separating from one large lump of points, and form four very distinct clusters. Female and male patients are distributed equally over the clusters. However, within each cluster both groups seem to form sub-clusters. Patients with an inactive DAS28 or TJC28 score are more frequent in the upper right cluster at the third time point. But they are also present in the other three clusters. The colorization with the Psoriasis classes does not show significant differences in distribution. Only the upper left cluster in the third time point contains relatively more PsV than PsA patients, but the overall number is small. RA negative and positive patients do not show any apparent pattern of distribution.

The average latent space representations for subgroups of the patients can be decoded back into the image space. Figure 15 shows the decoded averages for the two Psoriasis groups next to the averages of the same patients computed in the image space. In general, the averages from the image space appear blurry due to the large variation of position, shape, and size of the hands. In contrast,

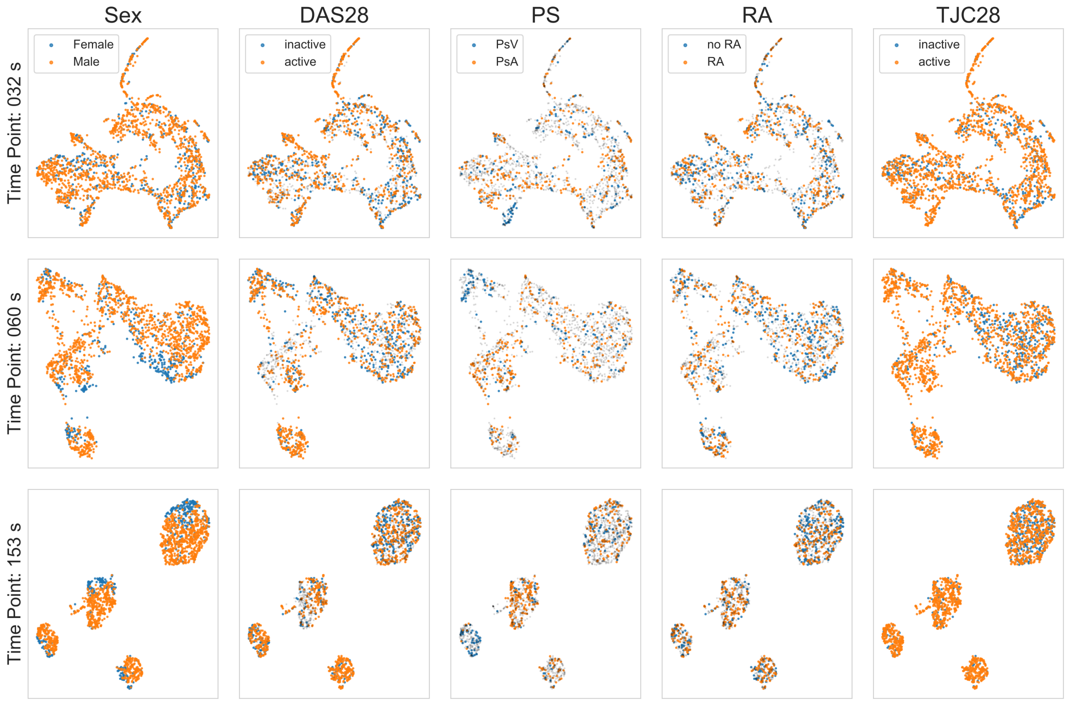


Figure 14: UMAP Visualization of Latent Space

UMAP Visualization of all patient's latent space representations at three different time points (rows). The points are colorized using five clinical parameters (columns). Since not each clinical parameter is applicable or available for each patient, data points with missing information are shown in pale gray as reference.

the decoded latent space averages appear clear, crisp, and subjectively natural. When comparing the average images between the two subgroups, the averages from the image space do not allow for any differential interpretation. However, the decoded latent space averages show higher signal intensities in the finger regions. The mean intensities along the central axis of the fingers are 0.6 and 0.7 in the decoded averages for PsV and PsA, respectively. This difference is also present in the mean intensities across all images of these two groups. The mean intensity computed over all PsV patient's images is 0.61 and 0.73 in PsA.

When comparing two groups, which are expected to not differ pathologically, the signal intensity difference disappears. Figure 16 shows the decoded averages for the female patients compared to the male patients. The signal distributions

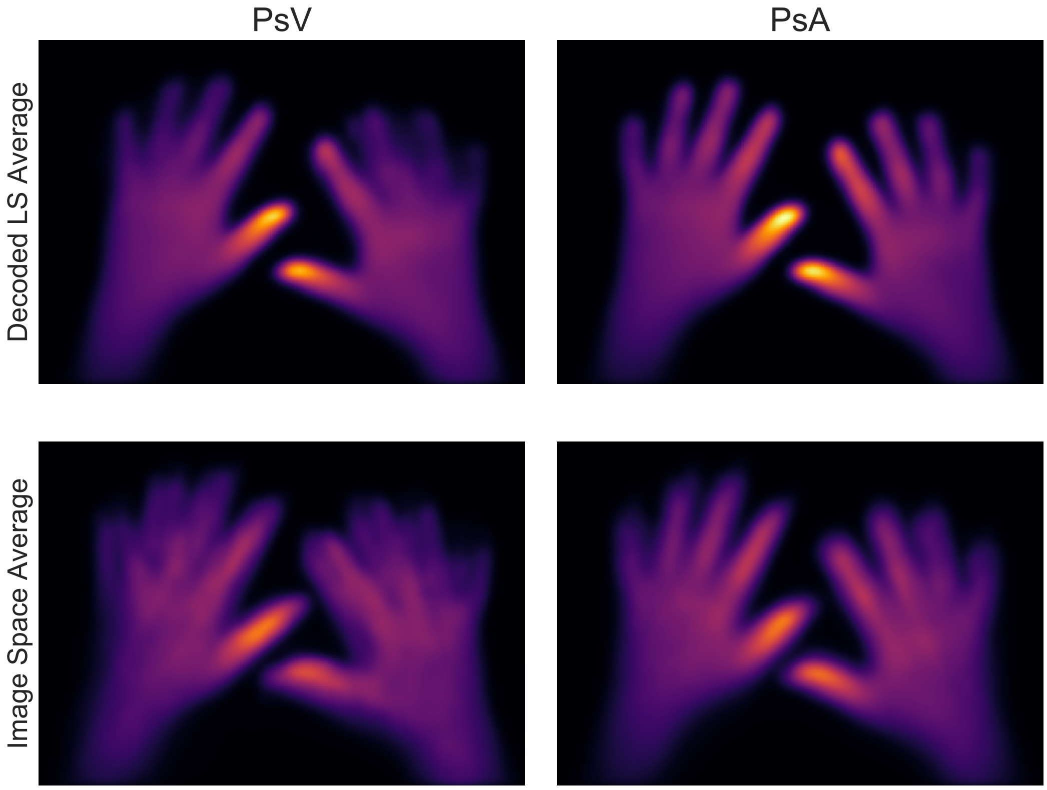


Figure 15: Comparison of generating average Images in Latent Space and Image Space

(top) Decoded images from the average latent space representations of PsV and PsA patients. (bottom) Average images generated from the original images of PsV and PsA patients.

only differ in the thumb areas. However, the size of the average female hand is only 80.2% of the size of the average male hand.

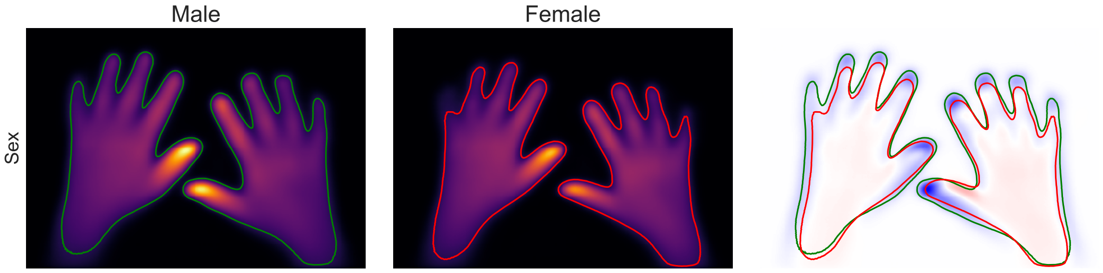


Figure 16: Decoded average Latent Space Representations of male and female patients

(left) Decoded images from the average latent space representations of all male patients (middle) Decoded images from the average latent space representations of all female patients (right) Overlay image of the outlines.

To further understand the latent space, a latent space representation can be varied in a single dimension and subsequently be decoded into an image. Figure 17 shows the effects of varying a single dimension of the latent space representation of a single patient. It has been varied within its observed range from minimum to maximum, and the resulting new representation has been decoded (left to right in the figure). With increasing the value, the signal intensities of the finger regions begin to increase until they appear much brighter.

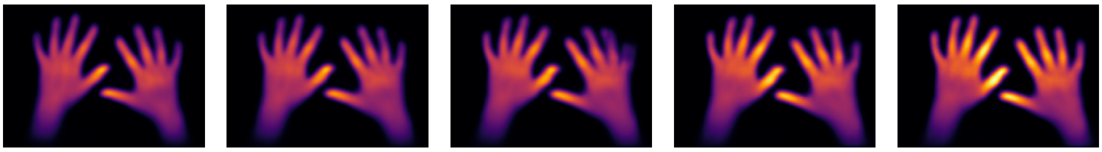


Figure 17: Effect of varying a single dimension

The decoded images based on a single latent space representation, that has been modified in a single dimension from its observed minimum (left) to its observed maximum (right).

The low-dimensional representation can be computed for each frame of a video. Figure 18 shows the average standard deviation across all encoded videos. Approximately half of the dimensions have very low standard deviations across the course of a video. Only five dimensions show high variations. This information can be used to explore the latent space further. Figure 19 shows the t-SNE and UMAP visualizations of all encoded videos with color-coded time. Using all dimensions, one large lump of points appears with individual trajectories of dense points. However, when computing UMAP on the five dimensions with the highest temporal variance, the structure changes drastically. Early time points form a small, dense lump of points on the left side. Later time points seem to spread in space. Interestingly, the latest time points are not on the rightmost side of the plot. Until 180 seconds into the video, the latent space representations spread to the right. Afterward, they are returning in the direction of the origin. However, the same effect is not present in the t-SNE projections. Using only the five dimensions with high temporal variance, there is no clear overall temporal behavior visible as in the UMAP projection.

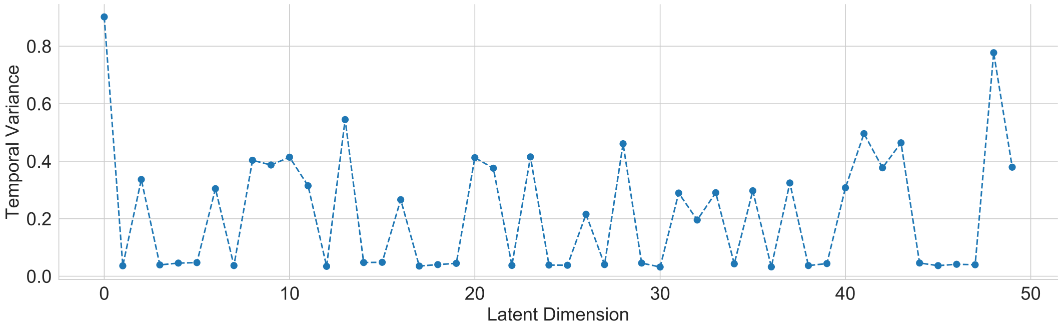


Figure 18: Temporal Variance of the low-dimensional representations

Average variance computed over all latent space trajectories for each individual latent space dimension. There are only a few dimensions that vary significantly over the duration of the videos. The majority remains relatively constant.

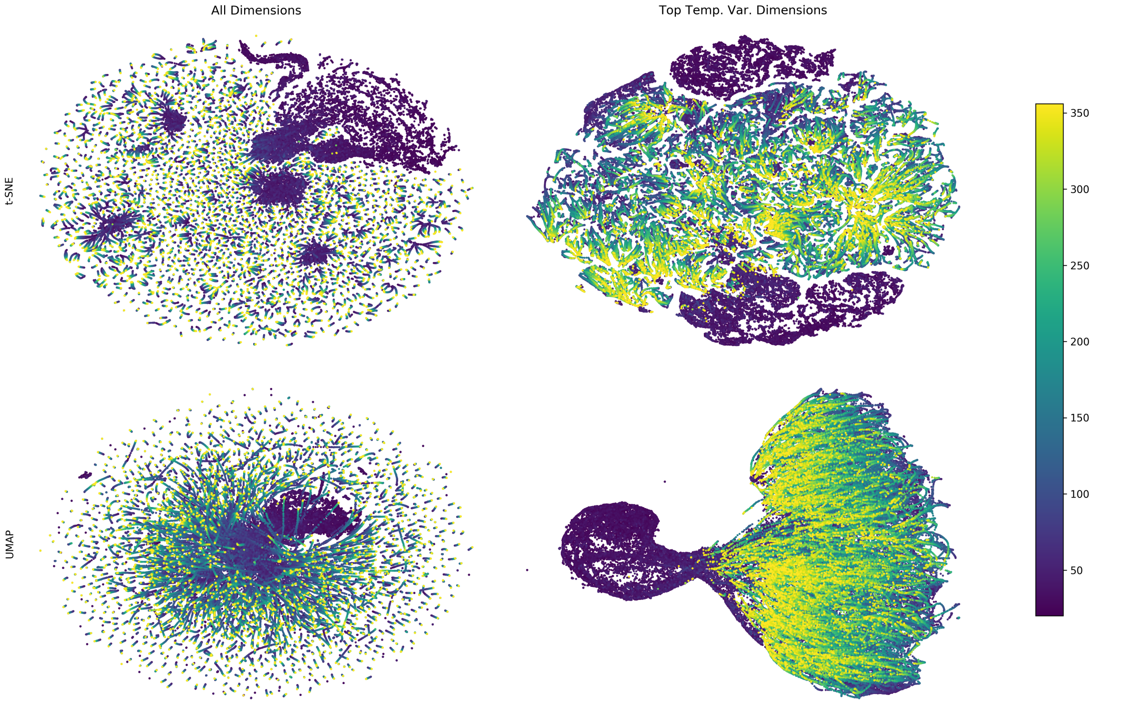


Figure 19: t-SNE and UMAP Visualization of Time

Projections of all latent space representations into 2D using t-SNE and UMAP (top and bottom), using the full latent space representations of all videos and frames (left), and using only the five dimensions with the highest average temporal variance (right).

2.6 Discussion

This chapter introduced an unsupervised approach to translate fluorescence images of hands into a low-dimensional representation. This representation forms the basis for further analyses. The dimensions can be divided into shape-related and signal distribution determining dimensions.

The reconstructions by the VAE model are generally good. They include important image features, such as inflammatory joints, even in a small-sized latent space. If these features appear in the reconstructed image, then all necessary information must be encoded in the latent space. The estimation of the intrinsic dimensionality agrees with the experimental results. Increasing the dimensionality of the latent space correlates with increasing reconstruction quality until the intrinsic dimensionality is reached. Additionally, the employment of a perceptual loss function during the training increases the reconstruction quality.

Hands, that deviate significantly in their shape from the average, show subjectively bad reconstructions. This is underlined by low correlation values and SSIM. The reason for this may be an underrepresentation of these cases in the data set.

Averaging subgroups in the latent space allows for the generation of visually pleasing summary images for these subgroups. Averaging subgroups in the image space is technically possible and straightforward but does not lead to informative images. Every patient has unique hands in terms of shape, size, and position. Thus, the hands are not aligned in the images, and the average signal intensity of a pixel becomes meaningless. However, it is in principle possible to standardize the shapes of the hands and then compute the average image. The comparison of these images with the decoded averages remains for future work. It is noteworthy, that the computed latent space average may be incorrect since the underlying manifold is not regarded.

The presented method reveals marginal differences between PsV and PsA. The decoded averages for both groups differ in signal intensity, as it is expected from the clinical definition of both diseases. However, the magnitude of differences does not meet the expectation. This is the outcome of the poor quality of the assigned labels and agrees with the underlying image data. Although the distinctive characteristic between PsV and PsA patients is the presence of inflammatory joints, a considerable amount of PsV patients shows signs of inflammation. A reliable label that indicates the actual presence of inflammatory joints may lead to more significant differences.

An indicator for a well-learned latent space is the UMAP projection, that incorporate the whole sequences. When restricting the UMAP projection on the dimensions with high temporal variance, the trajectories in latent space agree with the rapid influx and slow wash-out of the fluorescent marker ICG. When using the full dimensionality for UMAP, the variety of shapes distorts the interpretation. However, t-SNE was not able to show the same effect.

3

Enabling explainable Classification of Inflammatory Diseases in FOI with limited Data

3.1 Problem Formulation

As described in the previous chapter, analyzing high-dimensional data with a limited number of data points is difficult. When the training set is too small to represent the variety of each subgroup, it is hard to build a robust, generalizing classifier. Even the supervised training of a small CNN with just a few layers based on a few hundreds of images becomes infeasible, since the dimensionality of the input space and thus the degrees of freedom are too large. The problem becomes more difficult if the labels are ambiguous to some extent or they are not representing the ground truth. In medical image analysis, a classification between healthy and diseased patients is often desired. For mostly ethical reasons however, data sets of clinical trials often contain only diseased patients [96] [97] [98].

As described in the first chapter, the data set used in this thesis contains image data from patients of various diseases and different severities. The data basis for building a generalizing model, which robustly classifies patients into either diseased or healthy, is not given. However, there are groups in this data set, which should be distinguishable at least by definition. For example, it can be assumed that patients with Psoriasis (PsV) show different disease patterns than patients with Psoriatic Arthritis (PsA). The latter group is defined by an inflammatory involvement of the finger joints. Hence, both groups should be distinguishable. Simplifying a classification problem by focusing on only two groups of the data set is limiting the data-availability and worsening the ratio between the number of data points and parameters that have to be trained.

The predictive capability of a classification model strongly depends on the quality of the labels of the training data. In general, labels for medical imaging tasks require experienced physicians. Consequentially, the assigned labels reflect the experience and awareness of the respective physician. This dependency is one of the main reasons for a high interobserver or interrater variability in many medical imaging problems [99] [100]. One solution to this problem is to involve multiple raters and perform a majority vote on the labels. However, this is often not feasible due to financial constraints. Additionally, the quality of labels depends also on a proper problem formulation. Figuratively, if the label is the answer, then its quality depends also on the quality of the question. For example, if the rater must decide whether there is a cat or a dog in the image, the answer and thus the assigned label will most likely be appropriate. However, if the rater has to describe what is shown in the image, the quality of the labels may suffer severely, with respect to the binary classification task of cats and dogs. Even if the task is formulated very specific, the resulting labels in the previous example may be ambiguous, if something else than a cat or a dog is present in the image. Labels for classification tasks applied in medical imaging are often not trivial. Depending on the application, a robust distinction between “healthy” and “diseased” may simply be infeasible. The intuitive solution is to increase the number of classes, with the intention to better represent the reality. But by adding intermediate classes or different grades of the disease, the classification problem can become more complicated. First, this approach reduces the number of examples per class, with the resulting problems as described before. Second, the patients are often not uniformly distributed over the different classes. Classes, which are significantly over- or underrepresented, can introduce severe biases into the classification model [101] [102]. Although there are approaches to correct for such a bias, the classes and their distribution should already be considered during the design of the trial. An increasing finer granularity of the classes also increases the likelihood of a high interobserver or interrater variability since boundaries between certain degrees or grades of a disease can be fluent.

The data set in this study suffers from both problems. Aggravatingly, the disease assignment was done prior to the imaging procedure. Since the data set originates from a study with a different problem, the only disease-related label that is available is the patient’s primary disease during the enrollment of the study. Thus, these labels may correlate with the ground truth, but do not necessarily represent the ground truth. Not only is this impairing the predictive performance of a classification model, it is also limiting the assessment of the performance itself.

A different problem that arises in classification tasks is the understanding of the model’s decision. A classification model can achieve statistically high performance, but for the application in Medical Imaging, it is often a requirement to have plausible or traceable explanations for the individual classification. It is expected, that a model makes predictions or classifications based on medically relevant image features, just as an experienced human rater would do. The so-called Clever Hans phenomenon [103] is generally not desired and can lead to classifications based on background artifacts or left-over captions in the image itself [104]. Having an explainable classification does not only assure the correctness of an individual classification, it can furthermore help to gain insights into the data. As stated in the introduction of this thesis, FOI is a relatively new method that has been established in the research of inflammatory diseases only recently. Hence, the characteristic and distinctive phenotypes of these diseases are not fully understood at this point. Having a reliable, explainable classification model can assist in gaining new insights in this field.

This chapter describes an extension of the unsupervised learning approach presented in the previous chapter. The here proposed approach essentially uses the low-dimensional latent space to build a classifier. This way, the ratio between the dimensionality and the number of data points becomes more favorable. However, this alone is not reducing the problem of ambiguous labels. But since the VAE reduces the dimensionality drastically, it is possible to build a classifier for specific subgroups. Hence, this approach may also be considered as a semi-supervised learning approach. Although the initial situation with respect to data-availability and data quality may seem limited as described above, a sufficiently performing model can be used for further analyses. Assuming the model has learned the features that distinguish the chosen subgroups, it can be used to explain its decision and understand these features.

3.1.1 Related Work

As already introduced, the classification of medical images and images in general has been researched over the past decades and solved for many problems. Given reliable target values, transfer learning approaches can lead to well-performing classification models under low data-availability [44]. However, so-called semi-supervised VAEs have been investigated recently for image classification [105] [106] [107]. The architecture of the VAE in these approaches is comparable to the here proposed approach. The extensions for the classification task are employed using an additional fully connected layer [106], a Support Vector Machine [105], or predefined class centroids [107]. In summary, these approaches achieve higher classification performance compared to supervised CNNs like GoogLeNet with fewer data. With more data used for training, the supervised VAE approaches are outperformed.

3.2 Methods & Experiments

3.2.1 Extending the VAE for Classification

The first approach is an extension of the Neural Network of the VAE. Figure 20 illustrates three approaches that extend the VAE for classification purposes. After successful training of the VAE, the decoder is discarded, and the sampling layer is removed from the encoder. The truncated encoder model is then extended by a Global Average Pooling (GAP) layer and an additional fully connected classification layer with two output nodes, using a Softmax activation to emit class probabilities. The spatially resolved features from the last convolutions are reduced through the GAP layer to a single feature vector. This has been proven to be very effective for classification tasks [65]. This approach (denoted VAE-GAP) aims to build a model that is sensitive to the presence of image features from the convolutional layers.

In an alternative approach (denoted VAE-FC) the encoder is not truncated, and the classification layer is appended after the sampling layer. The motivation behind this approach is to make use of the learned latent space representations, which are the essential image features including their abstract spatial information.

In both approaches, the previously trained weights are frozen, and only the weights of the additional classification layer are trained. Although it is possible to include the classification layer in the primary model und train it simultaneously with the VAE, it is not feasible here due to the limited availability of labels. The training in two phases simplifies the focus on subgroups. In the first step, the VAE is learning a general representation of hands in FOI. Here the full potential of the data set is used, although it is already limited. While the focus on subgroups may worsen the problem of data-availability, the number of parameters, that have to be trained is reduced drastically, since only one layer has to be trained. Its number of parameters depends only on the number of classes or subgroups and the dimensionality of the latent space (VAE-FC) or the number of features from the last convolutional layer (VAE-GAP).

The second approach (VAE-RF) is not an extension of the Neural Network of the VAE itself, but implements a Random Forest based on the low-dimensional representations. Random Forest does not belong to the group of Deep Neural Networks, but it is a proven, powerful Machine Learning method, an ensemble method based on decision trees [108] [109] [110]. Again, the training is performed in two steps: Training of the VAE and training of the Random Forest model. An advantage of the Random Forest is the so-called Feature Importance, that can be computed. This may identify the important dimensions of the latent space, that differ between the chosen subgroups.

3.2.2 Explaining Decisions

A robust, generalizing classifier can be identified through several accuracy metrics and error estimates. In medical image analysis, however, the examining physician (and the patient) are not only interested in the model's classification or prediction.

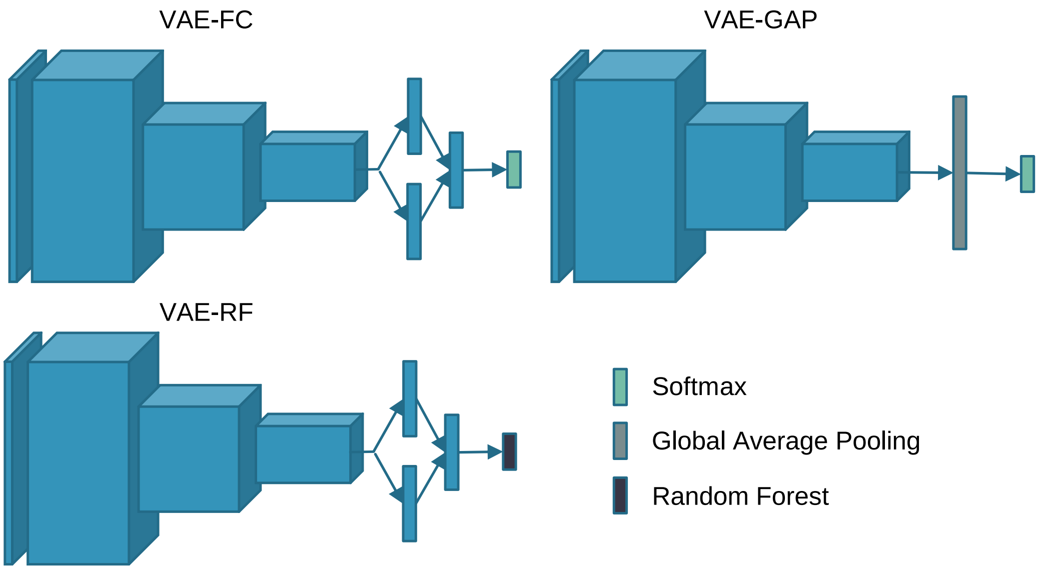


Figure 20: Architecture of the Classification Networks

(left) VAE-GAP (middle) VAE-FC (right) VAE-RF

3.2.2.1 Class Activation Maps

One way to understand the decision making of CNNs are Class Activation Maps (CAM) [111]. CAM aims to represent the activation of a classification CNN for a given class and a given input image. To achieve this, the activation output of the last convolution (or pooling) layer of the network, which still contains spatial information for each learned feature, is multiplied with the weight matrix w of the fully connected classification layer. The CAM for a given input image and class c is defined as

$$M_c(x,y) = \sum_k w_k^c \, f_k(x,y)$$

where $f_k(x,y)$ is the output of the last convolution for the given input image at feature k and spatial position (x,y). This weighted linear sum of the feature maps from the convolutional output is upscaled to the original image size to identify image regions, which generate the activation for the class c.

3.2.2.2 Occlusion Maps

Occlusion Maps are a simple way to identify image regions, which are important for a class prediction. Therefore, a small patch of the input image is occluded by setting the intensity values to zero, the average intensity of the image, or Gaussian noise. Then the occluded image is fed forward through the model and the class probability is computed. This procedure is performed repeatedly by sliding the patch over the complete image. For each position of the patch, the change in probability is calculated and shown as a 2D heatmap. Important image regions will show significantly lower probabilities for the respective class.

3.2.2.3 Local Interpretable Model-Agnostic Explanation

Local Interpretable Model-agnostic Explanations (LIME) [112] treat the model as a black box. In contrast to Occlusion Maps the model's prediction is evaluated based on permutations of a given input image. Therefore the image is over-segmented into many segments using quickshift [113]. Random combinations of these segments are classified using the model to compute the importance of each segment. The importance is determined by computing the variation of the output probability depending on the presence or absence of the image segment.

3.2.3 Control Experiment on synthetic Data

As stated at the beginning of this chapter, the success of a classification task highly depends on the quality of data. To reveal the full potential of the approaches above, an additional control experiment is conducted. This experiment follows the procedure as the other experiments, but is performed on synthetic data, which is derived from the FOI data set. The synthetic data set aims to represent an ideal scenario with two distinct groups. The synthetic "diseased" group is characterized by a bright spot on the area between MCP joints and carpus, representing a prominent inflammation of the metacarpal bones. The second group lacks this distinct feature. The data set is generated by randomly selecting 50% of the patients and manually adding a synthetic inflammation template on one of the patient's hands. The template is a filled circle with a radius of 25 pixels, that has been blurred with a Gaussian filter with large standard deviation ($\sigma = 10$). The signal intensity of the added template is scaled to 90% of the brightest signal in the individual image, so it does not introduce an additional scaling bias. The synthetic "healthy" group has not been altered.

This data set is an ideal control in many aspects. The respective labels can be considered accurate since all of the "diseased" patients actually have the desired disease pattern. Furthermore, all of the "healthy" patients lack this particular disease pattern. The chosen template resembles an actual inflammatory joint. However, the template is positioned between the MCP joints and carpal joints. There are only the metacarpal bones and no joints that can be already inflammatory. This reduces the chance of false-positives in the "healthy" group.

3.3 Results

This section presents the results of the classification experiments. At first, the above-introduced approaches are compared. These are extensions of the VAE model from the previous chapter, using additional layers and a Random Forest. Afterward, the classification is evaluated for different subgroups based on the available clinical data for the patients. The labels that define these subgroups vary in their quality and clinical relevance. The classification explanations are then evaluated and compared. Finally, the results from the control experiment with the synthetical FOI dataset are presented in terms of classification metrics and explainability.

3.3.1 VAE allows classification of subgroups

To investigate the classification capabilities of the approaches, different labels are used target variables. Here the classification of Rheumatoid Arthritis (RA) and Psoriasis (PsA/PsV) is investigated, as well as the sex, the Disease Activity Score (DAS28), and the Tender Joint Count (TJC28) of the patients. For simplicity, the continuous DAS28 and ordinal TJC28 have been binarized. The DAS28 is cut-off at a score of 3.2, which is the defined threshold between "inactive" and "moderate activity". The TJC28 is grouped into "no tender joints" ($TJC28 = 0$) and "tender joints present" ($TJC28 \geq 1$). Since the original data are videos, the classification capabilities are tested at different time points of the videos.

When extending the neural network of the encoder into a classification network, the resulting model achieves an accuracy of up to 62% for the classification of the two Psoriasis groups, depending on the chosen time point. However, when using only the pre-trained Convolutional layers followed by a GAP and a classification layer (VAE-GAP), the accuracy is most of the time close to 50%. Figure 21 shows the classification statistics for both extended encoder approaches. The VAE-GAP approach shows its highest accuracies during the early enhancement phase. The VAE-FC approach shows comparable accuracies during the early enhancement phase at single time points but is significantly superior only during the intermediate phase between 150s and 200s.

Using Random Forests, the prediction accuracy between the two Psoriasis groups varies around 70% in dependency of the time point and can reach up to 73.8% after 70 seconds into the video. Figure 22 shows the prediction accuracy of the VAE-RF model for selected time points and clinical parameters. The clinical parameter that can best be distinguished is the sex of the patient, with an accuracy of up to 87.84% after 100 seconds. The TJC28 can be classified as good as Psoriasis. But in contrast to Psoriasis, here the accuracy is slowly increasing

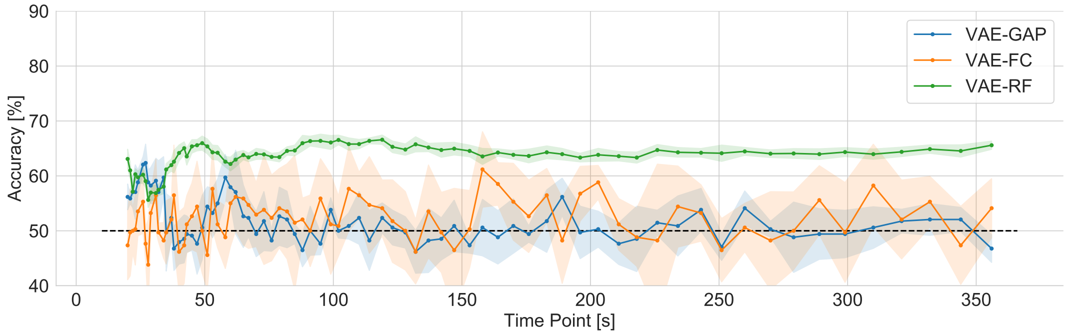


Figure 21: Accuracy over Time of the CNN models

The accuracy for the classification of PsV vs. PsA is shown for both methods, VAE-GAP (blue), VAE-FC (orange) and VAE-RF (green) in dependency of the time point of the video. The bold line denotes the mean, whereas the shaded area denotes the standard deviation. The VAE-RF achieves overall the best results.

until it reaches its maximum of 62.8% at 90 seconds. The classification accuracy of RA is very poor and does not exceed 60% at any time point. The accuracy for the DAS28 peaks around the same time as the TJC28 and achieves an accuracy of 62.9%, which is subsequently decreasing.

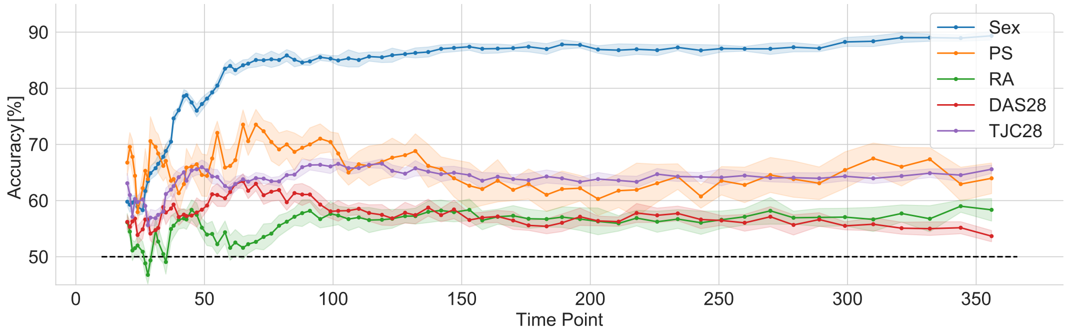


Figure 22: VAE-RF Statistics

The accuracy for the classification of clinical parameters using the VAE-RF model is shown in dependency of the time point of the video. The bold line denotes the mean, whereas the shaded area denotes the standard deviation.

When looking at the feature importance (FI) from the Random Forest in Figure 23, only a small number of dimensions of the latent space are relevant for each of the classification tasks. The interpretation of each FI for each clinical parameter is not trivial. However, there are a few noticeable correlations. The classification of DAS28 and TJC28 share their most important dimension. Also, RA and TJC28 have a few important dimensions in common. However, the patient's sex seems to rely on more important dimensions, than the other clinical parameters.

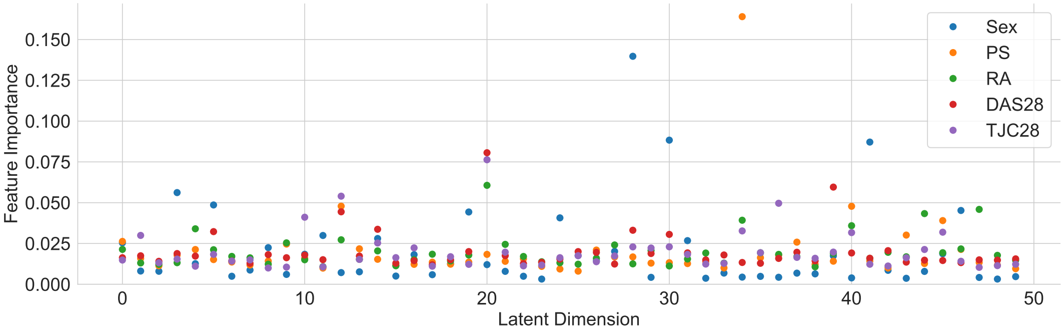


Figure 23: Random Forest Feature Importance

Random Forest Feature Importance for the five classification tasks. The classifications of clinically relevant labels share their most important dimensions with each other. The classification of the patient's sex is made by the shape of the hand and relies on a different set of dimensions.

3.3.2 Classifications can be explained

Figure 24 shows the computed Occlusion Maps and LIME results for two example patients from the two Psoriasis groups. Both patients are classified correctly into PsV and PsA by the VAE-RF model. The Occlusion Maps reveal that almost all parts of the hand decrease the probability of a PsA classification (blue areas) in the PsV patient. The PsA patient shows mostly neutral regions. The area of the index fingers and the thumbs, which appear relatively bright in the original image, favor the PsA classification (red areas). The results from LIME indicate mostly regions that favor a PsV classification (blue) in the PsV patient's hands. However, a few regions around the index finger on the left hand seem to favor a PsA classification. Most parts of the hands of the PsA patient show only weak tendencies towards any classification. Only the thumb of the right hand appears to be strongly indicating a PsA classification.

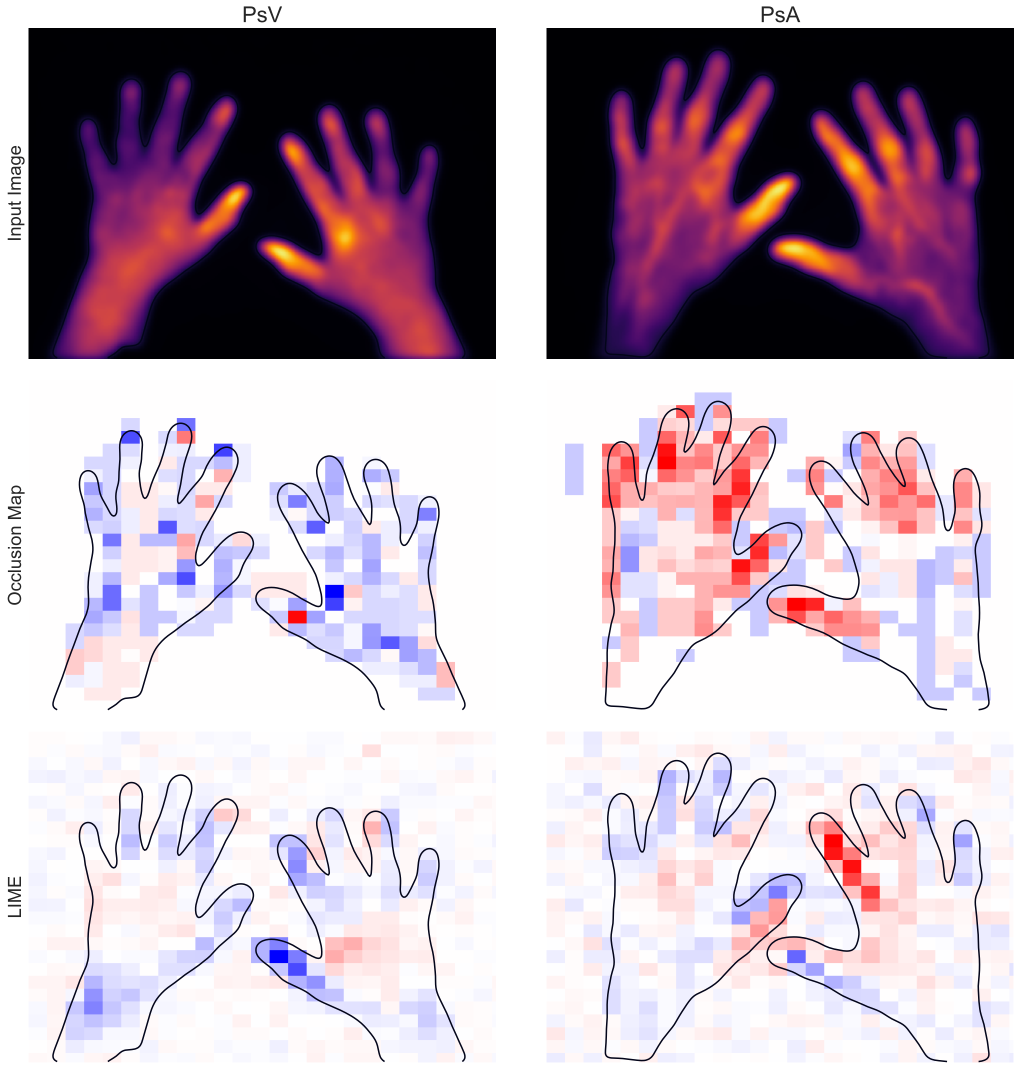


Figure 24: Classification Explanation

Classification Explanation using Occlusion Maps and LIME on images from a PsV (left) and PsA (right) patient. The original image is shown in the top. The Occlusion Maps (center) and LIME (bottom) indicate healthy-favoring regions in blue and disease-favoring regions in red.

3.3.3 Control Experiment confirms Feasibility of Pipeline

When repeating the experiments above with the synthetic data set, the classification with the trained encoder and subsequent Random Forest performs the best in terms of accuracy (96.50 ± 0.30, mean±standard deviation), area under the curve (AUC: 99.63 ± 0.04), and false-positive rate (FPR: 5.05 ± 0.57). Only its false-negative rate (FNR: 1.95 ± 0.35) is slightly outperformed by the VAE-GAP approach (1.55 ± 0.15). The VAE-FC approach achieves the worst performance in all four metrics (Accuracy: 93.65 ± 0.55, AUC: 98.55 ± 0.28, FNR: 3.55 ± 1.54, FPR: 9.15 ± 1.79). However, all approaches achieve better results on the synthetic data set than on the original FOI data set. The complete overview of classification metrics for all three approaches is shown in Figure 25.

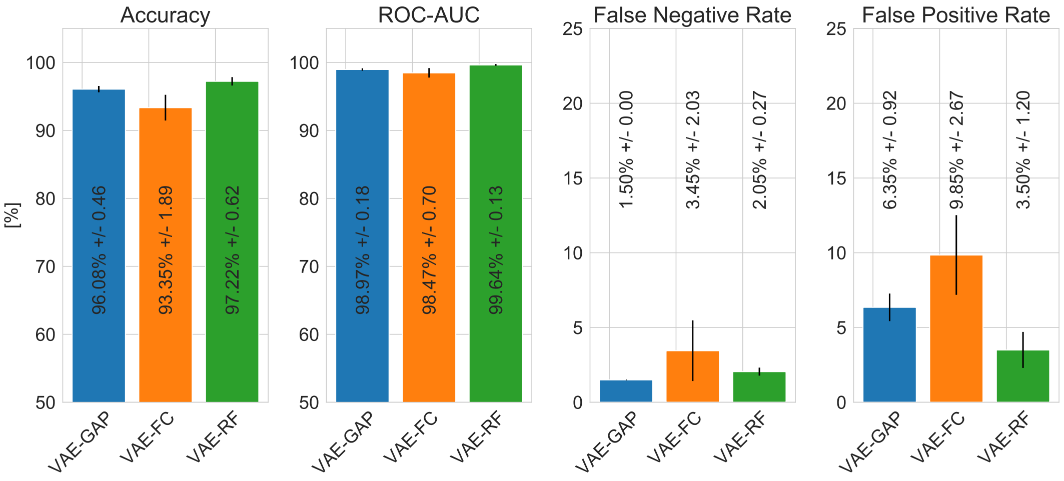


Figure 25: Classification Metrics for the Control Experiment

The classification performance measured in accuracy, ROC area under the curve, false-negative rate, and false-positive rate (from left to right) for the three classification approaches. All approaches use the encoder of the trained VAE with different extensions. VAE-GAP uses only the Convolution layers followed by a Global Average Pooling layer and a fully connected classification layer, VAE-FC uses the full encoder with a subsequent a fully connected classification layer, and the VAE-RF trains a Random Forest based on the latent space representation from the encoder.

The classification explanations for the control experiment are shown in Figure 26 for two randomly selected cases. Both methods, Occlusion Maps and LIME, mark the area of the inserted template of the "diseased" patient as highly favoring the "diseased" classification. The "healthy" patient's Occlusion Map is mostly covered in blue, suggesting that most parts of the hands are not favoring a “diseased” classification. However, the explanation with LIME suggests large areas of the hands as “diseased”-favoring. In contrast to the Occlusion Maps, the explanation maps computed with LIME indicate in both patients, that some parts of the image background contribute to the “healthy” classification, while other parts contribute to a “diseased” classification. In the few cases where the prediction is incorrect, the predicted probabilities are near the 50% threshold. An example with two failure cases and corresponding Occlusion Maps can be found in the Appendix. The example with a false-positive prediction shows a pattern that resembles the inflammation template, although it has not been inserted.

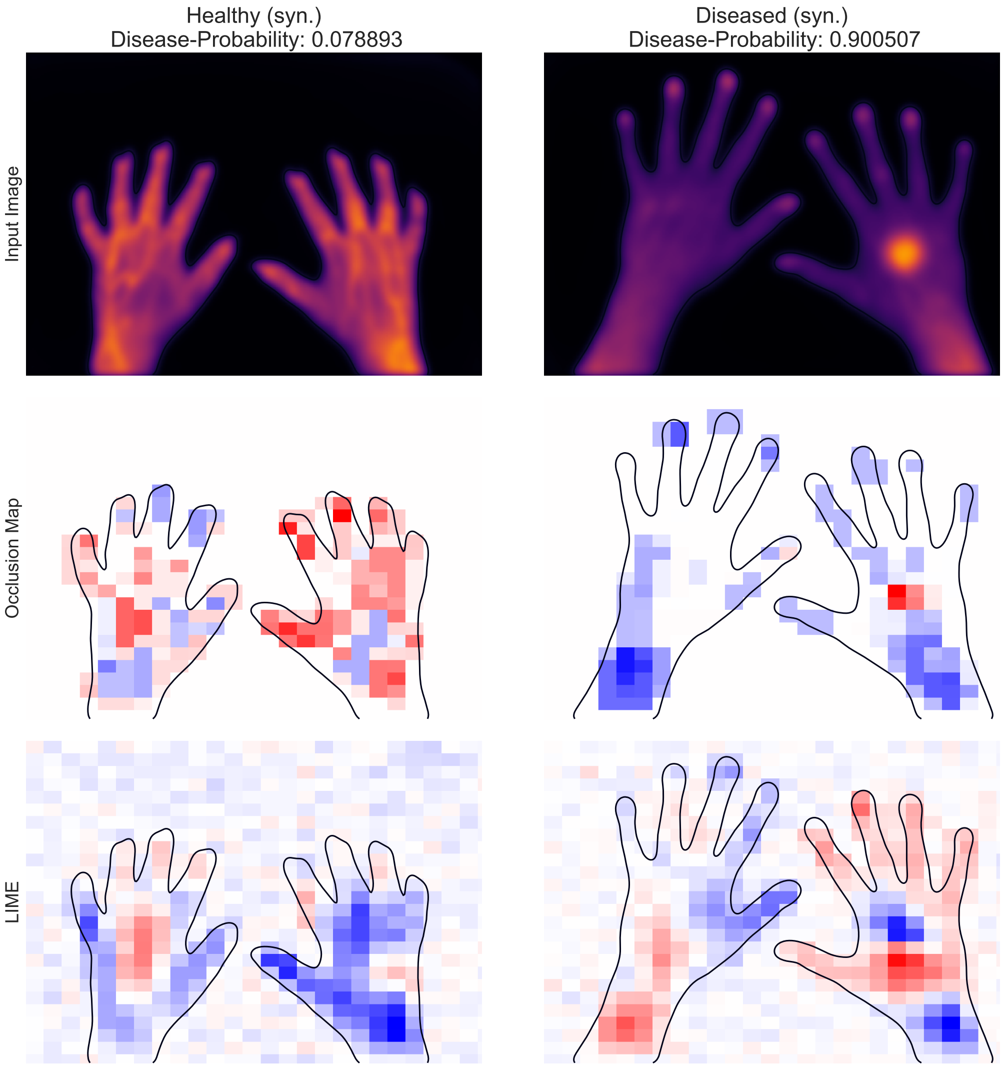


Figure 26: Classification Explanations for the Control Experiment

Classification Explanation using Occlusion Maps (center row) and LIME (bottom row) on images from a “healthy” (left) and “diseased” (right) patient. The inserted template for the synthetic “diseased” group can be seen on the left hand of the right patient.

3.4 Discussion

The here achieved classification accuracies of RA and PsA vary between 60% and 80%. Compared to other binary classification tasks, where accuracies close to 100% are not unusual, these values are not competitive at first sight. The success of a classification task depends on multiple factors, before the actual fitting of the model. First, the assigned labels in the data set must be correct. Second, the desired difference must be present in the data. As introduced at the beginning of this chapter, the quality of the assigned labels may be the main cause for the comparatively low performance. It is unknown how reliable the assigned diagnoses for each patient are. Furthermore, it is unclear how strong the actual difference in the images is. This is underlined by the fact that the VAE-RF model achieves a very good accuracy when trained on the target variable sex. While the categorization into the different diagnoses can be ambiguous and suffers a high inter-rater variability, the definition of sex is assumed to be less erroneous.

One observation from analysis of the feature importance is that the classification of the patient's sex mostly relies on distinct latent space dimensions than those of the other clinical parameters. The classifications of RA, PS, DAS28, and TJC28 are expected to be based mainly on their signal distributions since they are all intended to separate patients with and without inflammatory joints. In contrast, it can be assumed that the classification of the patient's sex is made only by the size and shape of the hands. This is supported by the different average body weights of male and female patients, which may affect the size and shape of their hands. A problem is the comparison of the feature importance for each clinical parameter. With an increasing number of clinical parameters and latent space dimensions, the task becomes too exhausting for a manual or visual inspection. A further improvement could be a correlation analysis or a hierarchical clustering to identify common important features and clinical parameters with similar phenotypes.

The classification explanations with Occlusion Maps and LIME deliver reasonable explanations of the Psoriasis classifications. The two methods do not give the same results, but the interpretations of both are also not conflicting. The interpretations of the explanations seem generally plausible. Areas, which favor a PsA classification, seem to correlate with relatively higher signal intensities in the input images. This agrees with the defining, distinguishing characteristic of the two Psoriasis groups. As introduced in the first chapter, it can be assumed that PsA patients have inflammatory joints and thus higher signal intensities in the finger.

The control experiment with the synthetical data set confirms the feasibility of the analysis pipeline under ideal conditions. As introduced at the beginning of this chapter, the synthetical data set has accurate labels that reflect the desired disease pattern. The classification experiment achieves very high accuracies with low false positive and false-negative rates. This highlights the importance of data quality. The overall tendencies from the previous experiments are confirmed, the combined approach of the VAE's encoder with subsequent Random Forest achieves the best results. An even better classification performance can be achieved with further tuning of the hyperparameters. However, this tuning was omitted, and training was performed with the same parameters as in the previous experiments in order to have comparable results. Further optimization of these parameters might also reduce the number of false predictions since their predicted probabilities are close to the 50% threshold. Without further optimization, either false-positives or false-negatives can be reduced by adjusting this threshold.

But based on this classification, interpretable and plausible explanations can be computed using Occlusion Maps. These highlight the area of the inserted template clearly. The explanation generated by LIME does not provide contradicting results, but their interpretation is less straightforward since they implicate importance in parts of the background. This effect can also be seen in the experiments with the original FOI data and may come from the fact that random combinations of image segments are combined and tested.

4

Data-driven Decomposition of Time-resolved Disease Pattern

4.1 Problem Formulation

The previous two chapters present approaches for exploration, summarization, and classification of medical images without considering time. Some diseases may differ only in their temporal behavior, and the underlying dynamical system may be unknown, complex, and non-linear.

Research of inflammatory diseases using FOI is very basic to date. As mentioned in the previous chapter, the analysis of the FOI data is performed manually. The influence of the time component of the data is mostly neglected. Either the individual frames are summed up entirely or within the three defined phases of signal distribution. Only a few studies perform a time-dependent analysis of the image data [114] [115]. However, these approaches only measure signal intensities of manually defined regions (e.g., finger joints) or compute descriptive statistics (e.g., average and maximum signal intensity) for each frame and plot these. An in-depth analysis of the underlying dynamics or relations between features and time is not performed.

There are several studies that apply methods from the field of Dynamical Systems on time-resolved medical imaging data. Recently, Casorso et al. [116] have applied the eigenmodes of autoregressive models [117] to functional MRI (fMRI) data, by comparing the fMRI data of brains in resting state versus brains under motor-task conditions. This approach is able to reveal motor areas, that are not found with classical methods. Kunert-Graf et al. [118] use Dynamic Mode Decomposition and compare the resulting modes between individual patients as well as groups of patients to reveal resting state networks in fMRI data.

This chapter introduces a new approach, which approximates the underlying dynamics with a linear operator in the low-dimensional latent space. The major extension to the previous approaches is the usage of two VAEs. The two VAEs each translate a frame at time point t and $t + 1$, respectively, into latent space. Additionally, a linear operator aims to predict the latent space representation from t to $t+1$. The introduction of a temporal dependency is aiming to regularize the latent space. This enforces the dynamical behavior to be linear in latent space. After successful training of the model, this linear operator can be used for further analyses. However, in contrast to other studies and practices, a single model and thus a single operator is trained. While this general approach of finding a linear operator, that describes the temporal behavior, is quite common, it is usually applied to data from a single experiment. This procedure is useful, when a manageable amount of experiments is analyzed and compared. In fluid dynamics, this procedure can be used to analyze the airflow around a given structure and compare the airflow behavior in dependency of the modifications of that structure. Given the number of cases and the poor quality of clinical labels in the FOI data set, the here presented procedure does not compare the characteristics of individual operators. It rather compares the individual deviations from the global linear operator. Furthermore, established methods are compared with the here presented approach. The proposed new architecture aims to solve the problem of finding a linear operator for time series propagation. The approach aims to achieve this by simultaneously solve the problem of learning a low-dimensional space in which this linear operator works well.

4.1.1 Related Work

Dynamical systems can be analyzed by a variety of methods such as Dynamic Mode Decomposition and its derivatives [119] [120]. However, this thesis focuses on data-driven approaches using Neural Networks. The combination of Neural Networks and Koopman operators has recently been investigated with time-lagged autoencoders or their derivatives [121] [122]. Time-lagged autoencoders extend the framework of autoencoders for time series such that for a given data point at time point t the succeeding data point at $t + 1$ is predicted. Lusch et al. [122] wrap the Koopman operator into an auxiliary network, that parametrizes continuous frequency spectra. Instead of approximating the Koopman operator, the auxiliary network approximates only the eigenfunctions of the hypothetical operator. The VAMPnet, proposed by Mardt et al. [121], aims to approximate a Markov State Model (MSM). This is realized by appending a Softmax layer to the encoder and thus encoding a state probability vector. By applying the variational approach for Markov processes (VAMP) and the derived VAMP-2 score, VAMPnets can be trained without a decoder network and thus without reconstruction loss. This approach is similar to a Deep Canonical Correlation Analysis (DCCA), where two non-linear functions in the form of neural networks are learned, whose representations are highly linearly correlated.

4.2 Linear Time Evolution Operators

In a variety of research areas, such as engineering and physics, the understanding of dynamical systems plays an important role. The behavior of such dynamical systems can be chaotic, non-linear, non-deterministic, or just unknown. While there are many different approaches to analyze these systems, they are often approximated using low-dimensional linear time evolution operators. A common example of such an operator is the transition matrix P of a Markov State Model (MSM). This operator propagates a state probability vector one time step further. By decomposing P into its eigenvalues and eigenvectors, stable and meta-stable distributions can be revealed. The application of MSMs has been successfully demonstrated for many problems including conformational dynamics. In the latter, MSMs have also been combined with Deep Neural Networks. Recently, Koopman operators have been combined with Deep Neural Networks to tackle similar problems. In contrast to MSMs, Koopman operators do not propagate state probabilities. Instead, they live in an observable space. This chapter aims to make use of the idea behind the Koopman Operator Theory by including a linear operator for time evolution into the VAE framework.

4.2.1 Koopman Operator Theory

The dynamics of dynamical systems are usually time-continuous processes. However, measurements of these systems happen mostly in discrete time steps. Hence,

$$x_{t+1} = \phi\left(x_t\right)$$

describes the evolution of the state $x \in \mathbb{R}^n$ from time point t one step to $t+1$. The dynamics of ϕ are usually high-dimensional and non-linear for real-world problems. Additionally, ϕ itself is usually unknown. The Koopman Operator Theory states that a dynamical system can be approximated linearly if transferred to an infinite-dimensional space [123]:

$$Kf\left(x\right) = f\left(\phi\left(x\right)\right) = (f \circ \phi)(x)$$

Thus, the Koopman Operator K can be used to evolve a system's observables in time:

$$f\left(x_{t+1}\right) = Kf(x_t)$$

In theory, f can be any function in L^∞. However, a low-dimensional discretization of K can be approximated with a corresponding f. If such a linear operator $\widehat{K}$ can be estimated with low error, it can be used for predicting or understanding time series. Additionally, f can be represented as a Neural Network, whose weights are chosen such that $\widehat{K}$ approximates K well enough. Due to the high dimensionality and the resulting requirement for high data-availability and high computational power, the approximation of such an operator has become feasible only recently. Alternative approaches to approximate Koopman operators are Dynamic Mode Decomposition and its derivatives [119] [120].

One way to interpret the dynamics is the decomposition of the linear operator into its eigenvalues and eigenfunctions of this operator [124] [125]. This spectral analysis can reveal the effects of different frequencies and is suitable for stationary, recurring and time-reversible processes. However, if the underlying process is assumed to be non-reversible, the Schur decomposition has been proposed as an alternative [126], if the matrix has the properties of a stochastic matrix. Alternatively, a singular value decomposition (SVD) is suggested [127] [128]. In general, the SVD of a $(m \times n)$ matrix M is defined as follows:

$$M = U\Sigma V^*$$

The diagonal entries of the matrix Σ contain the singular values, which correspond to the eigenvalues. They are the square roots of the non-zero eigenvalues of both MM^* and M^*M. The $(m \times m)$ matrix U and the $(n \times n)$ matrix V^* contain the left- and right-singular vectors, respectively. These singular vectors form a set of orthonormal vectors or basis vectors. In contrast to an eigenvalue decomposition, the SVD will result in real-valued singular values and vectors, if the matrix M is real-valued, which makes the subsequent analysis more manageable. For simplicity, the singular values are assumed to be already sorted in descending order in this chapter. The singular vectors with large singular values are denoted dominant singular vectors.

4.3 Methods & Experiments

In order to use a linear operator like the Koopman Operator within the VAE framework, the architecture has to be extended. Essentially, two copies of the VAE for the time points t and $t+1$, respectively, are connected in latent space. Figure 27 shows the extended architecture. The encoder of the VAE is representing the function f in the Koopman equation. The transfer operator $\widehat{K}$ is implemented as a fully connected layer. In contrast to regular practice in neural networks, the bias term is set to zero, and no activation function is used. This layer is attached to the latent space representation y_t of the input image x_t, which is generated by the encoder. The output $\widehat{y}_{t+1}$ estimates the latent space representation y_{t+1} of the succeeding frame x_{t+1}. The loss function is extended by adding the error of the estimated $\widehat{y}_{t+1}$ and the actual y_{t+1} from the encoder output of x_{t+1}:

$$Loss_K = \left\| \widehat{K} f(x_t) - f(x_{t+1}) \right\|_2^2 = \left\| \widehat{K} y_t - y_{t+1} \right\|_2^2 = \left\| \widehat{y}_{t+1} - y_{t+1} \right\|_2^2$$

It is crucial that the two copies of the VAE share the same weights. Otherwise, the function f in the Koopman equation would be different on both sides of the equation. It is noteworthy, that this approach of shared VAE weights does not double the number of parameters that must be optimized. Only the new parameters of the linear operator have to be fitted additionally. This increases the number of parameters only by the squared number of latent space dimensions. Thus, the increase in computational effort is manageable.

In this chapter, two experiments are conducted to evaluate the approach. The general procedure for these experiments begins with the training of the extended model (here denoted: Koopman-VAE) on the time series. After the model has converged to a low error level, the weights of the linear operator, which represent the transfer operator K, can be analyzed. The underlying dynamics are computed by evaluating the eigenfunctions of K on the trajectories in latent space.

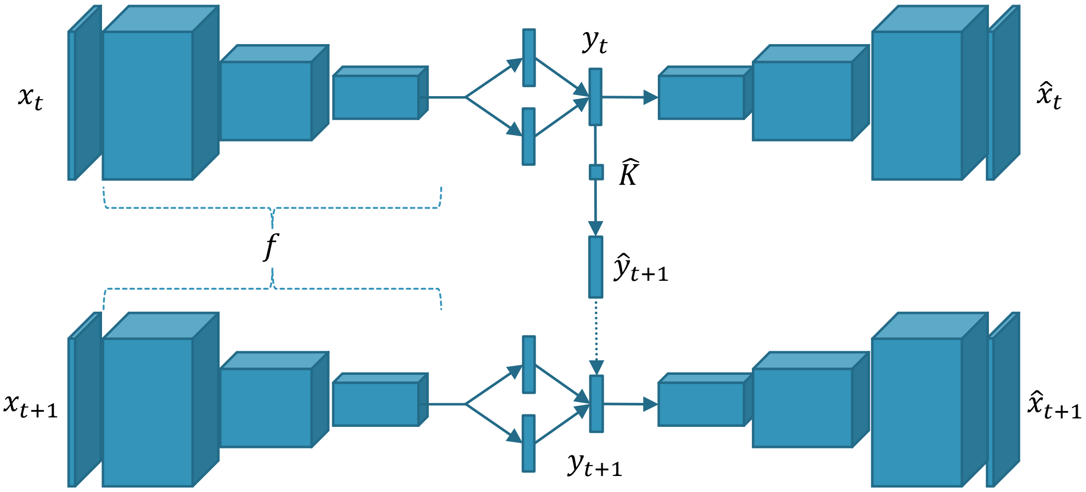


Figure 27: Architecture of the Koopman-VAE

The architecture of the Koopman-VAE is a combination of two VAEs which are interconnected in latent space by a fully connected layer. The first VAE handles the data at time point t, whereas the second VAE handles the succeeding time point $t + 1$. The weight of the additional layer represents the linear operator and propagates the latent space representation from the first to the second time point.

Two experiments are conducted, one with known dynamical behavior, one with unknown. First, the Lorenz system will be used. Here the dynamics are known, well described, and previously analyzed with different methods. The second experiment will be performed on the FOI data from the previous chapters. Here, the underlying dynamics are unknown. However, the results from the previous chapters about this particular data set and the results from this chapter about the method from the first experiment may help to evaluate the results.

4.3.1 Lorenz System

To validate the approach, the first experiment is performed with a toy data set, where the underlying dynamics are known. The Lorenz system is a non-linear system of ordinary differential equations with chaotic behavior. It is defined as follows:

$$\begin{aligned}\frac{dx}{dt} &= a\,(y-x)\\ \frac{dy}{dt} &= x\,(b-z)-y\\ \frac{dz}{dt} &= xy-cz\end{aligned}$$

With the constants set to $a = 10$, $b = 28$, and $c = \frac{8}{3}$ the system shows its known behavior with two attractors. Estimations of the dynamical properties using linear approximations have successfully been applied by employing DMD and SINDy [129] [130] [131].

Here, the Lorenz system is realized as a moving filled circle on an image plane. The trajectory of this circle is given by the (x, z)-coordinates of the Lorenz system, which represent the horizontal and vertical position on the image plane, respectively. Figure 28 shows a single frame together with the full trajectory of the circle. The system is simulated for 500 time points, and the trajectory is translated into image frames. Then the Koopman-VAE model is trained, and the dynamics are computed. Additionally, the trajectory and the signal intensity of the generated frames are distorted with Gaussian noise of low variance ($N(\mu = 0, \sigma = 1)$ and $N(\mu = 0, \sigma = 0.02)$, respectively).

This chapter aims to show that the combination of a VAE with a linear operator for propagation in time is feasible. Therefore, a detailed hyperparameter evaluation is omitted. A latent space dimensionality of eight has been proven to be good enough for this purpose and is used for the experiment here.

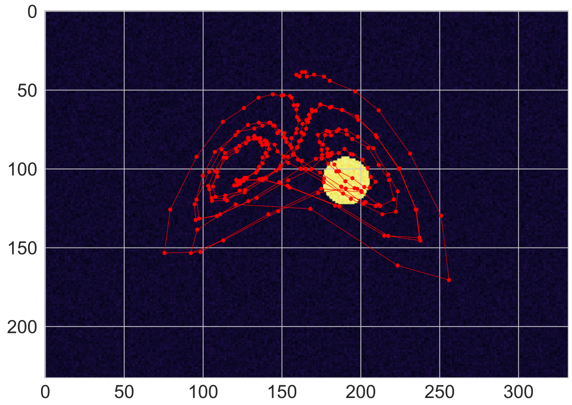


Figure 28: Trajectory of the Lorenz-System

The Lorenz system is here realized as an image problem. The computed trajectory is projected on the image plane. The position at each time point is realized as a filled circle moving along the trajectory.

4.3.2 FOI Data

The second experiment aims to discover the underlying dynamics of FOI data. In contrast to the previous example, the dynamic behavior is unknown. Here, the Koopman-VAE is trained on the frames of the FOI examination. The model is trained on the whole training set of patients. The resulting operator K and its eigen- and singular value decomposition are then used to analyze the system in general, and to evaluate the individual dynamics of each patient's time series. These individual dynamics can then be aggregated by different clinical parameters, like the diagnosed disease, in order to understand the effect of these factors on the dynamical behavior.

4.4 Results

This section shows the results of the previously introduced experiments. These aim to show the capabilities of the Koopman-VAE, which consists of two identical VAE models, that are joined with a linear operator in latent space. The analysis of the trained Koopman-VAE model on the Lorenz system is presented. In accordance with the result sections of the previous chapters, the reconstruction quality is elaborated. Subsequently, the results from the eigenvalue and singular value decomposition are shown and compared with the original data. The analysis of the FOI data is following the same scheme. However, due to a missing ground truth of the dynamics, a comparison with original coordinates is not possible. Instead, the revealed dynamics are compared between subgroups of patients. Furthermore, it is evaluated whether the transformations with eigen- and singular vectors enhance the classification abilities of the method.

The following notation is used: LS-n denotes the n-th dimension of the latent space trajectory. EV-n denotes the evaluation of the n-th eigenvector on the latent space trajectory, beginning with EV-0, which originates from the dominant eigenvector. Analogously, SV-n denotes the evaluations of the left singular vectors. The eigenvectors and singular vectors are evaluated on the latent space trajectories by using the dot product.

4.4.1 Koopman-VAE learns Lorenz System

As in the second and third chapters, also the extended model is able to reconstruct the input images with high similarity. The mean SSIM is 0.89, and the pixel-wise Pearson correlation is 0.85. The visual analysis of the reconstruction generally shows soft edges compared to the original image. Figure 29 shows the reconstructions of the full circle at selected positions. The quality of the reconstruction varies with the position of the circle in the frame. These are positions at which the circle is moving with higher velocities.

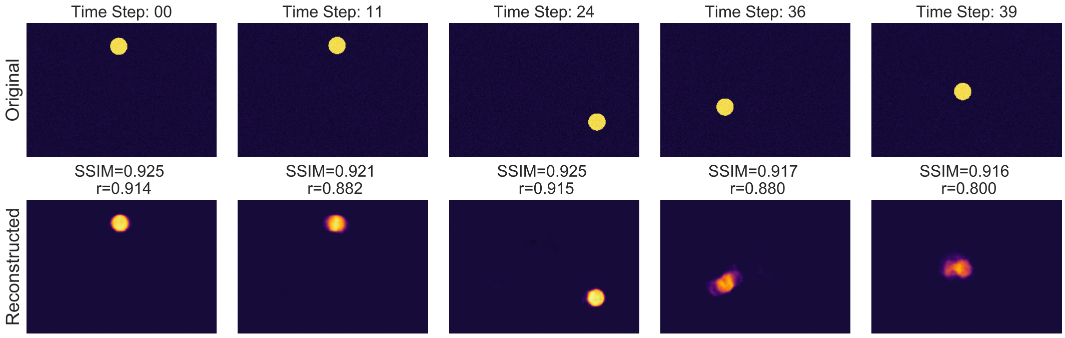


Figure 29: Reconstructions of the Lorenz System Frames

(top) Input frames (bottom) corresponding reconstructions by the VAE.

The latent space representations of the encoded Lorenz System are shown in dependency of the original coordinates in Figure 30, without considering time. Some dimensions correlate strongly with either the horizontal or vertical position (X- and Y-coordinate, respectively). Some dimensions correlate only partially. The horizontal position shows the strongest correlation with the second latent space dimension (LS-1, r=-0.79), whereas the vertical position correlates the most with LS-4 (r=-0.88). Besides apparent linear correlations, some latent space dimensions show a circular, almost spiral-like pattern (e.g., X vs. LS-7 and Y vs. LS-1).

Figure 31 compares the eigenvalues and singular values of the linear operator for the Lorenz system. The eigenvalues range between 1.014 and 0.903. Four eigenvalues are real-valued, the other are complex conjugates. All eigenvalues are close to the unit circle, except for the smallest one. Thus, their eigenvectors may represent metastable points in the latent space. The complex eigenvalues indicate a non-reversible process.

The singular values are distributed in a broader range, with a larger gap

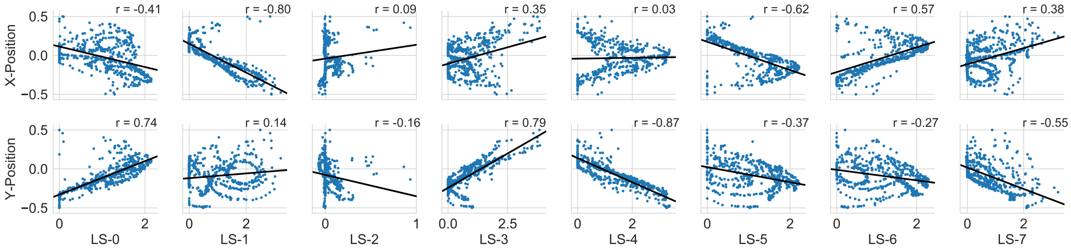


Figure 30: Latent Space representations of the Lorenz System

Relative horizontal (top) and vertical (bottom) position of the circle versus the individual dimensions of the latent space representation. The black line represents a linear regression, the correlation coefficient is annotated in the top right of each plot.

between the largest and second-largest singular values. Furthermore, the two smallest singular values are slightly detached, but less than the dominant singular value.

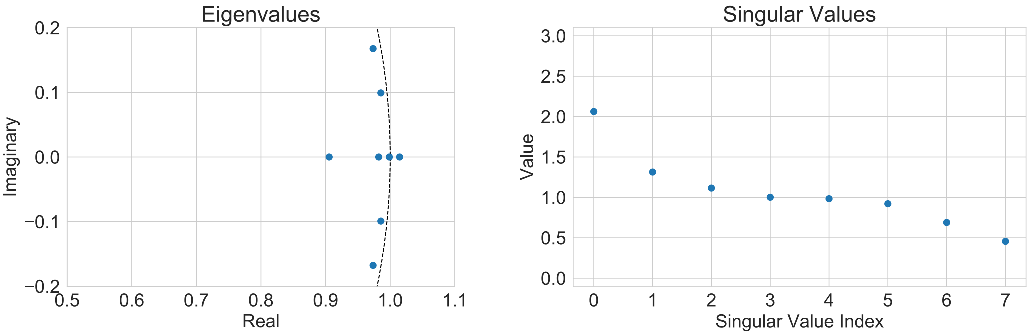


Figure 31: Eigenvalues and Singular Values

Eigenvalues (left) and Singular Values (right) of the linear operator that was trained on the Lorenz system.

The eigenvectors and their evaluation on the latent space trajectory are shown in Figure 32. Generally, the eigenvector evaluations show similar characteristics, each with recurring a pattern. Only the evaluations of the last eigenvector (EV-7) has a visually different pattern compared to the others. EV-0 and EV-4 reveal nearly identical temporal behavior.

Figure 33 shows the results from the singular value decomposition analogously to the results from the eigenvalue decomposition. In contrast to the evaluations

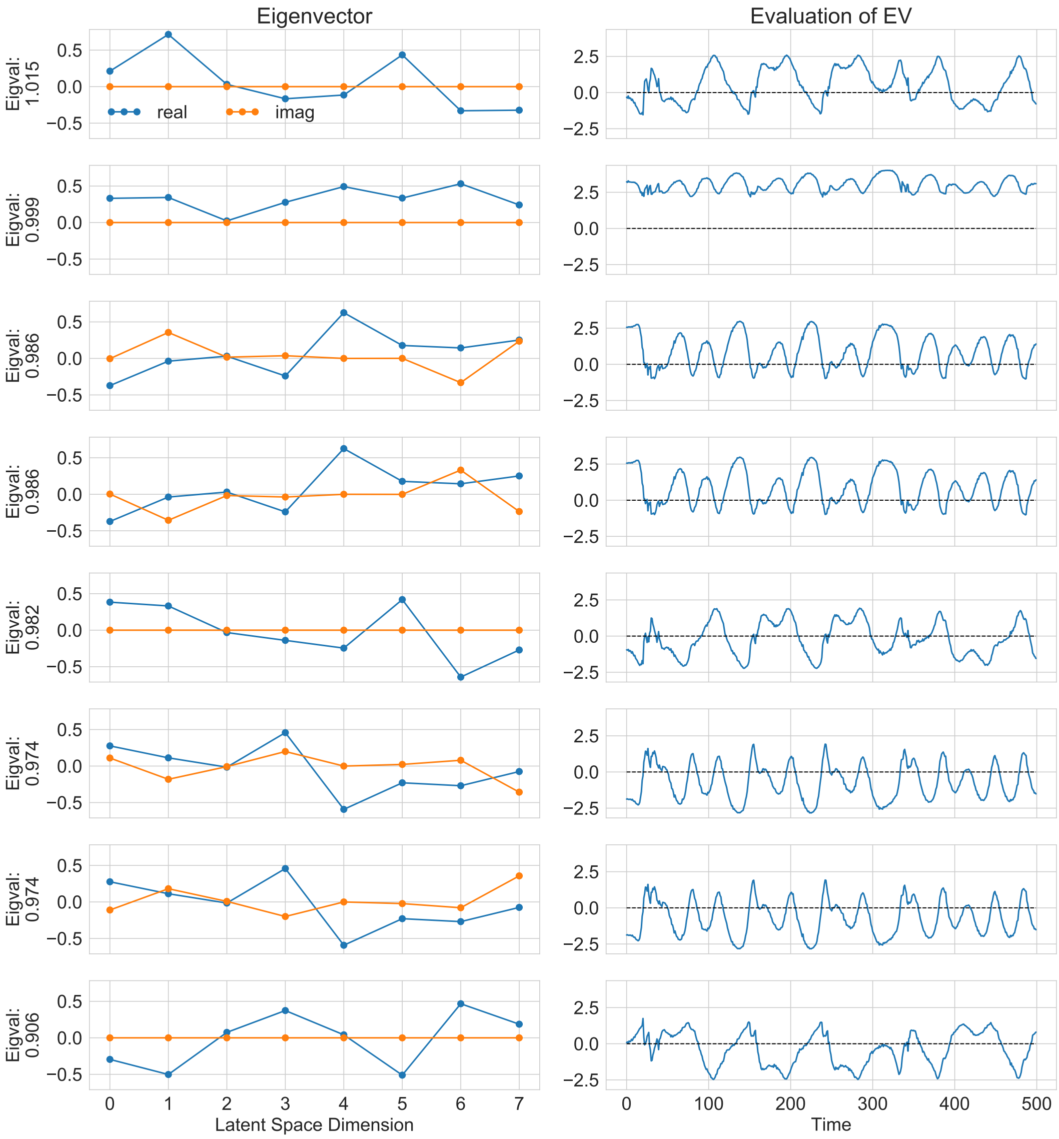


Figure 32: Eigenvalue Decomposition of the Lorenz System

Eigenvectors (blue: real part, orange: imaginary part) and their evaluation on the latent space representation of the Lorenz system (red, only real part shown).

of the eigenvectors on the latent space trajectory, each singular vector evaluation results in a unique temporal behavior with recurring patterns.

To validate the results from the decompositions, Figure 34 compares the results of the EVD and SVD with the original coordinates. Since the here used videos are a 2D projection of the Lorenz system, the exact position is known from the simulation. The original trajectory of X-Y-coordinates can be com-

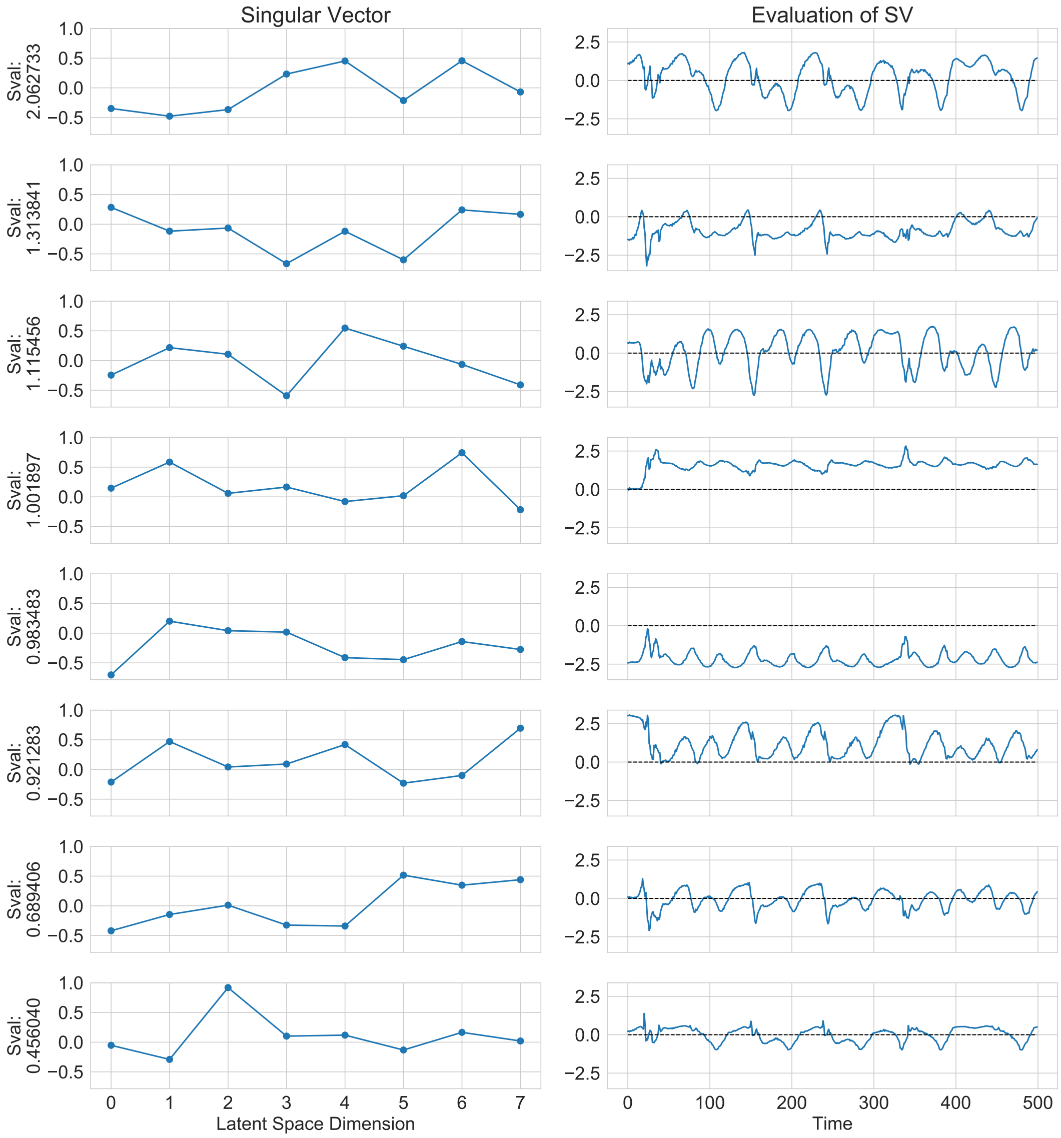


Figure 33: Singular Value Decomposition of the Lorenz System

Left and right singular vectors (left and center column, respectively) and the evaluation on the latent space representation of the Lorenz system (red).

pared with the trajectories in latent space as well as in the transformed spaces from the EVD and SVD. The absolute Pearson correlation coefficient is used, since a strong negative correlation may reveal the original coordinates as well, just with a different sign.

A brief visual analysis reveals that the five most correlating trajectories, in fact, resemble the X-coordinate. All shown trajectories have sharp spikes around time point 20, which are not present in the original trajectory. The horizontal position correlates most with EV-7, the evaluation of the eigenvector with the lowest eigenvalue (r=0.856). The dominant eigenvector generates the trajectory with the second-highest correlation (EV-0, r=-0.842). It also correlates strongly with SV-7 (r=0.823), which originates from the singular vector with the lowest singular value. The trajectory from the dominant singular vector (SV-0) shows a lower correlation of r=0.704 (see Appendix). The strongest correlation between the horizontal position and a latent space dimension is given by LS-1 (r=0.796).

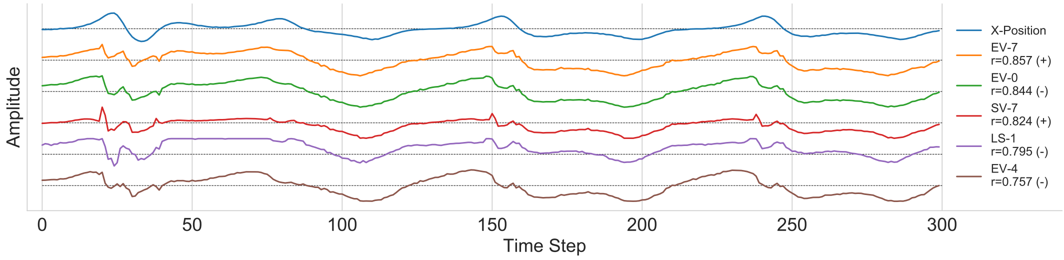


Figure 34: Comparison of Decomposition Methods with the Original Data

Original horizontal coordinate (X) of the Lorenz system and the evaluations of single eigenvectors and singular vectors, sorted by absolute correlation with the original coordinate (Top 5 only). Negative correlating trajectories have been flipped for a better visual comparison. Positive and negative correlations are indicated by the respective symbols behind the correlation coefficient. The full comparison can be found in the Appendix.

4.4.2 Koopman-VAE reveals Dynamics in FOI Data

The Koopman-VAE model that is trained on the FOI data is analyzed analogously to the model trained on the Lorenz system. Since the underlying dynamics are unknown, there are no original coordinates to compare.

The trained Koopman-VAE model reconstructs the images in comparable quality as the simple VAE model from the previous chapters. The average SSIM and pixel-wise correlation are 0.94 and 0.92, respectively. Although the Koopman-VAE has a more complex architecture compared to the initial VAE, the reconstructions are visually as good as in the second chapters. Thus, a comparative figure like Figure 11 and Figure 29 is omitted here for brevity.

Figure 35 shows the eigenvalues and the singular values of the linear operator trained on the FOI data. The lowest eigenvalue is 0.777, the largest is the complex conjugate pair 1.007+/-0.002i. In total there are only 8 real-valued eigenvalues and 21 complex conjugate pairs of eigenvalues. The majority of eigenvalues are close to an absolute value of 1, only the last seven are a little distant to the unit circle. The largest gap is between the second-lowest and lowest eigenvalue (0.823 and 0.777, respectively). The large number of complex eigenvalues indicates non-reversibility.

The singular values range between 2.130 and 0.409. Both, the largest and the lowest singular value are separated by a gap from the remaining 48 values. The latter form a slowly decaying plateau between 1.259 and 0.659, that shows no significant gaps.

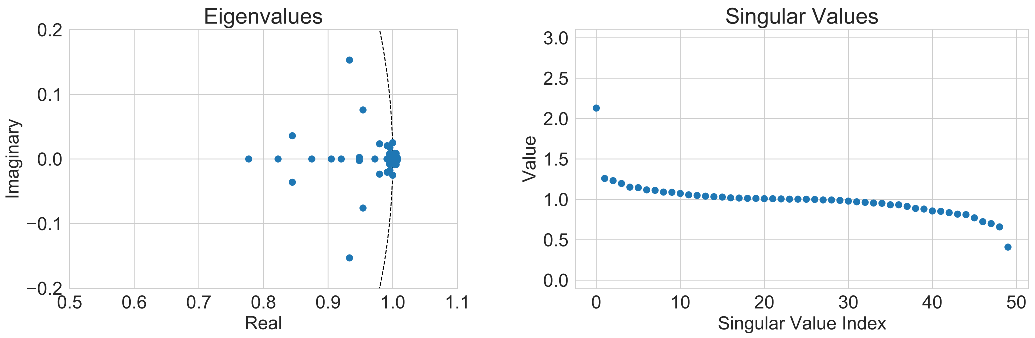


Figure 35: Eigen- and Singular Values for the FOI Data

Eigenvalues (left) and Singular Values of the linear operator that was trained on the FOI data

Figure 36 shows selected eigenvectors and their evaluation on the latent space representations of the FOI data, colored by the Psoriasis label. The eigenvectors are selected by their eigenvalues. The five highest and the five lowest are shown for brevity, a full overview can be found in the Appendix. The videos have been cropped in time for better comparison. Each video starts with the first frame where the ICG is present and ends after 200 seconds. The left part of Figure 36 shows that many of the eigenvectors have only a few distinct peaks in single dimensions, with the remaining dimensions close to zero (e.g., dimension 31 in the dominant eigenvector). The right part of Figure 36 shows that all evaluations show most of their activity in the first 60 seconds. Later time points are slowly and smoothly converging to zero. Two types of temporal behavior can be identified. The first pattern characterizes the eigenvectors with high eigenvalues. Starting at zero, the amplitude of the trajectories quickly increases in either a negative or positive direction and then slowly converges back to zero. The second pattern can be seen at the lowest eigenvalues. Here the amplitude initially diverges from zero, subsequently returns and overshoots zero, and then slowly converges back to zero.

When looking at the trajectories of the individual patients, there are only a few patients that exhibit outlying dynamics, which do not match the general patterns. Additionally, there are no visually significant differences between patients of the PsV and PsA subgroups. Generally, the difference is the largest during the early enhancement phase, approximately until time point 40. The average trajectories for both subgroups differ the most in EV-49 (Figure 38, top).

Figure 37 shows the results of the singular value decomposition. Selected left and right singular vectors are shown with their evaluation on the latent space representations. The singular vectors are selected by their singular values. As with the results from the EVD, only the five highest and the five lowest are shown, and the full overview is in the Appendix.

Unlike the eigenvectors, only a few singular vectors are characterized by distinct peaks. As seen in the singular vector evaluations of the Lorenz system, a variety of trajectory patterns can be seen in contrast to the eigenvector evaluations. The trajectory patterns of the eigenvector evaluations can be found here again, however, in different variations. SV-0 trajectories initially diverge from zero with a high amplitude until time point 10. Subsequently, the amplitude is decaying exponentially.

The trajectories of each singular vector are generally less homogeneous and show more outliers than those that are generated by the eigenvectors. While the majority of trajectories appear smooth, some trajectories demonstrate rippled temporal behavior. The differences between PsV and PsA are even smaller, compared to the eigenvector evaluations. The visually most significant difference of the average trajectories can be found in SV-47 (Figure 38, bottom), where PsV patients have, on average, a lower amplitude.

As already introduced in this section, a ground truth for the underlying dynamics of the FOI data is not available. Hence, the classification performance is evaluated using the transformations above. The same clinical parameters are used as in the previous chapter. Figure 39 shows the classification accuracies based on the trajectories in latent space and their transformations through eigenvectors and singular vectors these clinical parameters. There are no significant changes except for the label “Sex”, where the classification based on SV trajectories leads to better accuracies. The eigenvector evaluation leads to lower accuracies. At the time point of maximum accuracy based on LS trajectories (87.80%, SD: 1.35), the accuracies based on SV and EV trajectories are 89.35% (SD: 1.24) and 79.20% (SD: 2.06), respectively. However, the classification of PS, RA, and DAS28 based on EV trajectories shows marginally better accuracies at later time points of the FOI videos.

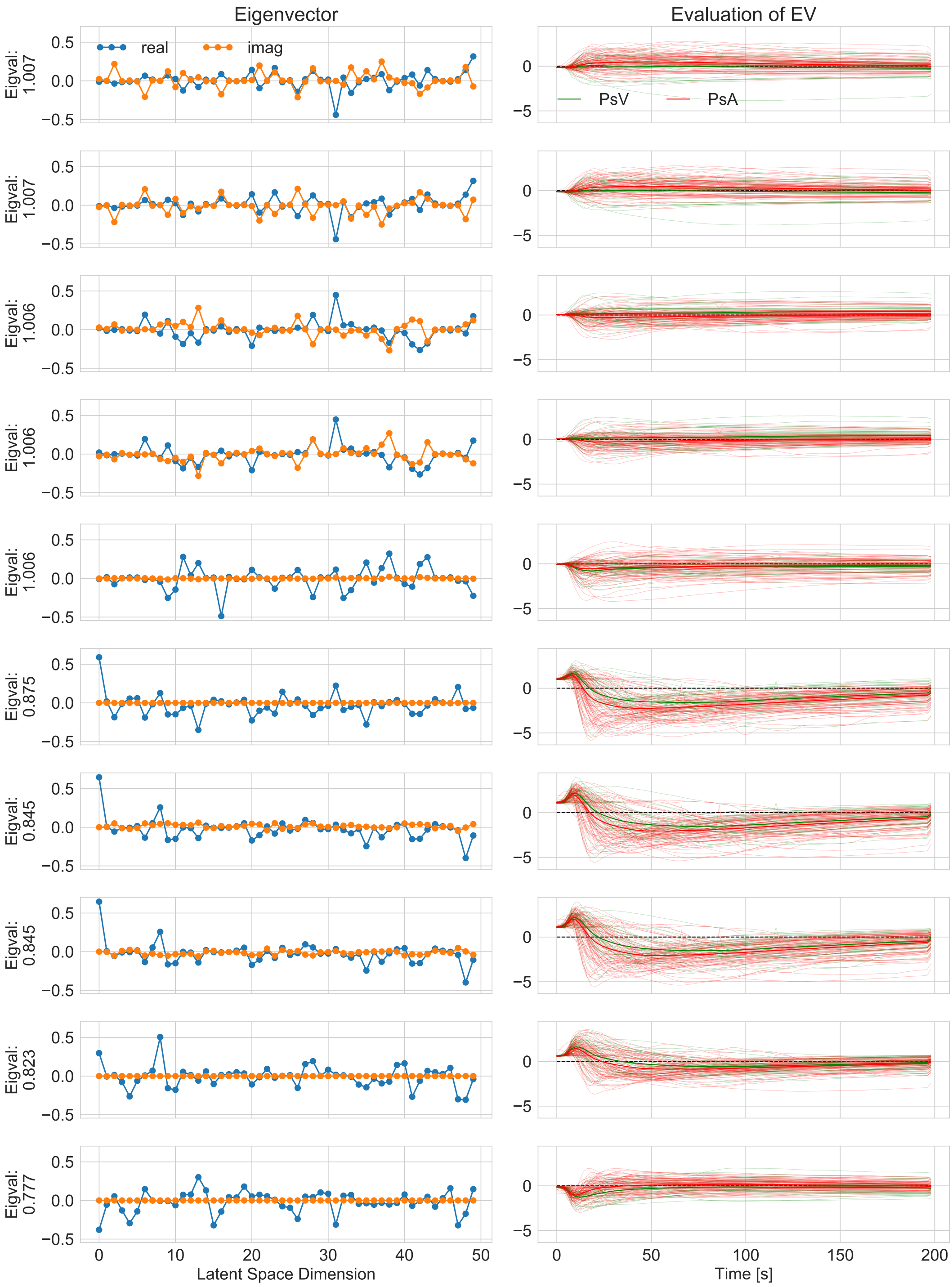


Figure 36: Eigenvectors and their evaluation on the Latent Space trajectories

Eigenvectors (left) and their evaluation on the latent space trajectories (right). Shown are only the eigenvectors with the five largest and the five smallest eigenvalues with their real and imaginary part (blue, orange). The evaluations are colored by the Psoriasis group (PsV: green, PsA: red). The averages for these groups are shown in dashed lines.

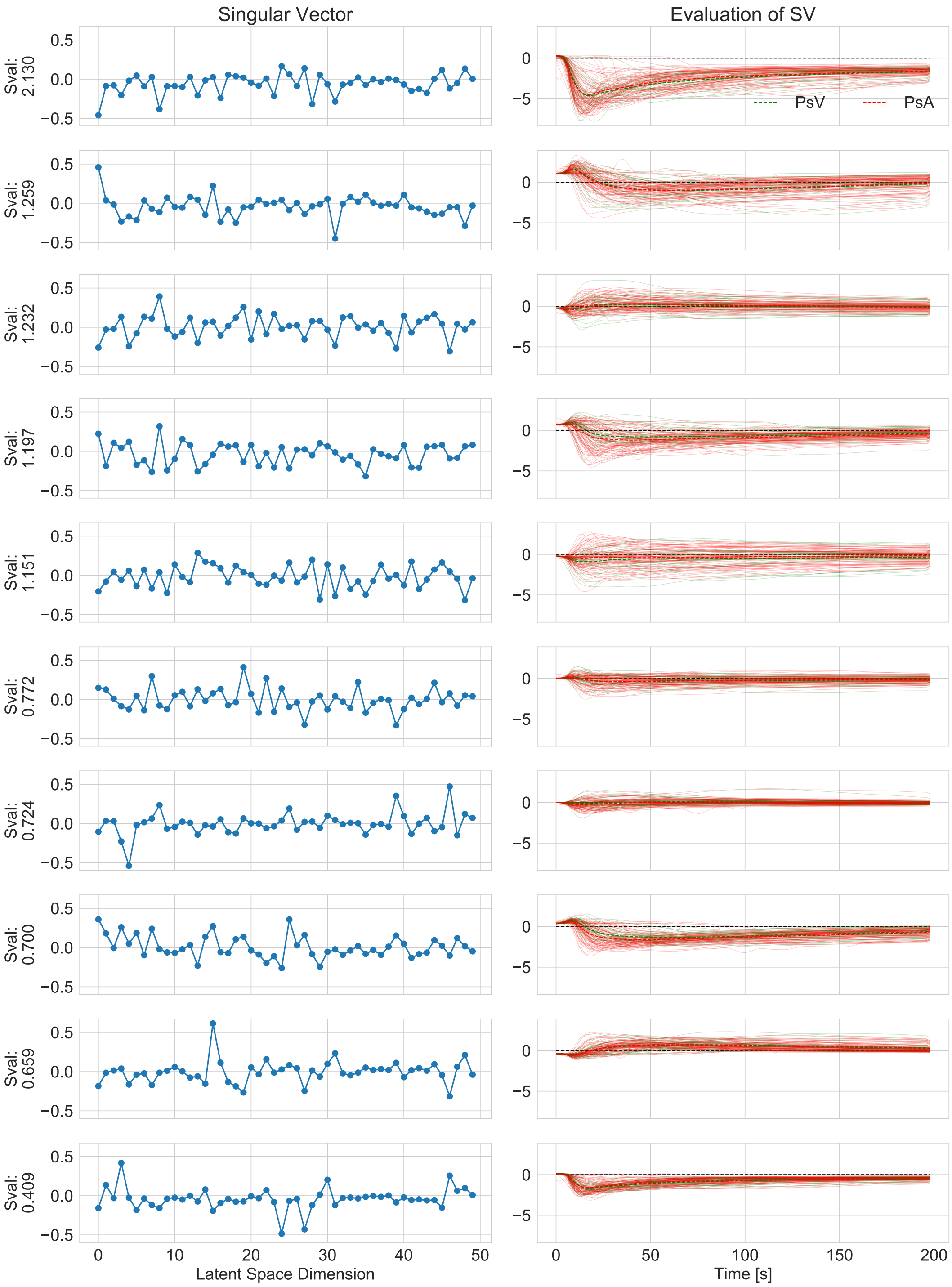


Figure 37: Results of the SVD with FOI Data

Left and right singular vectors (left, center), and the evaluation of the left singular vectors on the latent space trajectories (right). Shown are only the singular vectors with the five largest and the five smallest singular. The evaluations are colored by the Psoriasis group (PsV: green, PsA: red). The averages for these groups are shown in dashed lines.

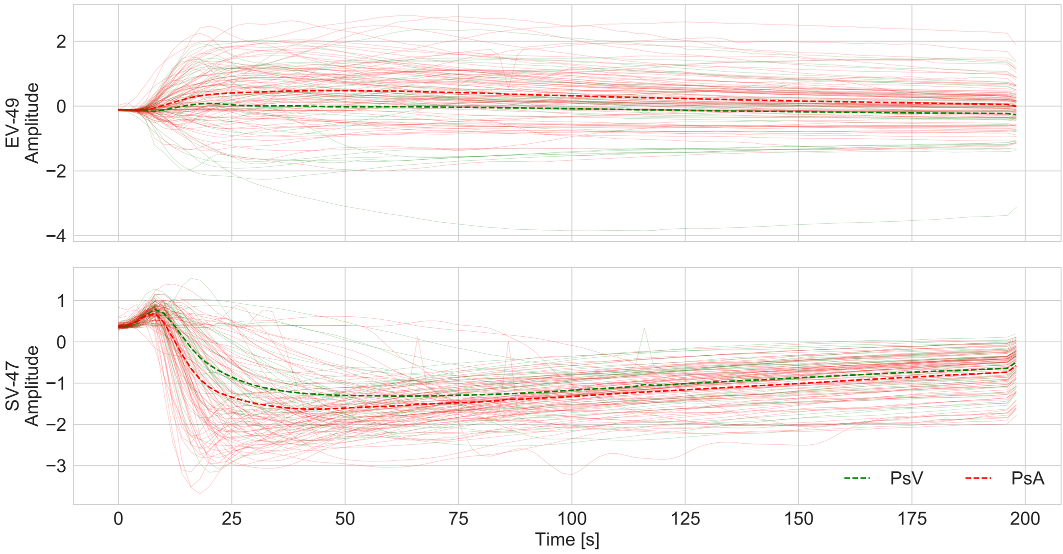


Figure 38: Evalutaion of EV-49 and SV-47 on the FOI Data

The evaluation of the least dominant eigenvector (EV-49, top) and the third least dominant singular vector (SV-47, bottom) on the latent space trajectories. The evaluations are colored by the Psoriasis group (PsV: green, PsA: red). The averages for these groups are shown in dashed lines.

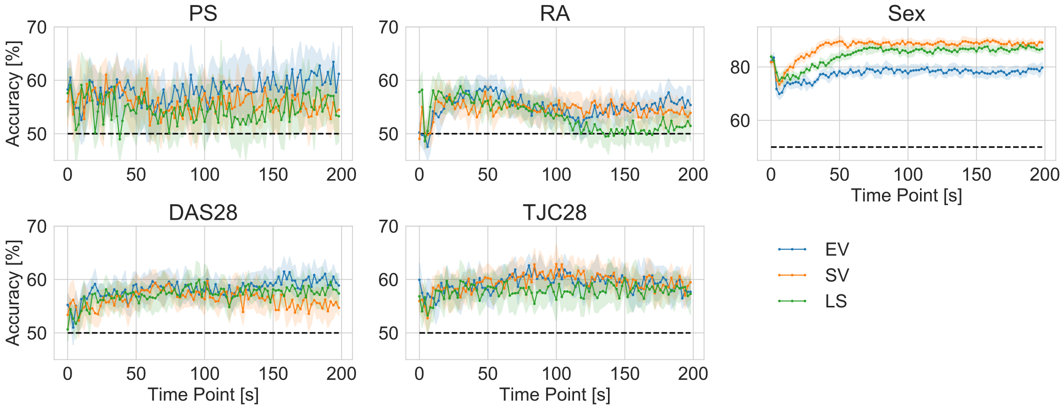


Figure 39: Classification on EV, SVD, LS

Average classification accuracies for five clinical parameters based on LS, EV, SV trajectories. Shaded areas represent standard deviation. Note: The Y-axis of the Sex plot is scaled differently.

4.5 Discussion

This chapter has demonstrated the capabilities of the unsupervised learning approach to reveal underlying dynamics in video data. The combination of a Variational Autoencoder (VAE) with a linear evolution operator proves to be feasible. The first experiment, which used the known Lorenz system to validate the results, successfully revealed the original coordinates of the system. Although the full potential of the Koopman-VAE model on the Lorenz system was not aimed to achieve here, the results seem plausible. Further optimization of the approach is necessary to investigate the full potential. The dependency of hyperparameters like the latent space dimensionality has not been investigated and may improve the results. The lack of reconstruction quality at time points of high velocities of the trajectory may come from underrepresentation since the moving circle spends less time at these positions.

The dominant eigenvector of the linear operator is slightly larger than 1. This may originate from numerical artifacts. The operator was found by optimizing a large neural network with three different loss terms. Hence, it may not fulfill the best numerical requirements. For example, there are negative entries, and neither the rows nor the columns have unit norm. Improving the network architecture, loss function may help as well as incorporating constraints.

The simple computation of the correlation coefficient is not the best way to discover correlations in time series, as it does not account for lags and other effects. However, given the low dimensionality, it is sufficient to sort the trajectories for visual comparison.

In summary, the FOI-Experiment demonstrated plausible results. The proposed Koopman-VAE model successfully learned and revealed the underlying dynamics of the FOI imaging application. Although the dynamical system and the resulting trajectories have a substantially less complex temporal behavior compared to the Lorenz system, the results confirm the general applicability of the approach, which is purely data-driven and without image analysis. A detailed investigation of the approach's capabilities to reveal the underlying dynamics of the FOI data is still necessary. Like the experiment with the Lorenz system, numerical artifacts may affect the quality of the analyses. Further optimization of the hyperparameters and the neural network are necessary to reveal the full potential of this approach, which was not the scope here.

The approach revealed that there are also differences between subgroups, although they are not significant. Especially the SVD of the linear operator and the transformation with singular vectors showed potential to reveal outliers. Trajectories that demonstrated rippled temporal behavior in SV-3 in contrast to the smooth majority of trajectories were later confirmed by manual inspection to show pulsating signal intensities in the thumbs. Trajectories, where the amplitude increased with a large delay, appeared to have a delayed injection of the contrast fluid. However, a confirmation of the reasons for the deviating behavior of the outliers was not possible due to missing documentation.

Due to the missing ground truth, it is not possible to compare the transformed trajectories with the actual underlying dynamics. Hence, it is not possible to correlate these trajectories comparable to the first experiment. Future experiments should at least incorporate two subgroups with a known difference in their dynamical behavior. These experiments can potentially leverage the interpretation of the eigenvectors and singular vectors, which is still outstanding. An intuitive way for the interpretation of these vectors is the transformation back into the image space using the VAE's decoder. However, it is not guaranteed that these vectors correspond to valid latent space representations of FOI images.

The classification based on the transformed latent space trajectories did not lead to clinically relevant improvements. As already mentioned in the previous chapter, the insufficient quality of labels may still play a significant role. Furthermore, it is unclear to which degree the diseases differ in their dynamics. The only significant change can be observed at the classification of the patient's sex. Here, the SVD improved the classification accuracy slightly. This is unsurprising, as a principal component analysis (PCA) is known to improve classification accuracies in some cases. A PCA is, in principle, an SVD of the data. The difference here is that the SVD is performed on the linear operator, not on the latent space trajectories. A speculative guess for the decreased performance of the EV trajectories is that the EVD contains mostly information about the dynamics. This is supported by the high correlation of EV trajectories in the first experiment (Figure 34). If the main difference between female and male patients is the size of the hands, which is constant during the imaging procedure, rather than the dynamics, this would explain the decreased classification performance.

Future experiments may aim to train separate operators for the respective subgroups. A comparison of eigenvalues and eigenvectors, as well as singular values and vectors, may reveal the dynamical differences between those groups. However, this would probably require more data of better quality.

5

Summary, Discussion & Outlook

5.1 Summary

This thesis has introduced a new data-driven analysis pipeline for disease assessment in time-resolved medical image data. The here presented approach focuses on a Variational Autoencoder, which reduces the dimensionality of the present data set by learning a latent space. This learned latent space enables and facilitates subsequent analyses. Three of these possible analyses have been demonstrated here.

First, subgroup analysis and data exploration have been proven to give feasible results. The average latent space representation for clinically relevant subgroups can be computed and decoded back into the image space. These feature-wise averages are superior over the pixel-wise averages and result in visually pleasing, interpretable images. The application of UMAP enables further investigation of the latent space and reveals clusters of similar patients. When comparing the latent space averages of subgroups, the differing dimensions can explain the difference in the image space. Varying a single dimension of a latent space representation and observing the effect on the decoded image can explain the feature that this dimension encodes.

Second, classification with limited data-availability has successfully been demonstrated. Several approaches have been compared. The two-step approach, that uses a Random Forest classifier on the latent space representations of the trained VAE, has been proven to be very effective on the synthetic FOI data set, and the actual FOI data with clinical target values. Additionally, this framework allows for explainable classifications via Occlusion Maps.

Third, a linear approximation of the underlying dynamics of the time-resolved data has been demonstrated. Two copies of the VAE have been joined with an additional layer that represents an approximation of a Koopman operator. Using the known Lorenz system, the Koopman-VAE has successfully revealed the underlying dynamics. Furthermore, the approach has been used to reveal the dynamics of time-resolved FOI data. This has enabled the discovery of patients with deviating temporal behavior.

Generally, this thesis has investigated the conditions, which lead to a well-trained model. The employment of a perceptual loss function improves the reconstruction quality significantly. It results in reconstructions that maintain the visual features of the original input data. The pixel-wise mean squared error leads to blurry reconstructions. The effect of weighing the Kullback-Leibler-Divergence in the global loss function has been evaluated.

The here proposed, novel approach has successfully been implemented as software prototype. This software is actively used by the project partner. It enables their researchers to conduct own machine learning experiments and to develope explainable classification models for phenotypes of further diseases. Therefore, new data is acquired by manual categorization. This is now possible with substantially lower effort, since only a small number of cases needs to be annotated. Preliminary results show that tendosynovitis can be detected with approximately 90% accuracy.

5.2 Discussion

The here presented analysis pipeline enables data exploration, explainable image classification, and approximation of the underlying dynamics. As introduced and elaborated in the individual chapters, each of these tasks has been solved or investigated before. However, this approach has investigated these tasks in the context of medical imaging and its accompanying problems. As stated in the previous chapters, these problems are mainly the limited data-availability, the insufficient quality of target values, and the absence of ground truth.

5.2.1 Reduction of Dimensionality enables previously infeasible Analyses

The key element of the here presented approach is the drastic reduction of dimensionality in combination with a weak reliance on labels. This is mainly realized via unsupervised learning of the VAE. Here, all available image information is used to learn a low-dimensional representation of the data, the latent space. With more than 200 thousand pixels, each image of the FOI data set contains a lot of information. But not the entire information is vital for the desired task. The VAE compresses the information by finding common features and encoding only the deviation of it. For example, each image shows two hands in a standardized arrangement. In the image, this is represented with a few thousand pixels. Whereas in the latent space, this can be encoded in only a few dimensions, that encode for the shape and size of the hands. Analogously, the signal distribution of the contrast fluid is stored very detailedly in the image. This could potentially be described with a few dimensions, that encode for the inflammatory joints or disease pattern.

Ideally, each dimension of the latent space encodes for a distinct feature. This disentanglement has not been fully achieved. But the different temporal variances in Figure 18 indicate, that only a few dimensions vary over the duration of the video and encode for the signal distribution, while the constant dimensions encode for the shape. To achieve the goal of disentanglement, the model can be extended to incorporate for prior knowledge. One extension can be the addition of a classification layer with sparsity promoting constraints or regularizations. In combinations with labels that represent pre-computed image features or measurements, this can enforce the VAE to encode these features in distinct dimensions of the latent space. An extension, which does not require additional labels, can be derived from the Koopman-VAE. Under the assumption, that shape and size of the hands do not vary between two time steps of a video, the two VAE copies can encode a common vector for these time-invariant features, and independent vectors for the signal distribution at t and $t + 1$. Similar approaches, that also use adversarial learning have been proposed recently [132].

The lack of disentangled dimensions may result in accurate average latent space representations. Here, the averages have been computed without regarding the manifold. Ideally, these computations incorporate Riemannian geometry. This problem also occurs in so-called shape spaces. Although these erroneous average latent space representations can be decoded and result in valid reconstructions, it is not guaranteed that these represent the correct feature-wise average image. The same problem applies to the eigenvectors and singular vectors, which thus cannot be decoded reliably.

5.2.2 Two-step Training simplifies explainable Classification under limited Data Availability

A characteristic problem of medical imaging data that occurs in each chapter is the insufficient quality of labels. This has been demonstrated successfully with the classification experiment using the synthetic FOI data set. The ideal condition of having accurate labels for two distinct classes has led to a well-performing classifier. However, under ideal conditions, the classification problem can be solved with less effort using a supervised classification network like the GoogLeNet or ResNet. The here presented approach can reveal the quality of the labels in early stages of the analysis, especially during the data exploration using UMAP.

The here presented approach has demonstrated to enable classification explanations. The explanations generated with LIME have not given fully interpretable results. In contrast, the Occlusion Maps have provided intuitive and interpretable explanations. This has been validated with the synthetic FOI data set, on which the inserted inflammation templates have been highlighted correctly and interpretable.

5.2.3 Underlying Dynamics can be revealed

As introduced in the previous chapter, there are more elaborate approaches to analyze and reveal underlying dynamics. However, it has been shown here, that a simple implementation of a linear operator can result in good approximations of the dynamics. The original coordinates of the Lorenz system have been revealed successfully. The FOI experiment lacks a ground truth for comparison, but it has been possible to identify cases with outlying temporal behavior. This has been possible through the decomposition of the linear operator into its eigenvectors and singular vectors. The approach further enables the comparison of the dynamics between subgroups. However, significant differences have not been revealed. Beyond the missing ground truth for the dynamics, the approach suffers from missing standardization of the data. As introduced before, the signal intensities of the FOI data are not normalized without a bias. Additionally, the temporal behavior is not normalized either. Besides the influences from the individual conditions of the patients, that are subject of this research, the influx and wash-out of the contrast fluid ICG is strongly dependent on body mass, heart rate, blood pressure, and non-physiological factors.

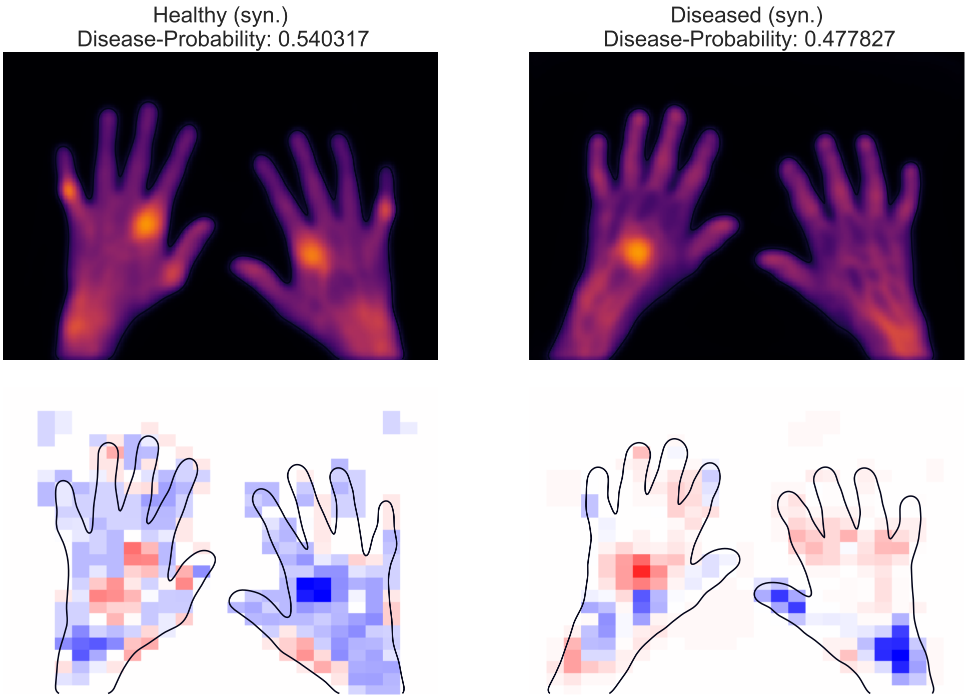
The here chosen implementation of the linear operator is adapted from the Koopman Operator Theory and approximated by minimizing the propagation error. When comparing the approach with recent research on the numerical approximation of Koopman Operators, it becomes clear that the here found operator may not fulfill all criteria of a Koopman Operator [133] [134]. Furthermore, the linear operator has the same dimensionality as the latent space, and an accurate modeling of the dynamics may require a higher dimensional operator.

5.3 Outlook

As already stated, the low-dimensional latent space representation enables a variety of further analyses. Future research can investigate the computation of disease networks. This would require a disentangled latent space, such that diseases-irrelevant information can be excluded. In combination with a reasonable distance metric, a network can be constructed that represents the variety and relations between the individual inflammatory diseases and their grades of severity. This network enables differential diagnoses. And in combination with longitudinal studies, such a network can enable an unbiased therapy response monitoring for individual patients. As stated in the individual discussions of the previous chapters, the performance can potentially be improved through hyperparameter optimization. This has been omitted to maintain comparable results. Similarily, a sensitivity analysis needs to be performed. Here, only the effect of the latent space size and the weight of the KL-loss have been investigated. Further robustness has been enforced through the addition of noise to the input data. Future work can further investigate the applicability of the approach on image data-driven in other domains. The VAE can be used on microscopy images of cancer cells to generate a phenotypic fingerprint of the cells, that describes the signal distribution of fluorescent markers. The effect on this fingerprint can then be observed in dependency of chemical treatment. Drug candidates that lead to similar phenotypes can be identified by clustering their phenotype in the latent space. Cell migration can be investigated with the proposed Koopman-VAE. Videos of cells migrating in a synthetical scratch wound assay are a common way to investigate angiogenesis and inter-cellular interactions [135] [136]. Different migration behaviors can potentially be identified.

Appendix

Synthetical Data Classification Failures



Healthy (syn.)
Disease-Probability: 0.540317

Diseased (syn.)
Disease-Probability: 0.477827

Figure 40: Two example cases of the synthetical FOI dataset with false prediction

Two examples from the synthetical FOI dataset (top) and corresponding Occlusion Map (bottom). The false-positive example (left) and the false-negative example both have probabilities for the “diseased” class near the 50% threshold.

Koopman-VAE

Full list of correlations (Lorenz)

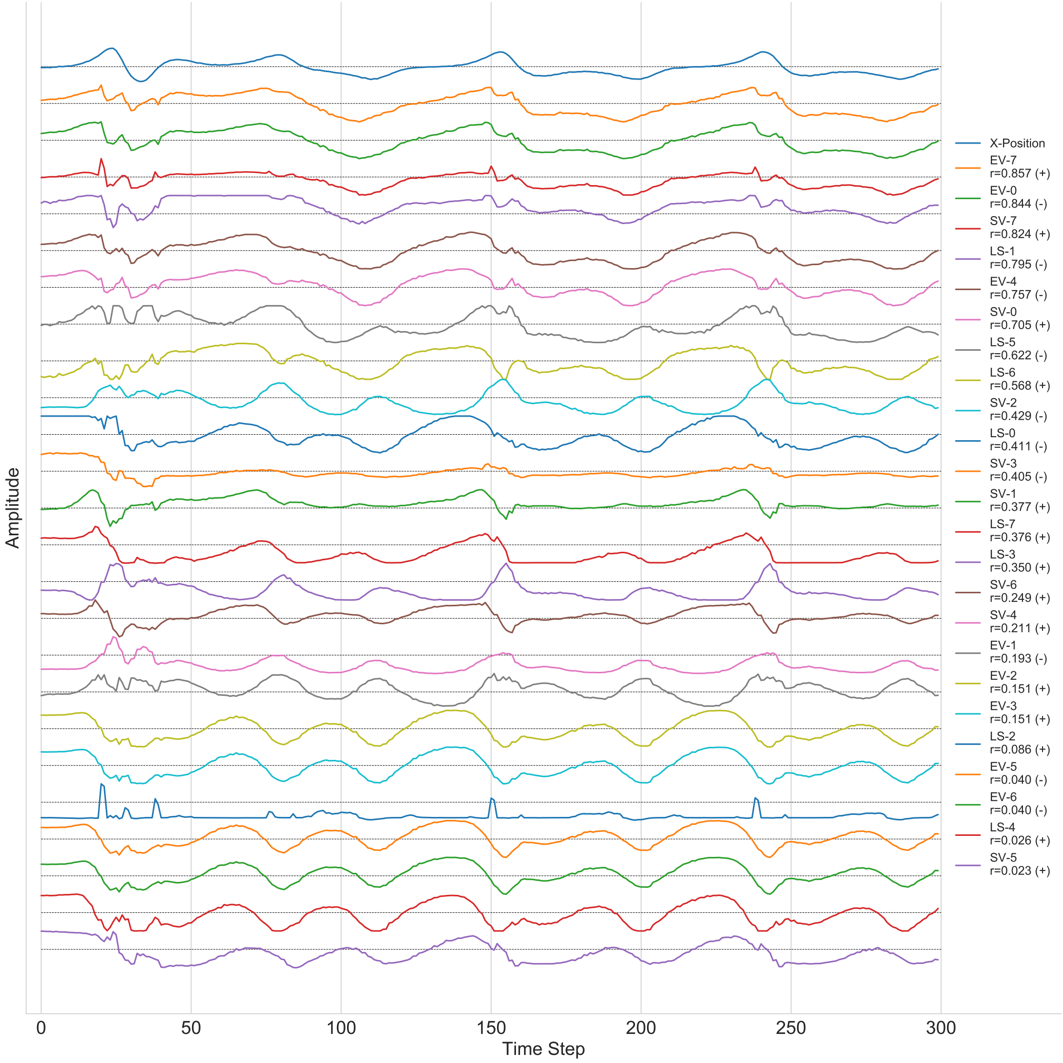


Figure 41: Full Comparison of Decomposition Methods with the Original Data (horizontal Component)

Original horizontal coordinate (X) of the Lorenz system and the evaluations of single eigenvectors and singular vectors, sorted by absolute correlation with the original coordinate. Negative correlating trajectories have been flipped for a better visual comparison.

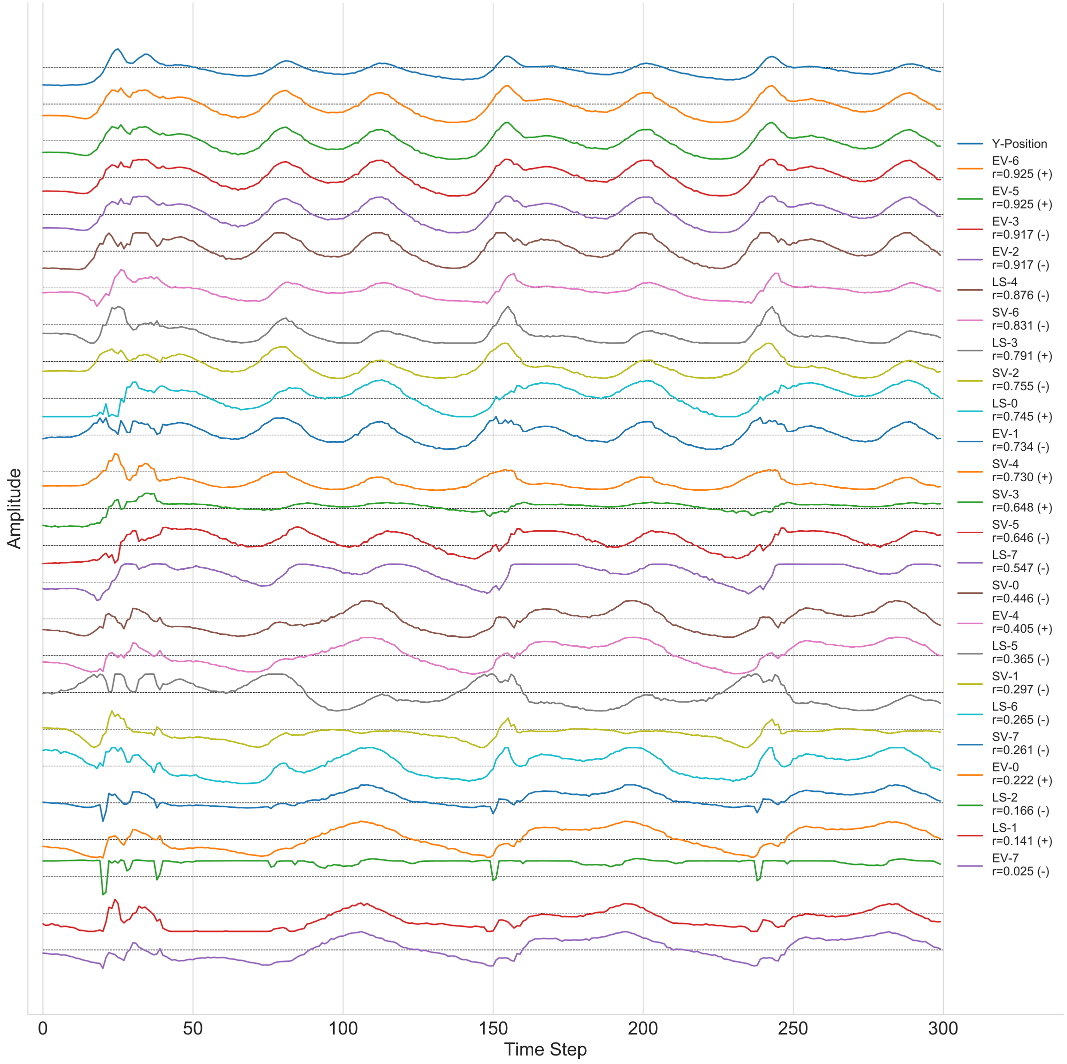
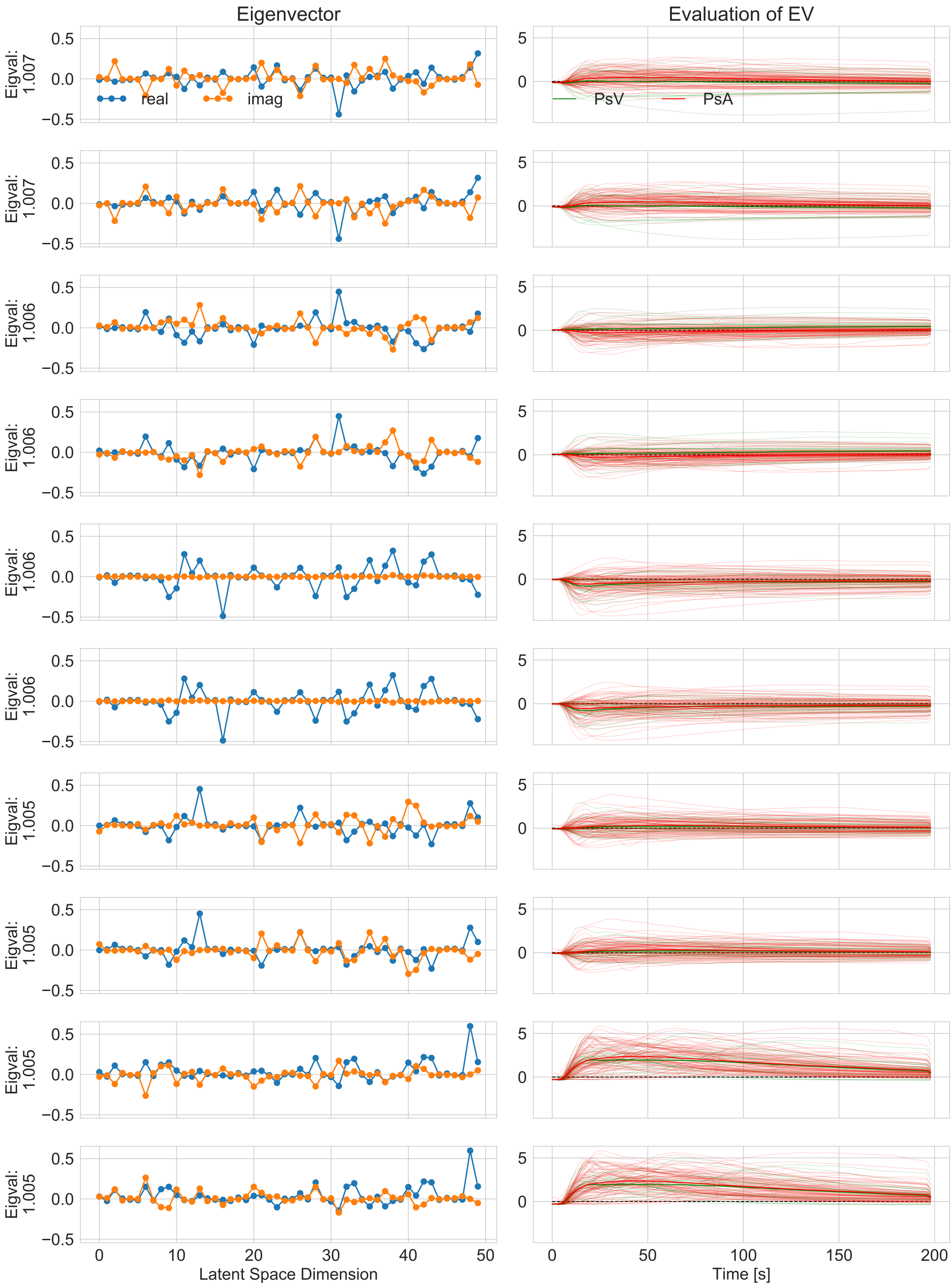
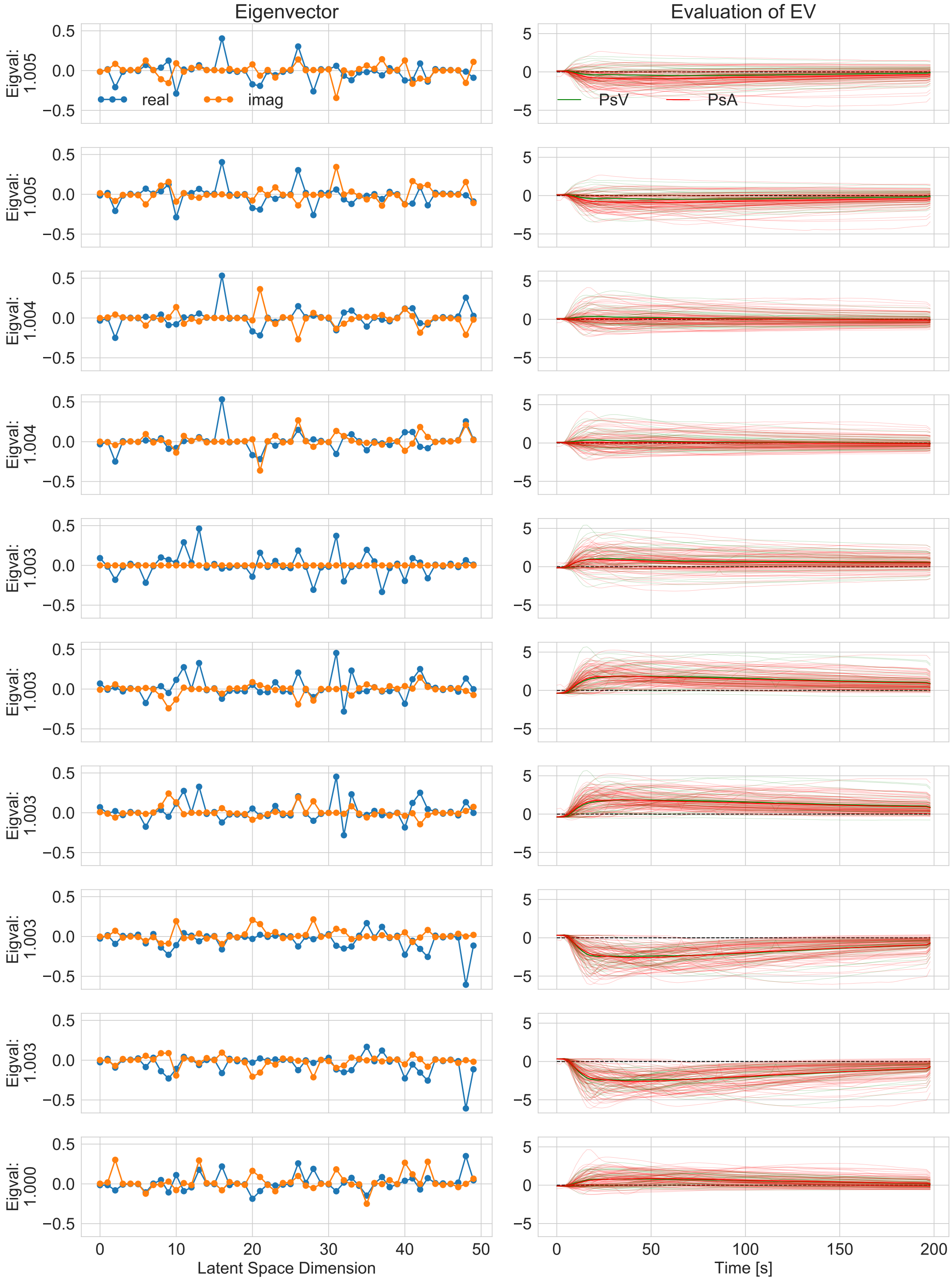
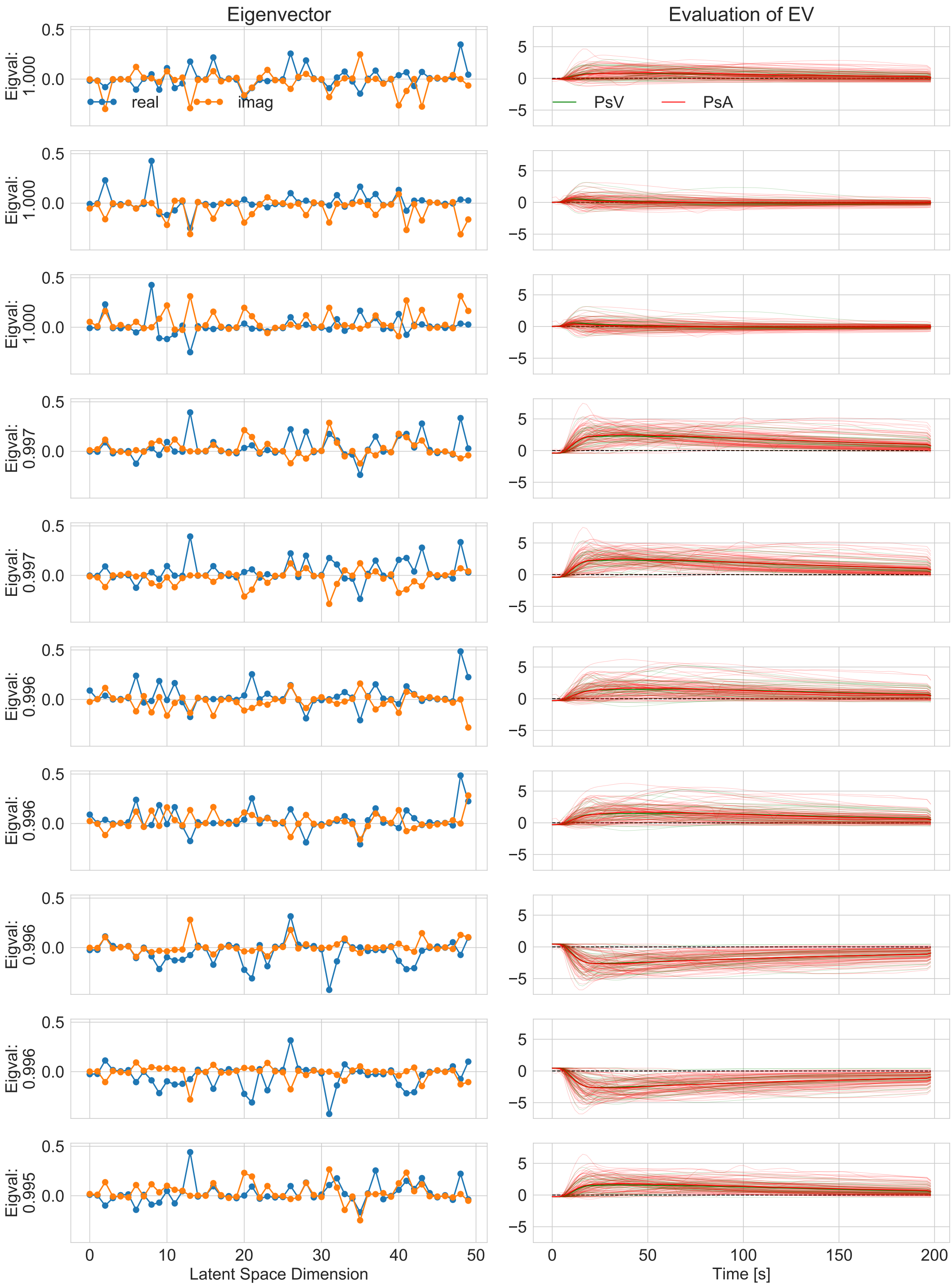


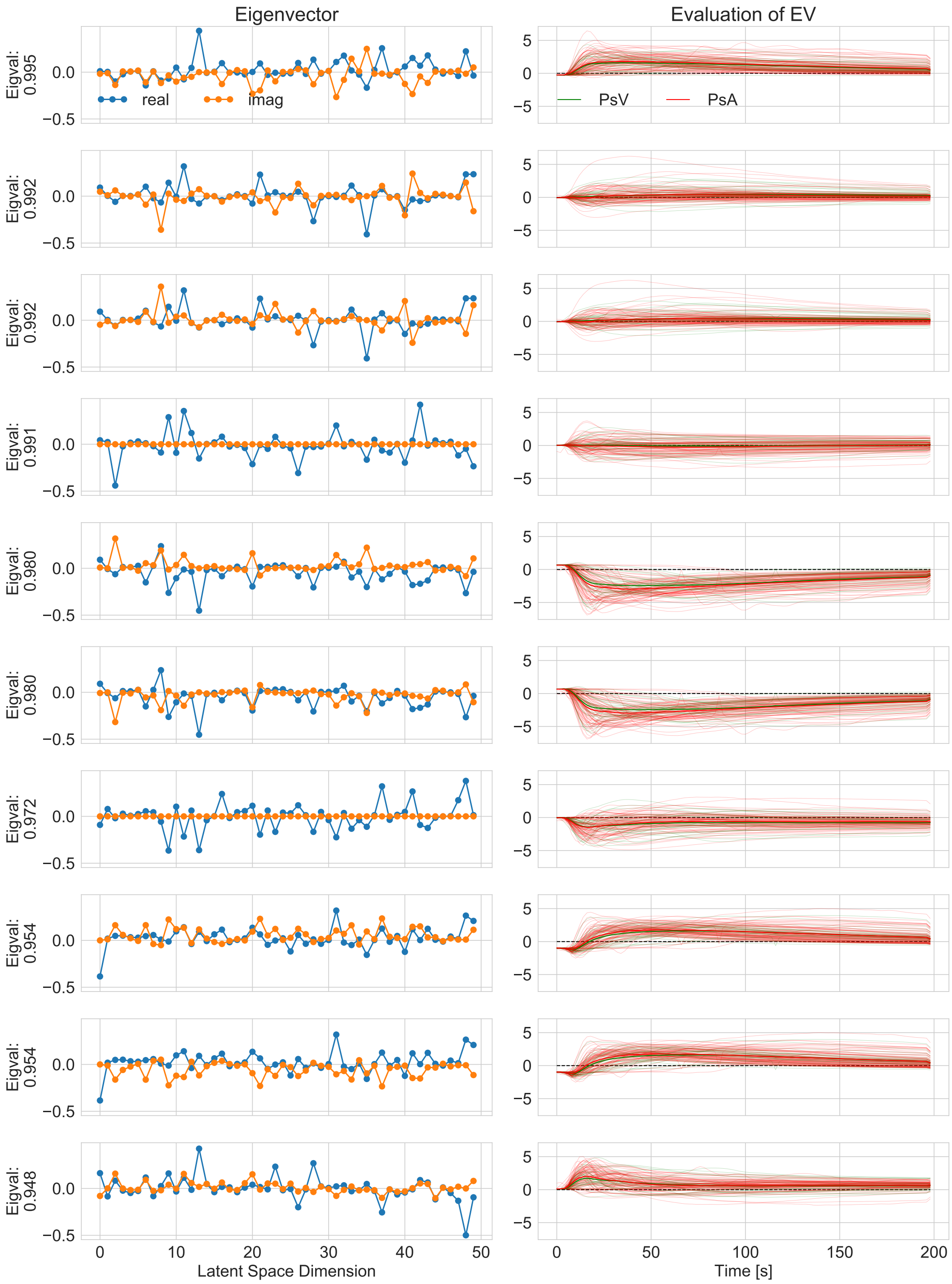
Figure 42: Full Comparison of Decomposition Methods with the Original Data (vertical Component) Original vertical coordinate (Y) of the Lorenz system and the evaluations of single eigenvectors and singular vectors, sorted by absolute correlation with the original coordinate. Negative correlating trajectories have been flipped for a better visual comparison.

Full dynamics of EVD




Eigenvector
Evaluation of EV
Eigval: 1.005
Eigval: 1.005
Eigval: 1.004
Eigval: 1.004
Eigval: 1.003
Eigval: 1.003
Eigval: 1.003
Eigval: 1.003
Eigval: 1.003
Eigval: 1.000
real
imag
PsV
PsA
0.5
0.0
−0.5
5
0
−5
0
10
20
30
40
50
Latent Space Dimension
0
50
100
150
200
Time [s]


Eigenvector
Evaluation of EV
Eigval: 1.000
Eigval: 1.000
Eigval: 1.000
Eigval: 0.997
Eigval: 0.997
Eigval: 0.996
Eigval: 0.996
Eigval: 0.996
Eigval: 0.996
Eigval: 0.995
real
imag
PsV
PsA
0.5
0.0
5
0
-5
0
10
20
30
40
50
Latent Space Dimension
0
50
100
150
200
Time [s]


Eigenvector
Evaluation of EV
Eigval: 0.995
Eigval: 0.992
Eigval: 0.992
Eigval: 0.991
Eigval: 0.980
Eigval: 0.980
Eigval: 0.972
Eigval: 0.954
Eigval: 0.954
Eigval: 0.948
real
imag
PsV
PsA
0.0
−0.5
5
0
−5
0
10
20
30
40
50
Latent Space Dimension
0
50
100
150
200
Time [s]

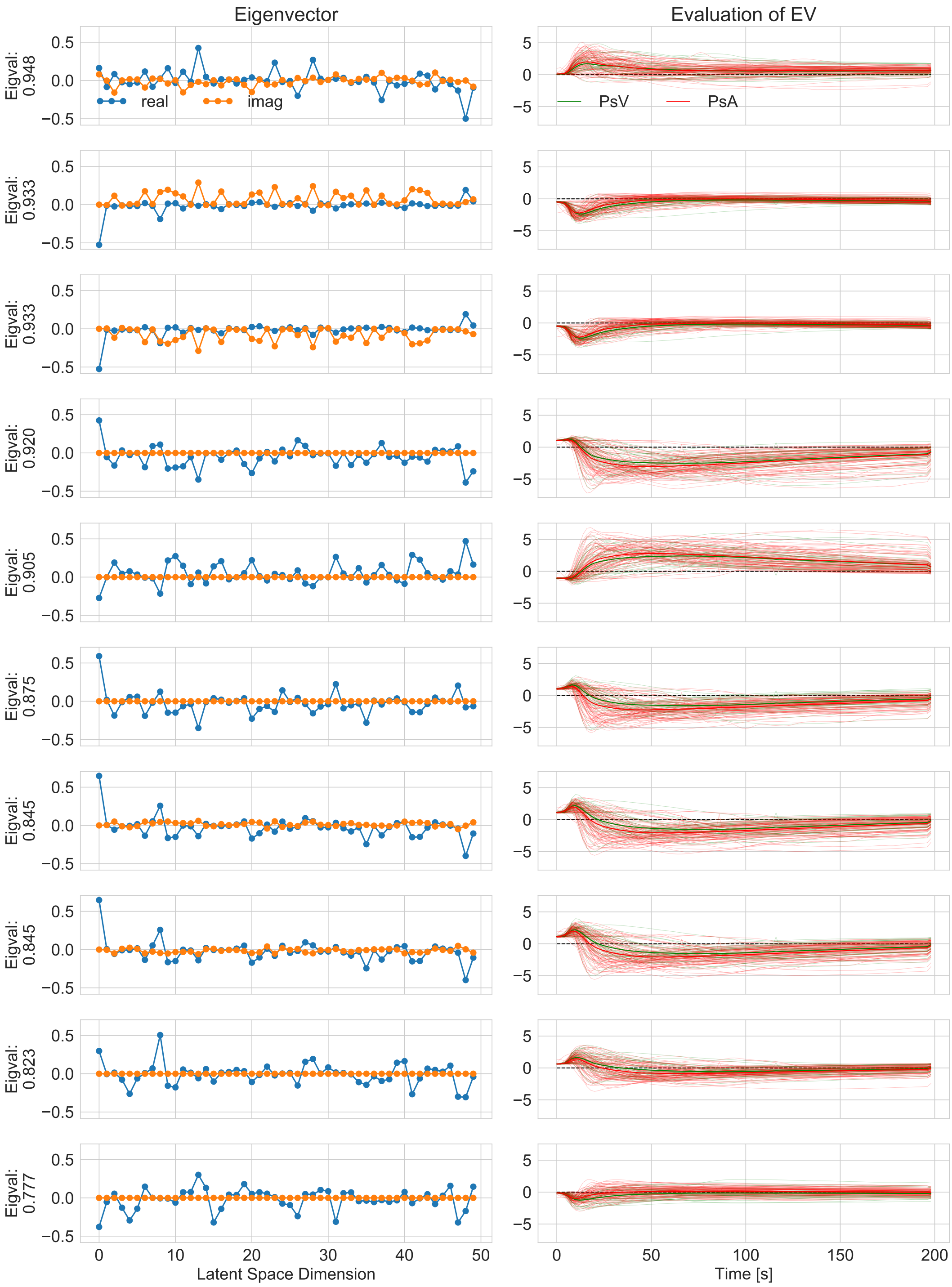
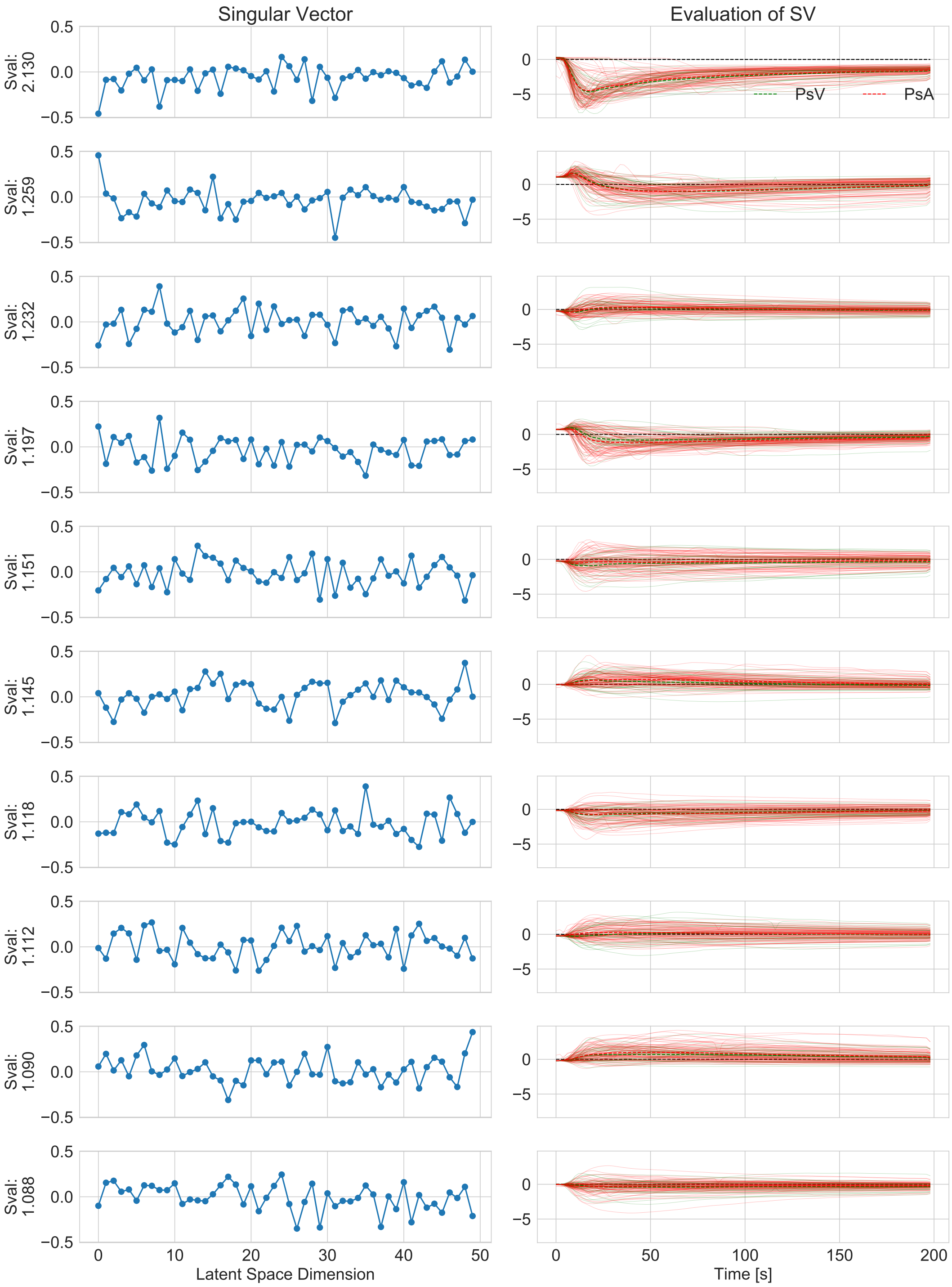
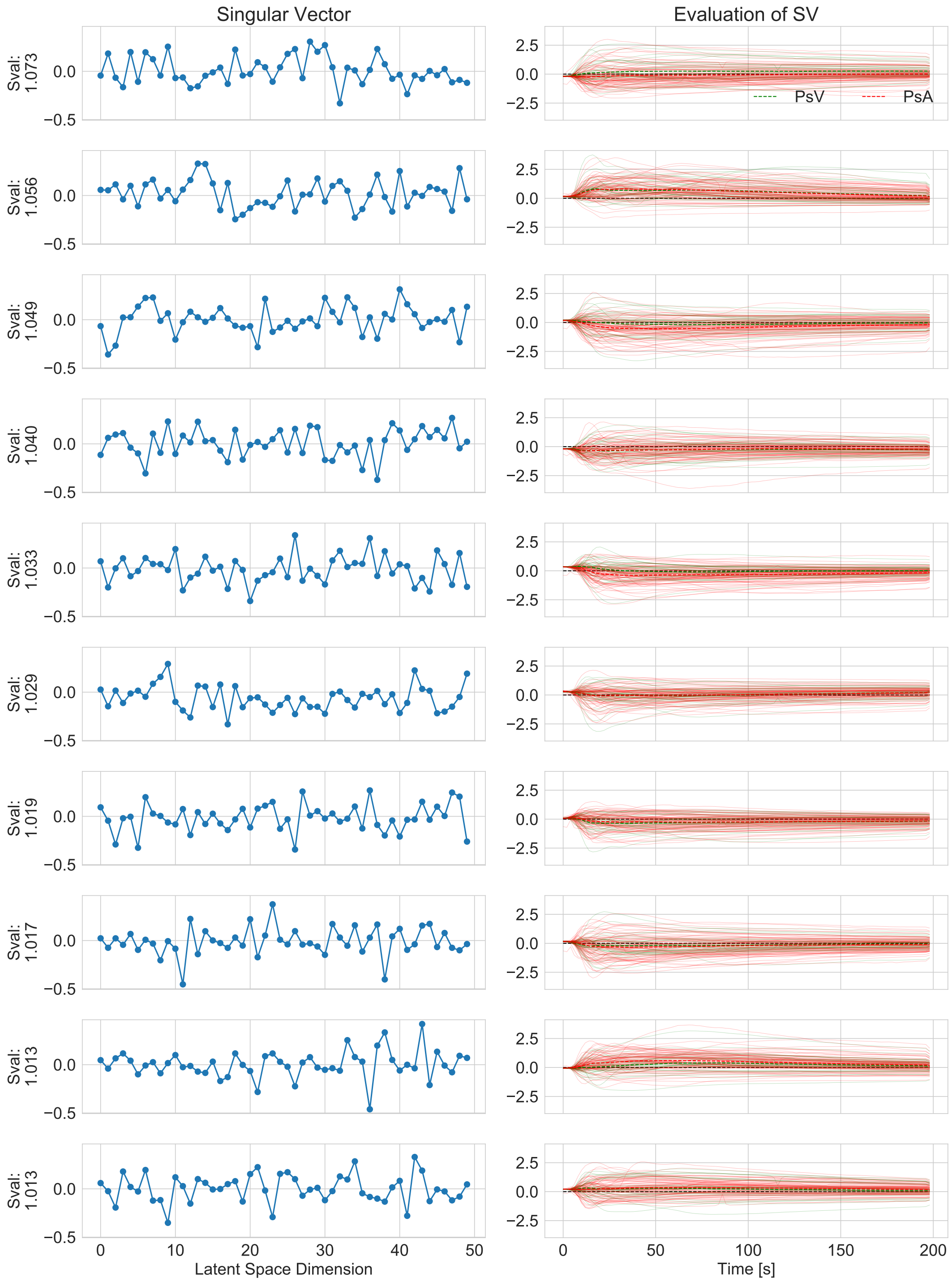


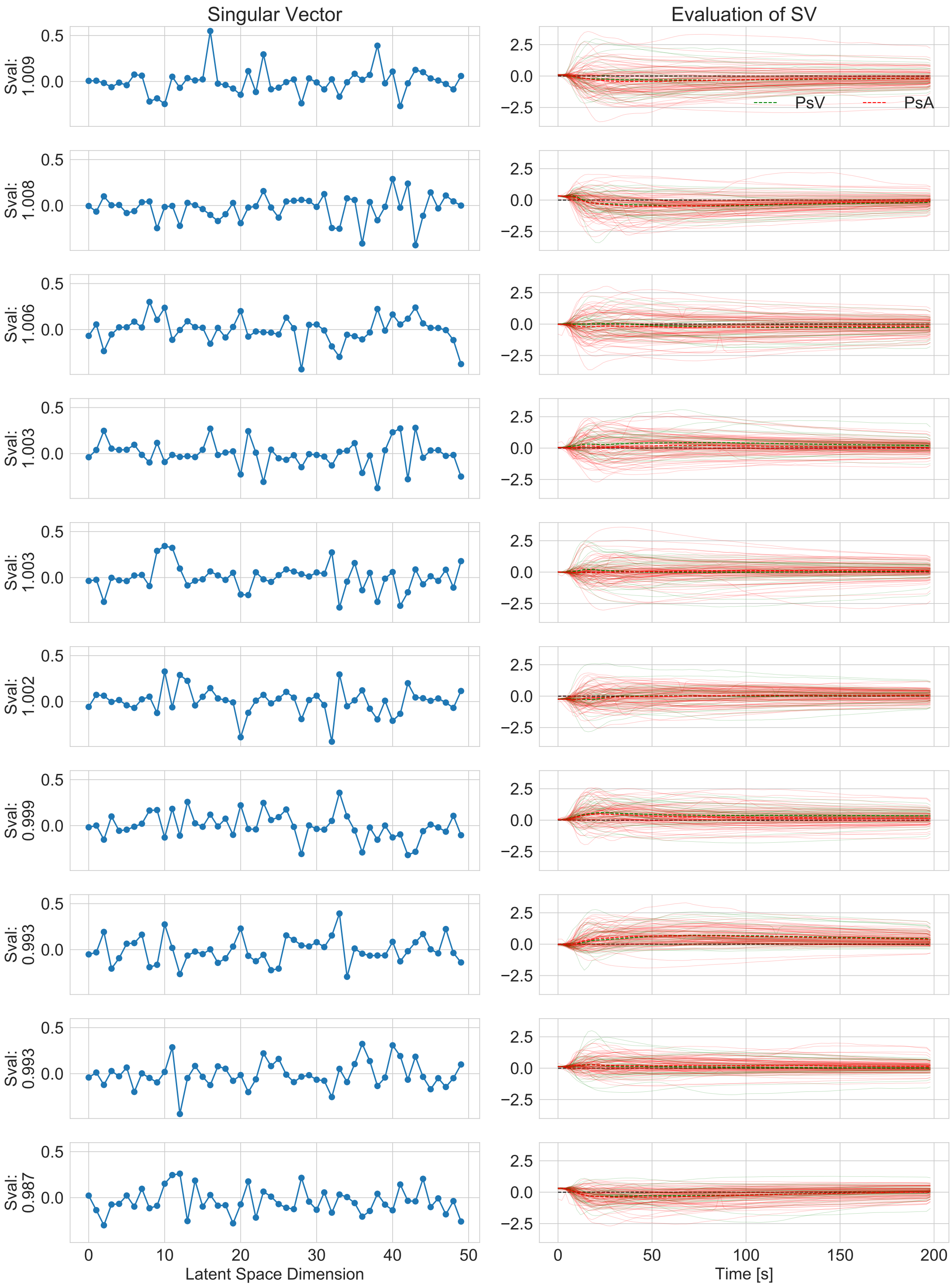
Figure 43: Eigenvectors and their evaluation on the Latent Space trajectories

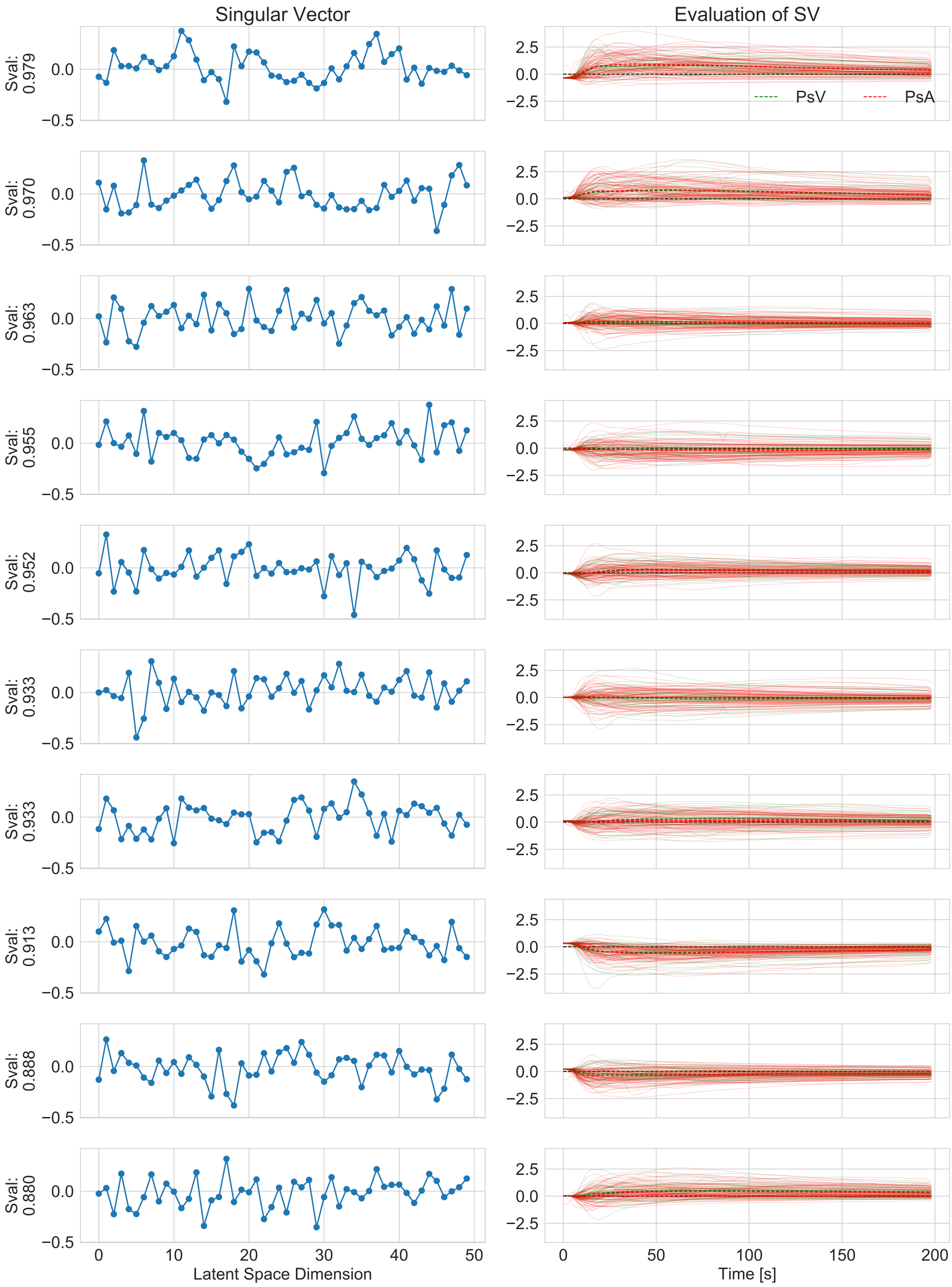
Eigenvectors (left) and their evaluation on the latent space trajectories (right). Shown are the eigenvectors with their real and imaginary part (blue, orange). The evaluations are colored by the Psoriasis group (PsV: green, PsA: red). The averages for these groups are shown in dashed lines.

Full dynamics of SVD




Singular Vector
Evaluation of SV
Sval: 1.073
Sval: 1.056
Sval: 1.049
Sval: 1.040
Sval: 1.033
Sval: 1.029
Sval: 1.019
Sval: 1.017
Sval: 1.013
Sval: 1.013
0.0
−0.5
2.5
0.0
−2.5
PsV
PsA
0
10
20
30
40
50
Latent Space Dimension
0
50
100
150
200
Time [s]


Singular Vector
Evaluation of SV
Sval:
1.009
Sval:
1.008
Sval:
1.006
Sval:
1.003
Sval:
1.003
Sval:
1.002
Sval:
0.999
Sval:
0.993
Sval:
0.993
Sval:
0.987
0.5
0.0
2.5
0.0
−2.5
PsV
PsA
0
10
20
30
40
50
Latent Space Dimension
0
50
100
150
200
Time [s]


Singular Vector
Evaluation of SV
Sval: 0.979
Sval: 0.970
Sval: 0.963
Sval: 0.955
Sval: 0.952
Sval: 0.933
Sval: 0.933
Sval: 0.913
Sval: 0.888
Sval: 0.880
0.0
−0.5
2.5
0.0
−2.5
PsV
PsA
0
10
20
30
40
50
Latent Space Dimension
0
50
100
150
200
Time [s]

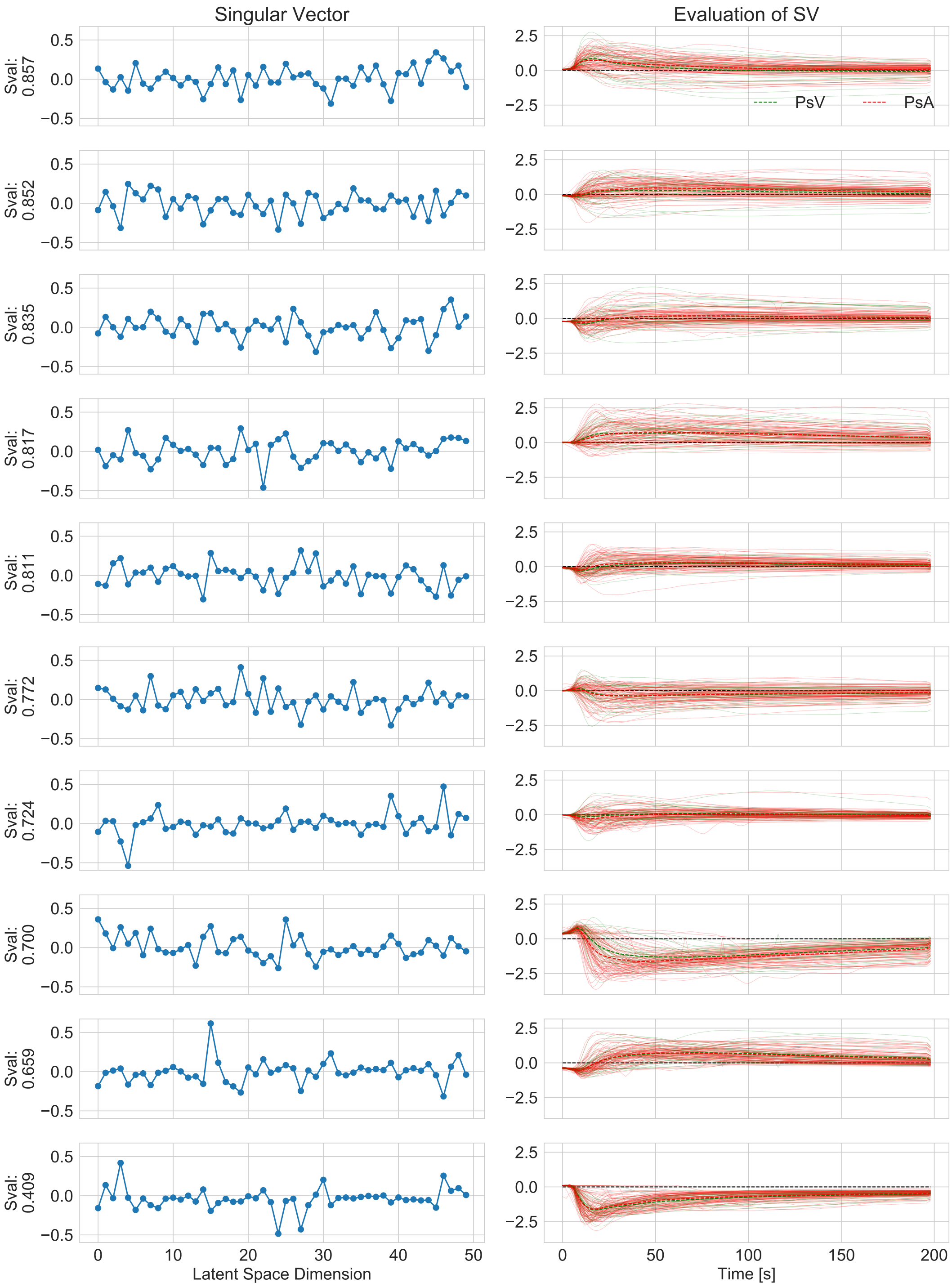


Figure 44: Singular Vectors and their evaluation on the Latent Space trajectories

Left and right singular vectors (left, center), and the evaluation of the left singular vectors on the latent space trajectories (right). The evaluations are colored by the Psoriasis group (PsV: green, PsA: red). The averages for these groups are shown in dashed lines.

Bibliography

[1] Kunio Doi. Diagnostic imaging over the last 50 years: Research and development in medical imaging science and technology. *Physics in Medicine and Biology*, 51(13), 2006.

[2] William R. Hendee, Gary J. Becker, James P. Borgstede, Jennifer Bosma, William J. Casarella, Beth A. Erickson, C. Douglas Maynard, James H. Thrall, and Paul E. Wallner. Addressing overutilization in medical imaging. *Radiology*, 257(1):240–245, 2010.

[3] Bruce J. Hillman and Jeff C. Goldsmith. The uncritical use of high-tech medical imaging. *New England Journal of Medicine*, 363(1):4–6, 2010.

[4] Anders Eklund, Thomas E Nichols, and Hans Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28):7900–7905, jul 2016.

[5] Glinda S. Cooper, Milele L.K. Bynum, and Emily C. Somers. Recent insights in the epidemiology of autoimmune diseases: Improved prevalence estimates and understanding of clustering of diseases. *Journal of Autoimmunity*, 33(3-4):197–207, 2009.

[6] Matthias Schneider and Klaus Krüger. Rheumatoide arthritis - Frühdiagnose und krankheitskontrolle. *Deutsches Arzteblatt International*, 110(27-28):477–484, 2013.

[7] Allan Gibofsky. Overview of epidemiology, pathophysiology, and diagnosis of rheumatoid arthritis. *American Journal of Managed Care*, 18(13):S295–302, 2012.

[8] Dilek Solmaz, Lihi Eder, and Sibel Zehra Aydin. Update on the epidemiology, risk factors, and disease outcomes of psoriatic arthritis. *Best Practice and Research: Clinical Rheumatology*, 32(2):295–311, 2018.

[9] Josef S. Smolen, Daniel Aletaha, and Iain B. McInnes. Rheumatoid arthritis. *The Lancet*, 388(10055):2023–2038, 2016.

[10] IB B McInnes and G Schett. The pathogenesis of rheumatoid arthritis. *New England Journal of Medicine*, pages 2205–19, 2011.

[11] Paul R. Burton, David G. Clayton, Lon R. Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P. Kwiatkowski, Mark I. McCarthy, Willem H. Ouwehand, Nilesh J. Samani, John A. Todd, Peter Donnelly, Jeffrey C. Barrett, Dan Davison, Doug Easton, David Evans, Hin Tak Leung, Jonathan L. Marchini, Andrew P. Morris, Chris C.A. Spencer, Martin D. Tobin, Antony P. Attwood, James P. Boorman, Barbara Cant, Ursula Everson, Judith M. Hussey, Jennifer D. Jolley, Alexandra S. Knight, Kerstin Koch, Elizabeth Meech, Sarah Nutland, Christopher V. Prowse, Helen E. Stevens, Niall C. Taylor, Graham R. Walters, Neil M. Walker, Nicholas A. Watkins, Thilo Winzer, Richard W. Jones, Wendy L. McArdle, Susan M. Ring, David P. Strachan, Marcus Pembrey, Gerome Breen, David St. Clair, Sian Caesar, Katherine Gordon-Smith, Lisa Jones, Christine Fraser, Elaine K. Green, Detelina Grozeva, Marian L. Hamshere, Peter A. Holmans, Ian R. Jones, George Kirov, Valentina Moskvina, Ivan Nikolov, Michael C. O'Donovan, Michael J. Owen, David A. Collier, Amanda Elkin, Anne Farmer, Richard Williamson, Peter McGuffin, Allan H. Young, I. Nicol Ferrier, Stephen G. Ball, Anthony J. Balmforth, Jennifer H. Barrett, D. Timothy Bishop, Mark M. Iles, Azhar Maqbool, Nadira Yuldasheva, Alistair S. Hall, Peter S. Braund, Richard J. Dixon, Massimo Mangino, Suzanne Stevens, John R. Thompson, Francesca Bredin, Mark Tremelling, Miles Parkes, Hazel Drummond, Charles W. Lees, Elaine R. Nimmo, Jack Satsangi, Sheila A. Fisher, Alastair Forbes, Cathryn M. Lewis, Clive M. Onnie, Natalie J. Prescott, Jeremy Sanderson, Christopher G. Mathew, Jamie Barbour, M. Khalid Mohiuddin, Catherine E. Todhunter, John C. Mansfield, Tariq Ahmad, Fraser R. Cummings, Derek P. Jewell, John Webster, Morris J. Brown, G. Mark Lathrop, John Connell, Anna Dominiczak, Carolina A. Braga Marcano, Beverley Burke, Richard Dobson, Johannie Gungadoo, Kate L. Lee, Patricia B. Munroe, Stephen J. Newhouse, Abiodun Onipinla, Chris Wallace, Mingzhan Xue, Mark Caulfield, Martin Farrall, Anne Barton, Ian N. Bruce, Hannah Donovan, Steve Eyre, Paul D. Gilbert, Samantha L. Hider, Anne M. Hinks, Sally L. John, Catherine Potter, Alan J. Silman, Deborah P.M. Symmons, Wendy Thomson, Jane Worthington, David B. Dunger, Barry Widmer, Timothy M. Frayling, Rachel M. Freathy, Hana Lango, John R.B. Perry, Beverley M. Shields, Michael N. Weedon, Andrew T. Hattersley, Graham A. Hitman, Mark Walker, Kate S. Elliott, Christopher J. Groves, Cecilia M. Lindgren, Nigel W. Rayner, Nicholas J. Timpson, Eleftheria Zeggini, Melanie Newport, Giorgio Sirugo, Emily Lyons, Fredrik Vannberg, Adrian V.S. Hill, Linda A. Bradbury, Claire Farrar, Jennifer J. Pointon, Paul Wordsworth, Matthew A. Brown, Jayne A. Franklyn, Joanne M. Heward, Matthew J. Simmonds, Stephen C.L. Gough, Sheila Seal, Michael R. Stratton, Nazneen Rahman, Sclerosis Maria Ban, An Goris, Stephen J. Sawcer, Alastair Compston, David Conway, Muminatou Jallow, Kirk A. Rockett, Suzannah J. Bumpstead, Amy Chaney, Kate Downes, Mohammed J.R. Ghori, Rhian Gwilliam, Sarah E. Hunt, Michael Inouye, Andrew Keniry, Emma King, Ralph McGinnis, Simon Potter, Rathi Ravindrarajah, Pamela Whittaker,

Claire Widden, David Withers, Niall J. Cardin, Teresa Ferreira, Joanne Pereira-Gale, Ingileif B. Hallgrimsdóttir, Bryan N. Howie, Chris C.A. Spencer, Zhan Su, Yik Ying Teo, Damjan Vukcevic, David Bentley, and Alistair Compston. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[12] Maria Kourilovitch, Claudio Galarza-Maldonado, and Esteban Ortiz-Prado. Diagnosis and classification of rheumatoid arthritis. *Journal of Autoimmunity*, 48-49:26–30, 2014.

[13] M. Kristen Demoruelle and Kevin D. Deane. Treatment strategies in early rheumatoid arthritis and prevention of rheumatoid arthritis. *Current Rheumatology Reports*, 14(5):472–480, 2012.

[14] J M Moll and V Wright. Psoriatic arthritis. *Seminars in Arthritis and Rheumatism*, 3(1):55–78, 1973.

[15] K. B. Sokoll and P. S. Helliwell. Comparison of disability and quality of life in rheumatoid and psoriatic arthritis. *Journal of Rheumatology*, 28(8):1842–1846, 2001.

[16] Alexis Ogdie and Pamela Weiss. The Epidemiology of Psoriatic Arthritis. *Rheumatic Disease Clinics of North America*, 41(4):545–568, 2015.

[17] D. D. Gladman, C. Antoni, P. Mease, D. O. Clegg, and O. Nash. Psoriatic arthritis: Epidemiology, clinical features, course, and outcome. *Annals of the Rheumatic Diseases*, 64(SUPPL. 2):14–17, 2005.

[18] Philip E. Stuart, Rajan P. Nair, Lam C. Tsoi, Trilokraj Tejasvi, Sayantan Das, Hyun Min Kang, Eva Ellinghaus, Vinod Chandran, Kristina Callis-Duffin, Robert Ike, Yanming Li, Xiaoquan Wen, Charlotta Enerbäck, Johann E. Gudjonsson, Sulev Kõks, Külli Kingo, Tõnu Esko, Ulrich Mrowietz, Andre Reis, H. Erich Wichmann, Christian Gieger, Per Hoffmann, Markus M. Nöthen, Juliane Winkelmann, Manfred Kunz, Elvia G. Moreta, Philip J. Mease, Christopher T. Ritchlin, Anne M. Bowcock, Gerald G. Krueger, Henry W. Lim, Stephan Weidinger, Michael Weichenthal, John J. Voorhees, Proton Rahman, Peter K. Gregersen, Andre Franke, Dafna D. Gladman, Gonçalo R. Abecasis, and James T. Elder. Genome-wide Association Analysis of Psoriatic Arthritis and Cutaneous Psoriasis Reveals Differences in Their Genetic Architecture. *American Journal of Human Genetics*, 97(6):816–836, 2015.

[19] Simon J. Bond, Vernon T. Farewell, Catherine T. Schentag, and Dafna D. Gladman. Predictors for radiological damage in psoriatic arthritis: Results from a single centre. *Annals of the Rheumatic Diseases*, 66(3):370–376, 2007.

[20] National Institute for Health Excellence and Care. Etanercept, infliximab and adalimumab for the treatment of psoriatic arthritis. 2010.

[21] Richard G. Langley, Boni E. Elewski, Mark Lebwohl, Kristian Reich, Christopher E.M. Griffiths, Kim Papp, Lluís Puig, Hidemi Nakagawa, Lynda Spelman, BárÃřur Sigurgeirsson, Enrique Rivas, Tsen-Fang Tsai, Norman Wasel, Stephen Tyring, Thomas Salko, Isabelle Hampele, Marianne Notter, Alexander Karpov, Silvia Helou, and Charis Papavassilis. Secukinumab in Plaque Psoriasis âĂŤ Results of Two Phase 3 Trials. *New England Journal of Medicine*, 371(4):326–338, 2014.

[22] Iain B. McInnes, Arthur Kavanaugh, Alice B. Gottlieb, Lluís Puig, Proton Rahman, Christopher Ritchlin, Carrie Brodmerkel, Shu Li, Yuhua Wang, Alan M. Mendelsohn, Mittie K. Doyle, Jacob Aelion, Gaspar Akovbyan, Daina Andersone, Audrey Bakulev, Vida Basijokiene, Andre Beaulieu, Charles Birbara, Erin Boh, Marc Bourcier, Jurgen Braun, Jan Brzezicki, Russell Buchanan, John Budd, Loreta Bukauskiene, Michael Burnette, Juan Canete Crespillo, Wayne Carey, Chandra Chattapadhyay, Dariusz Chudzik, Robert Cooper, Edit Drescher, Anna Dudek, Jan Dutz, Hisham El-Kadi, Ludwig Erlacher, Scott A. Fretzin, Juan A. Garcia Meijide, Emmanuel George, Nigel Gilchrist, Geoffrey Gladstein, Peter Gow, Winfried Graninger, Robert M. Griffin, Lyn Guenther, Wayne Gulliver, Stephen Hall, Dale G. Halter, Kathryn Hobbs, Elana Ilivanova, Pentti Jarvinen, Slawomir Jeka, Peter Jones, Majed Khraishi, Muza Kokhan, Attila Kovacs, Laszlo Kovacs, Alexey Kubanov, Rod Kunynetz, Richard Langley, Susan Lee, Craig L. Leonardi, Clode Lessard, Virginija Lietuvininkiene, Paul Lizzul, Alan Martin, Alexey Maslyanskiy, Robert T. Matheson, Helene Mikazane, Frederick T. Murphy, Peter Nash, Eugeny Nasonov, Frederico Navarro Sarabia, Alexander Orlov-Morozov, Leena Paimela, William Palmer, Kim Papp, Margarita Pileckyte, Gyula Poor, Yves Poulin, Ruben Queiro Silva, Ronald Rapoport, Audrey Rebrov, Maria Rell-Bakalarska, Phoebe Rich, Maureen Rischmueller, Bernadette Rojkovich, Cheryl Rosen, Les Rosoph, Clemens Scheinecker, Rudolph Schopf, Michael Sebastian, Stuart Seigel, Saeed Shaikh, Tom Sheeran, William J. Shergy, Valery Shirinsky, Evan L. Siegel, Howard Sofen, Wolfram Sterry, Jerzy Supronik, Zoltan Szabo, Ferenc Szanyo, Hasan Tahir, Jerry Tan, William Taylor, Vadim Temnikov, Diamant Thaci, Darryl Toth, Ilona Újfalussy, Heikki Valleala, Ronald Vender, Norman Wasel, Martin Willans, Jurgen Wollenhaupt, Omid Zamani, Ellen Zanetakis, Elena Zonova, David Zoschke, and Anna Zubrzycka-Sienkiewicz. Efficacy and safety of ustekinumab in patients with active psoriatic arthritis: 1 year results of the phase 3, multicentre, double-blind, placebo-controlled PSUMMIT 1 trial. *The Lancet*, 382(9894):780–789, 2013.

[23] Stephanie G Werner, Hans-Eckhard Langer, Sarah Ohrndorf, Malte Bahner, Peter Schott, Carsten Schwenke, Michael Schirner, Hans Bastian, Gudrun Lind-Albrecht, Bernward Kurtz, Gerd R Burmester, and Marina Backhaus. Inflammation assessment in patients with arthritis using a novel in vivo fluorescence optical imaging technology. *Annals of the Rheumatic Diseases*, 71(4):504–510, 2012.

[24] Stephanie G. Werner, Hans-Eckhard Langer, Peter Schott, Malte Bahner, Carsten Schwenke, Gudrun Lind-Albrecht, Felicitas Spiecker, Bernward Kurtz, Gerd R. Burmester, and Marina Backhaus. Indocyanine Green-Enhanced Fluorescence Optical Imaging in Patients With Early and Very Early Arthritis: A Comparative Study With Magnetic Resonance Imaging. *Arthritis & Rheumatism*, 65(12):3036–3044, 2013.

[25] Ariane Klein, Georg Werner Just, Stephanie Gabriele Werner, Prasad T. Oommen, Kirsten Minden, Ingrid Becker, Hans-Eckhard Langer, Dirk Klee, and Gerd Horneff. Fluorescence optical imaging and musculoskeletal ultrasonography in juvenile idiopathic polyarticular disease before and during antirheumatic treatmentÂă- a multicenter non-interventional diagnostic evaluation. *Arthritis Research & Therapy*, 19(1):147, 2017.

[26] Tomáš Etrych, Henrike Lucas, Olga Janoušková, Petr Chytil, Thomas Mueller, and Karsten Mäder. Fluorescence optical imaging in anticancer drug delivery. *Journal of Controlled Release*, 226:168–181, 2016.

[27] Moses Bio, Pallavi Rajaputra, Gregory Nkepang, and Youngjae You. Far-red light activatable, multifunctional prodrug for fluorescence optical imaging and combinational treatment. *Journal of Medicinal Chemistry*, 57(8):3401–3409, 2014.

[28] Srabani Bhaumik, Jeannette Depuy, and June Klimash. Strategies to minimize background autofluorescence in live mice during noninvasive fluorescence optical imaging. *Lab Animal*, 36(8):40–43, 2007.

[29] Sevim Kahraman, Ercument Dirice, Ahter Dilsad Sanlioglu, Burcak Yoldas, Huseyin Bagci, Metin Erkilic, Thomas S. Griffith, and Salih Sanlioglu. In vivo fluorescence imaging is well-suited for the monitoring of adenovirus directed transgene expression in living organisms. *Molecular Imaging and Biology*, 12(3):278–285, 2010.

[30] Marites P. Melancon, Yuetang Wang, Xiaoxia Wen, James A. Bankson, L Clifton Stephens, Samar Jasser, Juri G. Gelovani, Jeffrey N. Myers, and Chun Li. Development of a Macromolecular Dual-Modality MR-Optical Imaging for Sentinel Lymph Node Mapping. *Investigative Radiology*, 42(8):569–578, aug 2007.

[31] Olga Schenk, Yinghe Huo, Koen L. Vincken, Mart A. van de Laar, Ina H. H. Kuper, Kees C. H. Slump, Floris P. J. G. Lafeber, and Hein J. Bernelot Moens. Validation of automatic joint space width measurements in hand radiographs in rheumatoid arthritis. *Journal of Medical Imaging*, 3(4):044502, 2016.

[32] Yinghe Huo, Koen L. Vincken, Desiree Van Der Heijde, Maria J.H. De Hair, Floris P. Lafeber, and Max A. Viergever. Automatic Quantification of Radiographic Wrist Joint Space Width of Patients with Rheumatoid Arthritis. *IEEE Transactions on Biomedical Engineering*, 64(11):2695–2703, 2017.

[33] Yinghe Huo, Koen L. Vincken, Max A. Viergever, and Floris P. Lafeber. Automatic joint detection in rheumatoid arthritis hand radiographs. *Proceedings - International Symposium on Biomedical Imaging*, pages 125–128, 2013.

[34] Mikkel Østergaard, Fiona McQueen, Charlotte Wiell, Paul Bird, Pernille Bøyesen, Bo Ejbjerg, Charles Peterfy, Frédérique Gandjbakhch, Anne Duer-Jensen, Laura Coates, Espen A. Haavardsholm, Kay Geert A. Hermann, Marissa Lassere, Philip O'Connor, Paul Emery, Harry Genant, and Philip G. Conaghan. The OMERACT Psoriatic Arthritis Magnetic Resonance Imaging Scoring System (PsAMRIS): Definitions of key pathologies, suggested MRI sequences, and preliminary scoring system for PsA hands. *Journal of Rheumatology*, 36(8):1816–1824, 2009.

[35] Mikkel Østergaard, Charles G. Peterfy, Paul Bird, Frédérique Gandjbakhch, Daniel Glinatsi, Iris Eshed, Espen A. Haavardsholm, Siri Lillegraven, Pernille Bøyesen, Bo Ejbjerg, Violaine Foltz, Paul Emery, Harry K. Genant, and Philip G. Conaghan. The OMERACT rheumatoid arthritis magnetic resonance imaging (MRI) scoring system: Updated recommendations by the OMERACT MRI in arthritis working group. *Journal of Rheumatology*, 44(11):1706–1712, 2017.

[36] Jaap Fransen and Piet L C M van Riel. The Disease Activity Score and the EULAR Response Criteria. *Rheumatic Disease Clinics of North America*, 35(4):745–757, 2009.

[37] Marco Amedeo Cimmino, Francesca Barbieri, Mikael Boesen, Francesco Paparo, Massimiliano Parodi, Olga Kubassova, Raffaele Scarpa, and Giuseppe Zampogna. Dynamic contrast-enhanced magnetic resonance imaging of articular and extraarticular synovial structures of the hands in patients with psoriatic arthritis. *Journal of Rheumatology*, 39(SUPPL. 89):44–48, 2012.

[38] Valentin S. Schäfer, Wolfgang Hartung, Patrick Hoffstetter, Jörn Berger, Christian Stroszczynski, Martina Müller, Martin Fleck, and Boris Ehrenstein. Quantitative assessment of synovitis in patients with rheumatoid arthritis using fluorescence optical imaging. *Arthritis Research and Therapy*, 15(5), 2013.

[39] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:1–9, 2015.

[41] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *Arxiv*, page 12, 2016.

[42] Claudia Beleites, Ute Neugebauer, Thomas Bocklitz, Christoph Krafft, and Jürgen Popp. Sample size planning for classification models. *Analytica Chimica Acta*, 760(June 2012):25–33, 2013.

[43] S.J. Raudys and A.K. Jain. Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):252–264, mar 1991.

[44] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding Transfer Learning for Medical Imaging. (NeurIPS), 2019.

[45] Mahadevappa Mahesh. *The Essential Physics of Medical Imaging, Third Edition.*, volume 40. jun 2013.

[46] Richard L. Morin and Mahadevappa Mahesh. The Importance of Spatial Resolution to Medical Imaging. *Journal of the American College of Radiology*, 15(8):1127, 2018.

[47] Deepak Soekhoe, Peter van der Putten, and Aske Plaat. On the Impact of Data Set Size in Transfer Learning Using Deep Neural Networks. 2:50–60, 2016.

[48] Hongkee Yoon, Jae Hoon Sim, and Myung Joon Han. Analytic continuation via domain knowledge free machine learning. *Physical Review B*, 98(24):1–7, 2018.

[49] Christopher M. Childs and Newell R. Washburn. Embedding domain knowledge for machine learning of complex material systems. *MRS Communications*, pages 806–820, 2019.

[50] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*, 19(1):221–248, 2017.

[51] Hristina Uzunova, Sandra Schultz, Heinz Handels, and Jan Ehrhardt. Unsupervised pathology detection in medical images using conditional variational autoencoders. *International Journal of Computer Assisted Radiology and Surgery*, 14(3):451–461, 2019.

[52] Thorsten Twellmann, Anke Meyer-Baese, Oliver Lange, Simon Foo, and Tim W. Nattkemper. Model-free visualization of suspicious lesions in breast MRI based on supervised and unsupervised learning. *Engineering Applications of Artificial Intelligence*, 21(2):129–140, 2008.

[53] Y Taigman, M Yang, and M.A. Ranzato. Deepface: Closing the gap to human -level performance in face verification. *CVPR IEEE Conference*, pages 1701–1708, 2014.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *CVPR IEEE Conference*, 2014.

[55] Holger R Roth, Le Lu, Jiamin Liu, Jianhua Yao, Ari Seff, Kevin Cherry, Lauren Kim, and Ronald M Summers. Efficient False Positive Reduction in Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation BT - Deep Learning and Convolutional Neural Networks for Medical Image Computing: Precision Medicine, High Performance and Larg. pages 35–48. Springer International Publishing, Cham, 2017.

[56] Choong Ho Lee and Hyung-Jin Yoon. Medical big data: promise and challenges. *Kidney Research and Clinical Practice*, 36(1):3–11, 2017.

[57] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[58] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, dec 1989.

[59] Kurt Hornik. Approximation Capabilities of Multilayer Neural Network. *Neural Networks*, 4(1989):251–257, 1991.

[60] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The Expressive Power of Neural Networks: A View from the Width. (Nips):1–9, 2017.

[61] Boris Hanin. Universal Function Approximation by Deep Neural Nets with Bounded Width and ReLU Activations. (1):1–9, 2017.

[62] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86, 1998.

[63] Y LeCun, B Boser, J S Denker, D Henderson, R E Howard, W Hubbard, and L D Jackel. Handwritten Digit Recognition with a Back-Propagation Network. *Advances in neural information processing systems*, pages 396–404, 1990.

[64] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012.

[65] Min Lin, Qiang Chen, and Shuicheng Yan. Network In Network. pages 1–10, 2013.

[66] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, pages 609–616, 2009.

[67] M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 586–591. IEEE Comput. Sco. Press, 1991.

[68] Pablo Navarrete and Javier Ruiz-Del-Solar. Analysis and comparison of eigenspace-based face recognition approaches. *International Journal of Pattern Recognition and Artificial Intelligence*, 16(7):817–830, 2002.

[69] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *Arxiv.Org*, 7(3):171–180, 2015.

[70] H Robbins and S Monro. A Stochastic Approximation Model. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[71] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *IEICE Transactions on Information and Systems*, E101D(4):1207–1208, dec 2014.

[72] Robin Winter, Floriane Montanari, Andreas Steffen, Hans Briem, Frank Noé, and Djork-Arné Clevert. Efficient multi-objective molecular optimization in a continuous latent space. *Chemical Science*, 10(34):8016–8024, 2019.

[73] Richard H R Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789):947–951, jun 2000.

[74] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep Sparse Rectifier Neural Networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 315–323, 2011.

[75] Steffen Eger, Paul Youssef, and Iryna Gurevych. Is it Time to Swish? Comparing Deep Learning Activation Functions Across NLP tasks. 2019.

[76] George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Joel Shor, and Michele Covell. Full Resolution Image Compression with Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5306–5314, 2017.

[77] Lovedeep Gondara. Medical Image Denoising Using Convolutional Denoising Autoencoders. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 241–246. IEEE, dec 2016.

[78] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. 2013.

[79] Yingzhen Li and Richard E. Turner. R\'enyi Divergence Variational Inference. (Nips):1–9, feb 2016.

[80] Adji B. Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David M. Blei. Variational Inference via Chi-Upper Bound Minimization. In *Advances in Neural Information Processing Systems*, number 30, pages 2732–2741, 2017.

[81] Hongyoon Choi, Seunggyun Ha, Hyejin Kang, Hyekyoung Lee, and Dong Soo Lee. Deep learning only by normal brain PET identify unheralded brain anomalies. *EBioMedicine*, 2019.

[82] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[83] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[84] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6 2014.

[85] Pavlo M. Radiuk. Impact of Training Set Batch Size on the Performance of Convolutional Neural Networks for Diverse Datasets. *Information Technology and Management Science*, 20(1):20–24, 2018.

[86] Marina Gomtsyan, Nikita Mokrov, Maxim Panov, and Yury Yanovich. Geometry-Aware Maximum Likelihood Estimation of Intrinsic Dimension. pages 1–17, 2019.

[87] Elizaveta Levina and Peter J Bickel. Maximum Likelihood Estimation of Intrinsic Dimension. In L K Saul, Y Weiss, and L Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 777–784. MIT Press, 2005.

[88] Zhou Wang and Alan C Bovik. Multi-Scale Structural Similarity For Image Quality Assessment. *Proceedings of the IEEE Asilomar Conference on Signals*, 2(Ki L):1398–1402, 2003.

[89] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi Twitter. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *CVPR*, pages 4681–4690, 2017.

[90] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9907 LNCS:649–666, 2016.

[91] Phillip Isola, Jun Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:5967–5976, 2017.

[92] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. (2015):1–14, 2015.

[93] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context Encoders: Feature Learning by Inpainting. 2016.

[94] Laurens van der Maaten and Hinton Geoffrey E. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 164(2210):10, 2008.

[95] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018.

[96] Giuseppe Pasqualetti, Giovanni Gori, Corrado Blandizzi, and Mario Del Tacca. Healthy volunteers and early phases of clinical experimentation. *European Journal of Clinical Pharmacology*, 66(7):647–653, 2010.

[97] David J. Pinato, Chara Stavraka, Mark Tanner, Audrey Esson, Eric W. Jacobson, Martin R. Wilkins, and Vincenzo Libri. Clinical, Ethical and Financial Implications of Incidental Imaging Findings: Experience from a Phase I Trial in Healthy Elderly Volunteers. *PLoS ONE*, 7(11), 2012.

[98] Janet Stocks, Neena Modi, and Robert Tepper. Need for Healthy Control Subjects when Assessing Lung Function in Infants with Respiratory Disease. *American Journal of Respiratory and Critical Care Medicine*, 182(11):1340–1342, dec 2010.

[99] C. Fink, C. Alt, L. Uhlmann, C. Klose, A. Enk, and H. A. Haenssle. Intra- and interobserver variability of image-based PASI assessments in 120 patients suffering from plaque-type psoriasis. *Journal of the European Academy of Dermatology and Venereology*, 32(8):1314–1319, 2018.

[100] Dror Koltin, Clodagh S. O'Gorman, Amanda Murphy, Bo Ngan, Alan Daneman, Oscar M. Navarro, Cristian Garcia, Eshetu G. Atenafu, Jonathan D. Wasserman, Jill Hamilton, and Marianna Rachmiel. Pediatric thyroid nodules: Ultrasonographic characteristics and inter-observer variability in prediction of malignancy. *Journal of Pediatric Endocrinology and Metabolism*, 29(7):789–794, 2016.

[101] Mu Zhou, Lawrence O. Hall, Dmitry B. Goldgof, Robert J. Gillies, and Robert A. Gatenby. Imbalanced learning for clinical survival group prediction of brain tumor patients. *Medical Imaging 2015: Computer-Aided Diagnosis*, 9414:94142K, 2015.

[102] E. Burnaev, P. Erofeev, and A. Papanov. Influence of resampling on accuracy of imbalanced classification. *Eighth International Conference on Machine Vision (ICMV 2015)*, 9875(Icmv):987521, 2015.

[103] Oskar Pfungst. Clever Hans (the Horse of Mr. Von Osten). *J. Philos. Psychol. Sci. Method*, 8:663–666, 1911.

[104] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1–8, 2019.

[105] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. *Advances in Neural Information Processing Systems*, (Nips):2360–2368, 2016.

[106] Chris Varano and Lytton Ave. Disentangling Variational Autoencoders for Image Classification. 2017.

[107] Qiuyu Zhu and Ruixin Zhang. A Classification Supervised Auto-Encoder Based on Predefined Evenly-Distributed Class Centroids. pages 1–17, 2019.

[108] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[109] Leo Breiman. Randomizing outputs to increase prediction accuracy. *Machine Learning*, 40:229–242, 2000.

[110] Leo Breiman. USING ADAPTIVE BAGGING TO DEBIAS REGRESSIONS. *Technical Report, UCB*, 547:1–16, 1999.

[111] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[112] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 2016.

[113] Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5305 LNCS(PART 4):705–718, 2008.

[114] Jan Neumann, Christoph Schmaderer, Sebastian Finsterer, Alexander Zimmermann, Dominik Steubl, Anne Helfen, Markus Berninger, Fabian Lohöfer, Ernst J. Rummeny, Reinhard Meier, and Moritz Wildgruber. Non-invasive quantitative assessment of microcirculatory disorders of the upper extremities with 2D fluorescence optical imaging. *Clinical Hemorheology and Microcirculation*, 70(1):69–81, 2018.

[115] Radin Adi Aizudin Bin Radin Nasirudin, Reinhard Meier, Carmen Ahari, Matti Sievert, Martin Fiebich, Ernst J. Rummeny, and Peter B. Noël. Preliminary clinical results: an analyzing tool for 2D optical imaging in detection of active inflammation in rheumatoid arthritis. *Medical Imaging 2011: Biomedical Applications in Molecular, Structural, and Functional Imaging*, 7965:796512, 2011.

[116] Jeremy Casorso, Xiaolu Kong, Wang Chi, Dimitri Van De Ville, B. T.Thomas Yeo, and Raphaël Liégeois. Dynamic mode decomposition of resting-state and task fMRI. *NeuroImage*, 194(February):42–54, 2019.

[117] Arnold Neumaier and Tapio Schneider. Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software*, 27(1):27–57, 2001.

[118] James M. Kunert-Graf, Kristian M. Eschenburg, David J. Galas, J. Nathan Kutz, Swati D. Rane, and Bingni W. Brunton. Extracting Reproducible Time-Resolved Resting State Networks using Dynamic Mode Decomposition. *bioRxiv*, page 343061, 2018.

[119] Clarence W Rowley, Igor Mezić, Sheverin Bagheri, Philipp Schlatter, and D Henningson. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, 641:115–127, 2009.

[120] Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010.

[121] Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. VAMPnets for deep learning of molecular kinetics. *Nature Communications*, 9(1):1–11, 2018.

[122] Bethany Lusch, J Nathan Kutz, and Steven L Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. Technical report, 2018.

[123] B. O. Koopman. Hamiltonian Systems and Transformation in Hilbert Space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, may 1931.

[124] Igor Mezić. Spectral Properties of Dynamical Systems, Model Reduction and Decompositions. *Nonlinear Dynamics*, 41(1-3):309–325, aug 2005.

[125] Igor Mezić. Analysis of Fluid Flows via Spectral Properties of the Koopman Operator. *Annual Review of Fluid Mechanics*, 45(1):357–378, 2013.

[126] Marcus Weber. Eigenvalues of non-reversible Markov chains - A case study. Technical Report 17-13, ZIB, Takustr. 7, 14195 Berlin, 2017.

[127] Frank Noé. Machine Learning for Molecular Dynamics on Long Timescales. pages 1–27, 2018.

[128] Hao Wu and Frank Noé. Variational approach for learning Markov processes from time series data. (Dmd):1–30, 2017.

[129] Steven L. Brunton, Bingni W. Brunton, Joshua L. Proctor, Eurika Kaiser, and J. Nathan Kutz. Chaos as an intermittently forced linear system. *Nature Communications*, 8(1):1–8, 2017.

[130] Hassan Arbabi and Igor Mezić. Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the Koopman operator. *SIAM Journal on Applied Dynamical Systems*, 16(4):2096–2126, 2017.

[131] An Xin-lei and Zhang Li. Dynamics analysis and Hamilton energy control of a generalized Lorenz system with hidden attractor. *Nonlinear Dynamics*, 94(4):2995–3010, 2018.

[132] Zhilin Zheng and Li Sun. Disentangling Latent Space for VAE by Label Relevant/Irrelevant Dimensions. pages 12192–12201, 2018.

[133] Stefan Klus, Péter Koltai, and Christof Schütte. On the numerical approximation of the Perron-Frobenius and Koopman operator. *Journal of Computational Dynamics*, 3(1):51–79, 2016.

[134] Stefan Klus, Feliks Nüske, Sebastian Peitz, Jan-Hendrik Niemann, Cecilia Clementi, and Christof Schütte. Data-driven approximation of the Koopman generator: Model reduction, system identification, and control. pages 1–29, 2019.

[135] Andreas Benn, Christian Hiepen, Marc Osterland, Christof Schütte, An Zwijsen, and Petra Knaus. Role of bone morphogenetic proteins in sprouting angiogenesis: Differential BMP receptor-dependent signaling pathways balance stalk vs. tip cell competence. *FASEB Journal*, 31(11):4720–4733, 2017.

[136] Martina Fischer, Paul Rikeit, Petra Knaus, and Catherine Coirault. YAP-mediated mechanotransduction in skeletal muscle. *Frontiers in Physiology*, 7(FEB), 2016.