INAUGURAL – DISSERTATION

zur

Erlangung der Doktorwürde

der

Gesamtfakultät für Mathematik, Ingenieur- und Naturwissenschaften

der

Ruprecht – Karls – Universität

Heidelberg

vorgelegt von

Lehmann, David Hermann, M.Sc.

aus Ostercappeln

Tag der mündlichen Prüfung : ……

Prediction of ventricular pressure and diagnosis from cardiac magnetic resonance imaging using artificial neural networks

Supervisor: Prof. Dr. Vincent Heuveline
Co-Supervisor: Prof. Dr. Benjamin Meder

Abstract

Diastolic heart failure is the most common cause of heart insufficiency worldwide. Diagnosis is typically established via hemodynamic measurements during invasive cardiac catheterization. However, this is associated with interventional risks for the patient. In this dissertation artificial intelligence (AI) models are proposed to predict left-ventricular filling pressures based on non-invasive cardiac magnetic resonance imaging (MRI).

A total cohort of 66,936 patients receiving cardiac catheterization, including 11,699 cardiac MRI examinations, was investigated. The developed AI model could distinguish between elevated and normal filling pressures, providing valuable information on the heart's diastolic function. The novel approach was found superior to established echocardiographic biomarkers and human experts.

A secondary AI model was developed to automatically diagnose various types of cardiomyopathies from cardiac MRI. The detectable disease patterns were: hypertrophic, dilated and ischemic cardiomyopathy, cardiac amyloidosis and control. The AI only required a single MRI frame to reach a diagnosis, which could enable less time-intensive MRI protocols in the future.

Both AI applications were introspected by attention mapping revealing the AI's approach to solve those tasks, thus contributing to explainability. The AI model behind the filling pressure prediction was validated in three independent hospitals involving multiple MRI manufacturers, protocols and models.

In essence, artificial neural networks can predict filling pressure and diagnosis from cardiac MRI, representing a highly scalable approach in the face of overburdened health systems, potentially impacting future diagnosis and treatment strategies.

Zusammenfassung

Die diastolische Herzinsuffizienz ist weltweit die häufigste Ursache für eine Herzinsuffizienz. Die Diagnose wird in der Regel durch hämodynamische Messungen während einer invasiven Herzkatheteruntersuchung gestellt. Dies ist jedoch mit interventionellen Risiken für den Patienten verbunden. In dieser Dissertation wird künstliche Intelligenz (KI) zur Vorhersage des linksventrikulären Füllungsdrucks auf der Grundlage der nicht-invasiven kardialen Magnetresonanztomographie (MRT) vorgeschlagen.

Insgesamt wurde eine Kohorte von 66.936 Patienten untersucht, bei denen eine Herzkatheteruntersuchung durchgeführt wurde, einschließlich 11.699 kardialen MRT-Untersuchungen. Das entwickelte KI-Modell konnte zwischen erhöhtem und normalem Füllungsdruck unterscheiden und damit wichtige Informationen über die diastolische Funktion des Herzens liefern. Der neu-entwickelte Ansatz erwies sich als überlegen gegenüber etablierten echokardiografischen Biomarkern und menschlichen Experten.

Ein sekundäres KI-Modell wurde entwickelt, um verschiedene Arten von Kardiomyopathien anhand von kardialen MRTs automatisch zu diagnostizieren. Die erkennbaren Krankheitsbilder waren: hypertrophe, dilatative und ischämische Kardiomyopathie, kardiale Amyloidose und Kontrollgruppe. Die KI benötigte nur ein einziges MRT-Bild, um eine Diagnose zu stellen, was zukünftig weniger zeitintensive MRT-Protokolle ermöglichen könnte.

Beide KI-Anwendungen wurden durch Aufmerksamkeitslokalisierung untersucht, welche den Ansatz der KI zur Lösung dieser Aufgaben offenbart und zur Erklärbarkeit beiträgt. Das KI-Modell für die Füllungsdruckvorhersage wurde in drei unabhängigen Krankenhäusern unter Verwendung mehrerer MRT-Hersteller, Protokolle und Modelle validiert.

Im Wesentlichen können künstliche neuronale Netze den Füllungsdruck und die Diagnose aus kardialen MRTs vorhersagen, was angesichts der überlasteten Gesundheitssysteme einen hochgradig skalierbaren Ansatz darstellt, der sich möglicherweise auf künftige Diagnose- und Behandlungsstrategien auswirkt.

Table of Contents

1 Introduction

Cardiovascular diseases, which often result in heart failure (HF), present the leading cause of mortality worldwide and are a significant contributor to reduced quality of life [Mensah et al., 2019]. They constitute a major burden on global healthcare systems. Especially, diastolic heart failure is the most prevalent underlying cause of heart insufficiency resulting in substantial morbidity and death [Tsao et al., 2018].

Diastolic heart failure, also known as heart failure with preserved ejection fraction (HF-pEF), is a condition where the heart's ventricles become stiff and do not relax properly during the diastolic phase of the heart cycle, resulting in inadequate filling of the heart with blood. This impairs the heart's ability to pump blood effectively, even though the ejection fraction is normal [Mandinov et al., 2000, Smiseth and Tendera, 2008].

Diastolic heart failure presents unique challenges in management compared to heart failure with reduced ejection fraction (HF-rEF). In recent years, there have been significant advances and shifts in treatment strategies for HF-pEF, reflecting a deeper understanding of the disease's pathophysiology and a focus on personalized medicine [Kirchhof et al., 2014, Spertus et al., 2021, Shah et al., 2024]. Traditionally, treatment for diastolic heart failure has focused on managing symptoms and controlling comorbid conditions with diuretics, antihypertensives and Aldosterone antagonists [McDonagh et al., 2021].

Recent years have seen a shift towards more targeted therapies and novel approaches based on evolving research and clinical trials. Initially used for diabetes management, SGLT2 inhibitors have shown promise in HF-pEF. They aid in decreasing hospitalization rates and enhancing the quality of life for heart failure patients. The DELIVER and EMPEROR-Preserved trial demonstrated significant benefits of Dapagliflozin and Empagliflozin medication in reducing heart failure hospitalizations in HF-pEF patients [Solomon et al., 2021, Anker et al., 2021].

Besides the remarkable progress in the medical treatment options of diverse HF-pEF etiologies, the early and accurate identification of patients and the precise characterization of diastolic function parameters remain difficult tasks [Lehmann et al., 2024]. The gold standard for defining diastolic dysfunction is undoubtedly the invasive intracardiac hemodynamics measured during cardiac catheterization [Dal Canto et al., 2022]. In this procedure, a thin tube (catheter) is inserted into an artery at the patient's groin or wrist and steered through the blood vessels to the left heart chamber. While this can measure the diastolic properties of the heart with the utmost precision, invasive procedures are expensive, not ubiquitous available and carry the risk of potentially severe complications. The risk of major complications during diagnostic cardiac catheterization is 1 %, and the incidence of mortality is 0.05 % for diagnostic purposes [Tavakol et al., 2012].

As a result, echocardiography and cardiac biomarkers are commonly utilized in daily clinical routines for diagnosing diastolic dysfunction without requiring surgical procedures [Vieillard-Baron et al., 2019, Nadar and Shaikh, 2019]. Since the echocardiographic assessment is based on indirect quantification of wall-motion and flow-velocities, the assessment

is frequently challenged by different anatomical, structural and functional co-conditions, often resulting in contradictory information [Mottram and Marwick, 2005, Lehmann et al., 2024]. The natriuretic peptides are critical biomarkers used extensively in the diagnosis and management of HFpEF, as the associated myocardial stress stimulates their release. However, this biomarker also comes with considerable limitations, as its levels may be elevated by other cardiac conditions beyond HFpEF and are affected by factors such as atrial fibrillation, renal dysfunction and obesity [Januzzi Jr and Myhre, 2020].

Artificial intelligence and scalable digital solutions, transforming healthcare globally, with even stronger implications anticipated in the future, are the pristine paradigm to address these modern challenges [van Smeden et al., 2022, Vardas et al., 2022]. Especially, Convolutional Neural Networks (CNNs) have particularly excelled in processing visual information, such as medical imaging data [Avendi et al., 2016, Esteva et al., 2017, Poplin et al., 2018, Zhang et al., 2019]. Artificial neural networks are composed of millions of neurons organized into layers, providing the network with the ability to recognize patterns far beyond human cognition and therefore enabling solutions for previously unsolvable tasks [Goodfellow et al., 2016].

In this dissertation, a non-invasive method to predict important diagnosis of patients undergoing cardiac magnetic resonance imaging (cMRI) and to obtain the left ventricular end-diastolic pressure (LVEDP) by using artificial intelligence is presented. The LVEDP relates both acutely and chronically to clinical conditions that affect ventricular function. It is an important biomarker to assess the extent and severity of the diastolic impairment [Mielniczuk et al., 2007, Du et al., 2015].

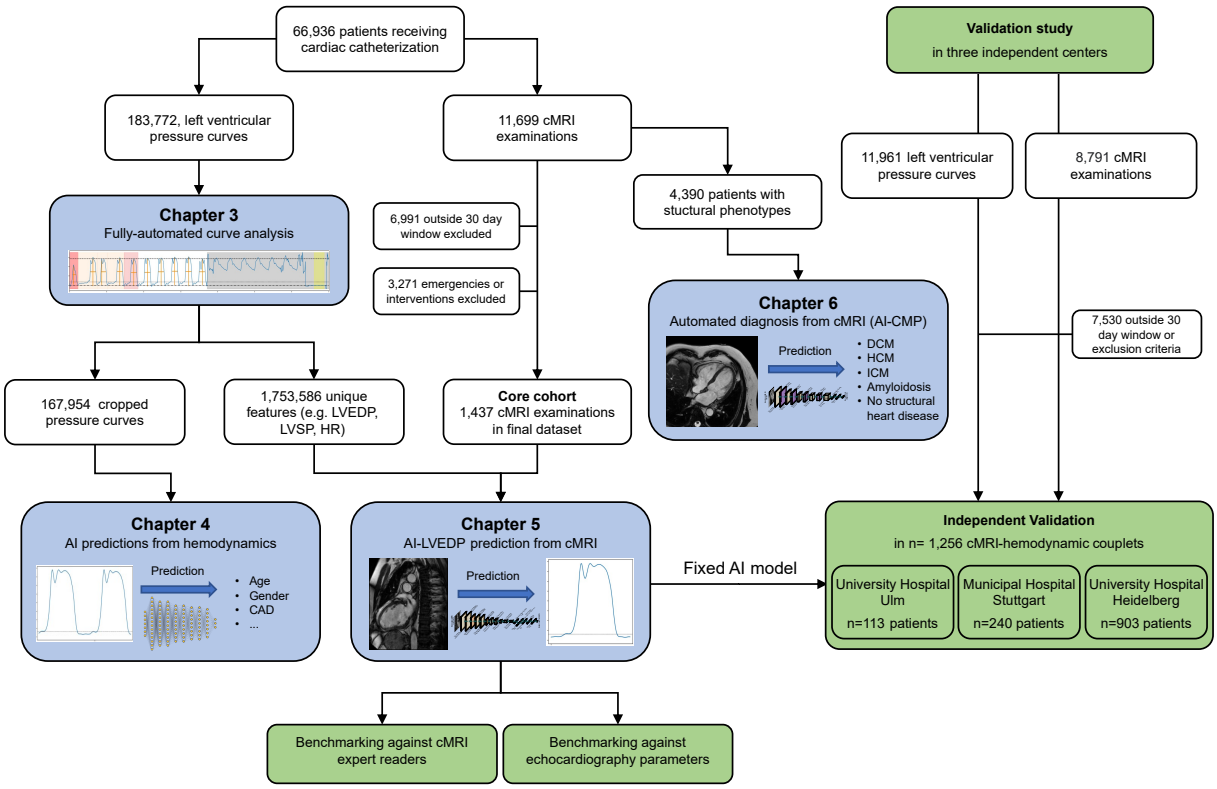


Figure 1.1: Structure of the dissertation.

The integration of this methodology could have a positive impact on treatment strategies and stimulate further scientific advances by enabling the comprehensive characterization of

patients' systolic and diastolic functions, as well as cardiac diagnoses, within a singular cMRI examination [Lehmann et al., 2024]. This novel approach would further increase the already extensive capabilities of cMRI, which is recognized as a Class I diagnostic approach according to recent guidelines [Arbelo et al., 2023]. Addressing diastolic dysfunction and its contributing factors at an earlier stage would greatly benefit a significant number of patients, providing increased precision in the care of this challenging patient group.

In order to achieve this objective, a patient collective that underwent both invasive intracardiac pressure measurements and cMRI examination was selected. Deep neural networks were trained on the cMRI images to predict not only several diagnoses but also LVEDP as an important marker of diastolic function [Chung et al., 2015]. The AI responsible for the LVEDP prediction was named "AI-LVEDP". The AI model developed to predict patient's diagnosis is referred to as "AI-CMP".

AI-LVEDP was benchmarked against human experts, cardiac biomarkers and echocardiographic parameters. For both AI agents, a comprehensive sensitivity analysis was conducted and AI-introspection was performed to assess the neural network's approach for predicting hemodynamics and diagnosis. Finally, AI-LVEDP was extensively validated in a multicenter, multi-vendor and multi-protocol study, including data from maximum-care hospitals located in Heidelberg, Ulm and Stuttgart.

An overview of the dissertation structure is presented in Figure 1.1, highlighting the different utilized electronic health record resources and the, in the scope of this work, developed algorithms and AI systems. In Chapter 2, the functionality of artificial neural networks is introduced, including concepts and various techniques, i.e. model regularization and introspection methodology, which were fundamental to the success of this machine learning endeavor.

In Chapter 3, an algorithm was developed to automatically annotate the hemodynamic left ventricular pressure curves, which were recorded during the cardiac catheterization. The main objective of the algorithm was to robustly extract the LVEDP, which was later used as the ground truth objective to train AI-LVEDP. However, the functionality was extended to provide a comprehensive analysis of the pressure curves, contributing to a deeper understanding of cardiac hemodynamics and enabling the search for novel biomarkers. Based on these findings AI models capable of inferring patient's sex, age and diagnosis of coronary artery disease from a ventricular pressure curve were developed (Chapter 4), giving intriguing insights into the hidden information content of the curves.

In Chapter 5, the development and evaluation of AI-LVEDP is discussed. This included careful selection of training data, since all patients with events likely affecting LVEDP between cardiac MRI and hemodynamics had to be excluded. Moreover, thorough benchmarking against gold-standard methods, introspection and independent external validation was conducted.

The design of AI-CMP is outlined in Chapter 6, where CMP is short for cardiomyopathy. AI-CMP was trained to diagnose hypertrophic cardiomyopathy, dilated cardiomyopathy, ischemic cardiomyopathy and cardiac amyloidosis, which constitute the most common myocardial diseases [Brieler et al., 2017]. Additionally, the learning task included identifying patients without any structural heart disease.

Finally, in Chapter 7, the potential impact and implications of this dissertation and foundations for future research are discussed.

The research findings presented in Chapter 3, 4, 5 and 6 have been partially published in *The Lancet Digital Health* as part of this dissertation's academic contribution [Lehmann et al., 2024].

2 Artificial Neural Networks

Artificial neural networks, or simply neural networks (NN), are transforming industries by enabling machines to learn from data and perform tasks that previously required human intelligence. Their potential for automation and rapid processing of data at a large scale continues to drive innovation and opens up new possibilities across various fields [Voulodimos et al., 2018, Devlin et al., 2018, Jaganathan et al., 2019].

Commonly, the term neural network is used synonymously with terms such as artificial intelligence, machine learning or deep learning. However, these terms have a distinct hierarchical relationship (Figure 2.1).

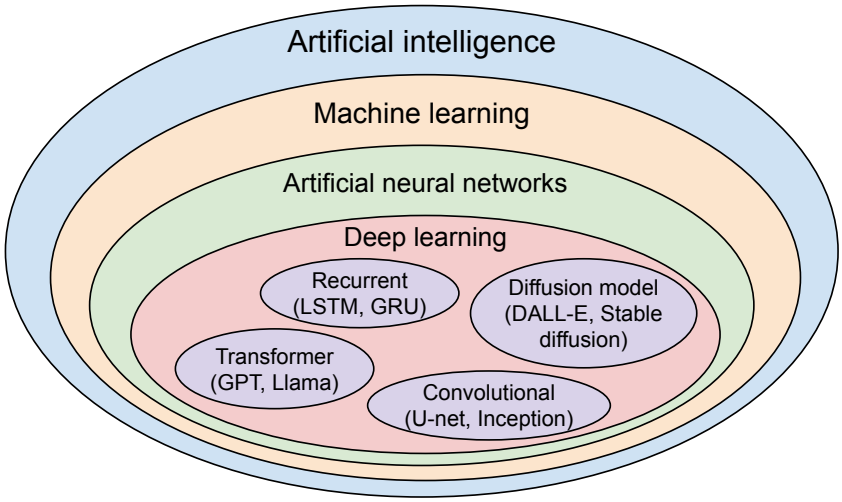


Figure 2.1: Artificial intelligence - hierarchical overview.

Artificial intelligence (AI) is a broad domain comprising systems that can extract intelligence. Intelligence itself is a vague entity, which is not trivial to define. Definitions range from the capability of thought processes and reasoning to the mimicking of human behavior [Russell and Norvig, 2016].

Machine learning is a sub-domain of AI requiring the algorithms to learn without being explicitly programmed [El Naqa and Murphy, 2015]. Consecutively, neural networks constitute the sub-domain of neurologically-inspired machine learning models. They comprise the collective of brain-inspired algorithms, although methods are highly mathematically idealized and further evolved, often leaving only little conformity with the biological foundation [Yang and Wang, 2020].

The highly popularized term of deep learning is a sub-field of artificial neural networks [Campesato, 2020]. Formally, deep neural networks are neural networks with a high number of neuron layers [Schmidhuber, 2014]. Advances in hardware (e.g. graphic processing units) and the realization that neural networks scale well with large amounts of data have contributed significantly to the success of deep learning [Owens et al., 2008].

A recurring pattern in deep learning is the creation of hierarchical representations. Challenging machine learning problems require sophisticated representations in order to enable adequate solutions. Neural networks with multiple layers learn different levels of representations. By recombining simple features, the artificial neural network can model relationships of any complexity [LeCun et al., 2015, Lehmann, 2018].

This is the basis for complex deep learning architectures such as large convolutional neural networks (U-net, Inception) [Ronneberger et al., 2015, Szegedy et al., 2016], recurrent neural networks inspired by memorization mechanisms [Hochreiter and Schmidhuber, 1997], diffusion models for text-to-image generation [Marcus et al., 2022, Rombach et al., 2022] and transformer-based large language models (Chat GPT, Llama) [Singh et al., 2023, Touvron et al., 2023].

Despite these very stunning and overwhelming applications, the essence of all machine learning models is simple pattern recognition and interpolation. This chapter gives an overview on the functionality of artificial neural networks and prominent associated techniques.

2.1 Introduction to Artificial Neural Networks

An artificial neural network consists of individual neurons, which are the basic component [Nielsen, 2015, Choi et al., 2020]. The artificial neuron is motivated by the biological neuron. However, it is highly mathematically idealized. Biological neurons naturally model neuron activation over time (e.g. spiking), whereas the artificial neuron does not describe temporal dependencies. An illustration of an artificial neuron is shown in Figure 2.2. The neuron has an

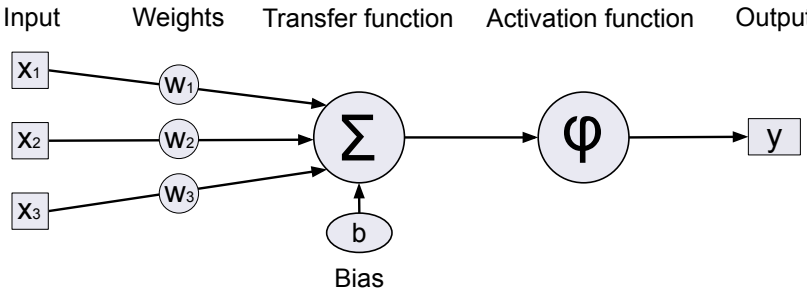


Figure 2.2: Artificial neuron.

input x. The respective output y of the neuron can be formalized as

$$y = \varphi\Big(\sum_i w_i x_i + b\Big), \tag{2.1}$$

where w is the weight vector of the neuron. The bias b is an offset added to the product of weight vector and input. This scalar is then processed by an activation function φ to produce the output y.

Artificial neurons are aggregated into layers, each of which consists of multiple individual neurons. In the case of a fully-connected neural network, each neuron is connected to all neurons in the preceding and succeeding layer (Fig. 2.3). Neurons within the same layer are typically not interconnected. The first layer is the input layer and models the input dimensions of the network.

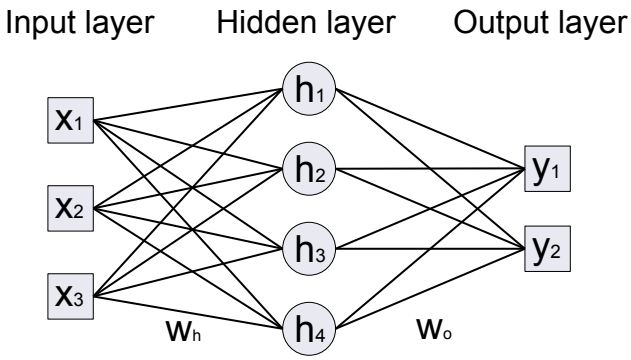


Figure 2.3: Artificial neural network.

The last layer, also known as the output layer, provides the model prediction. The intermediate layers are referred to as hidden layers [Yadav et al., 2015]. The outputs of the hidden layers are frequently called the latent space [Pan et al., 2008]. The latent space captures the essential features or characteristics of the input data in a compressed form. The model output $\vec{y}$ for a fully-connected neural network with a single hidden layer and an input layer activation $\vec{x}$ is given by

$$\vec{y} = \varphi_o(W_o \cdot \varphi_h(W_h \cdot \vec{x} + \vec{b}_h) + \vec{b}_o). \tag{2.2}$$

The functions φ_h and φ_o represent the activation functions for the hidden layer and the output layer. In neuron layers, the weight vectors of individual neurons expand to weight matrices. These weight matrices comprise all the connection weights to the preceding layer. Specifically, W_h denotes the weight matrix of the hidden layer, while W_o represents the weight matrix of the output layer. Each bias vector b includes a bias value for each neuron in its respective layer [Haykin, 1999]. The process of producing the output of a neural network is known as forward propagation, as each activation is passed on to the subsequent layer to generate the next layer activation.

2.2 Training Artificial Neural Networks

A key characteristic of artificial neural networks is their ability to learn. The learning process itself consists of adapting the strength of the connections between the neurons. Neural networks are trained through the process of backpropagation. The backpropagation algorithm is a gradient-based optimization method for calculating the parameter updates of neural networks [Rumelhart et al., 1986]. In order to quantify the feedback, a loss function has to be specified. The loss function evaluates the discrepancy between the model output and the ground truth label. The explicit comparison of the model prediction to the ground truth is categorized as supervised learning [Niculescu-Mizil and Caruana, 2005].

The output is produced by propagating the input through the network, with the network input as initial activation. The loss function L takes network output y and the label $\hat{y}$ as arguments and generates the model loss. Commonly used metrics to quantify this distance are mean squared

error (MSE), mean absolute error (MAE) and cross-entropy

$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad L_{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad L_{CE} = -\sum_{i}\sum_{j}\hat{y}_{i,j}\log(y_{i,j}). \tag{2.3}$$

Here, i is the sum over all training instances. Cross entropy is often used in a multiclass setting, implying multiple output neurons [Grandini et al., 2020]. Hence, j denotes the sum over each individual output neuron. The objective of the training procedure is to minimize the loss by optimizing the parameter configuration, with gradient descent enabling this process. The derivatives of the loss function provide the neurons with the necessary feedback to modify their weights and biases accordingly. This is achieved through the calculation of the appropriate derivatives of the loss function. As a result, the loss travels backwards through the network. The update rules for weights w and biases b are

$$w_{t+1} = w_t - \eta\frac{dL(w_t)}{dw_t} \qquad \text{and} \qquad b_{t+1} = b_t - \eta\frac{dL(b_t)}{db_t}. \tag{2.4}$$

The learning rate, denoted by the hyperparameter η, controls the size of the gradient update. The dynamic programming paradigm and the chain rule are utilized to compute the derivatives of the loss function effectively [Bellman, 1954]. With the chain rule for derivatives, commonly computed parts of the total derivative can be separated and stored [Deisenroth et al., 2020]. This leads to a considerable reduction in calculation complexity, as partial derivatives can be reused and do not have to be calculated redundantly.

In general, loss functions in neural networks are non-convex, resulting in few theoretical guarantees regarding convergence [Goodfellow and Vinyals, 2014]. The non-convex nature of the loss function, combined with a multitude of parameters, leads to numerous local minima. The gradient descent algorithm adjusts weights and biases to follow the direction of the steepest descent on the multidimensional error surface. Given the typical non-convexity of the loss function, gradient descent can only ensure convergence to a local minimum. Even the convergence to global minima (zero training loss) is feasible in deep over-parametrized neural networks [Du et al., 2019, Oymak and Soltanolkotabi, 2020]. However, the non-convexity is not a limitation since global convergence is not a desirable constellation as it implies a lack of generalization (overfitting).

2.3 Activation functions

The activation function is a crucial component of an artificial neuron, as it introduces non-linearity into the system. Without an activation function, the weight matrices from multiple neuron layers can be multiplied together to form a single weight matrix that represents the entire network. As a result, the multilayer network would effectively be reduced to an equivalent single-layer network, which can only perform a linear combination of inputs [LeCun et al., 2015, Lehmann, 2018]. A commonly used activation function is the rectified linear unit (ReLU)

$$\mathrm{ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}. \tag{2.5}$$

The advantage of ReLU is the fast calculation, as it is often evaluated as zero. However, ReLU units can be fragile during training and can ”die” [Lu et al., 2019]. A large gradient in a ReLU neuron can update its weights so drastically that it will never activate on any datapoint again. As a result, the gradients of the neuron’s parameters will remain zero permanently.

The scaled exponential linear unit (SELU) does not inherit this problem and is proven to have self-normalizing properties. Based on the Banach fixed-point theorem it is proven that SELU activation

$$\mathrm{SELU}(x) = \begin{cases} s \cdot x & \text{if } x > 0 \\ s \cdot a \cdot (e^x - 1) & \text{if } x \leq 0 \end{cases}, \text{ where a}\approx 1.673 \text{ and s}\approx 1.051. \tag{2.6}$$

causes activations propagated through many network layers to converge towards zero mean and unit variance [Klambauer et al., 2017].

Both activations presented are not affected by the vanishing gradient problem that occurs with functions such as sigmoid or tangens hyperbolicus, which compress their inputs into a restricted output range. Thereby, the gradient can decrease exponentially as it is passed backwards through each layer, which leads to inadequately small signals for parameter optimization [Hochreiter, 1991, Hochreiter et al., 2001].

In multiclass classification it is typical to use one neuron per class in the output layer. In such cases the target vectors are one-hot encoded and the preferred activation function is softmax activation [Bishop, 2006]. The softmax activation transforms an unprocessed activation vector x by applying the exponential function to each value. Another key property is that the sum of all values in the softmax vector equals one [Graves and Schmidhuber, 2005]. Hence, the softmax output can be interpreted as a pseudo-probability. This is ensured by normalizing each exponential value with the sum of all exponentials. The softmax activation of neuron i for an input vector $\vec{x}$ is given by

$$\mathrm{softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}. \tag{2.7}$$

The exponential scaling increases the discrepancy between input activations. This often results in the predominant pseudo probability being close to 1.0 and hence making the neural network appear very decisive.

2.4 Parameter initialization

Research suggests that many attempts in the past to train deep neural networks have likely failed due to poor initialization schemes [Sutskever et al., 2013]. A well-regarded approach for initializing weight matrices is Glorot Uniform Initialization [Glorot and Bengio, 2010]. Also known as Xavier Initialization, this method uses values sampled from a uniform distribution to set the weight matrix values, with biases initialized to zero

$$\mathrm{Glorot} \sim \mathrm{Uniform}\left[-\frac{\sqrt{6}}{\sqrt{n_{in}+n_{out}}}, \frac{\sqrt{6}}{\sqrt{n_{in}+n_{out}}}\right]. \tag{2.8}$$

Where n_{in} stands for the number of input neurons and n_{out} for the number of output neurons in the corresponding layer. The initialization weights are drawn from a uniform distribution with boundaries set at $\frac{\sqrt{6}}{\sqrt{n_{in}+n_{out}}}$ for the upper limit and $-\frac{\sqrt{6}}{\sqrt{n_{in}+n_{out}}}$ for the lower limit.

Glorot Uniform Initialization helps in preserving the gradient magnitude during backpropagation. By maintaining constant variance of gradients, Glorot Uniform Initialization addresses the vanishing and exploding gradient problem. Consistent gradient flow ensures that all layers learn effectively, making the training process faster, more stable and efficient [Hanin, 2018].

2.5 Optimization

Optimization in the context of neural networks is the process of systematic adjustment of model parameters, such as weights and biases, to achieve minimal loss [Abdolrasol et al., 2021]. Basic weight updates through gradient descent, as stated in Equation 2.4, are infrequently used to optimize neural networks. To enhance convergence speed and reach superior optima, supplementary techniques are typically applied. Two widely used enhancements are learning rate decay and momentum.

At first, it is assumed that the optimum is far away from the initialization, hence the gradient updates should be large to accelerate convergence. Momentum increases the size of the current parameter update dependent on previous gradients by adding an additional term to the update rule

$$w_{t+1} = w_t - \eta \frac{dL(w_t)}{dw_t} - \alpha \frac{dL(w_{t-1})}{dw_{t-1}}. \tag{2.9}$$

In this expression, L denotes the loss function, $\frac{dL(w_{t-1})}{dw_{t-1}}$ indicates the gradient update from the earlier time step, and α is a hyperparameter that controls the sizing of the momentum term.

Momentum is based on the principle that a large gradient update in the previous step suggests that the next gradient update should also be large. Empirical studies have shown that using a momentum term can significantly enhance convergence speed in neural networks [Rumelhart et al., 1986]. Moreover, momentum is advantageous for escaping poor local minima within the non-convex loss function surface, often resulting in a model with better overall performance [Sutskever et al., 2013].

Learning rate decay, a second commonly used technique, involves gradually decreasing the learning rate as the training advances. This is particularly useful when the neural network's weight configuration is nearing a good minimum, where smaller parameter updates are more effective for precise convergence. Conversely, a high learning rate could lead to oscillation around the minimum in the same setting [Lehmann, 2018]. RMSprop (Root Mean Square propagation) builds upon this intuition and introduces adaptive learning rates for each individual trainable parameter [Hinton et al., 2012a].

Adam (Adaptive moment estimation) is a widely used optimization algorithm for training machine learning models and the preferred choice for optimization in this dissertation. Adam combines ideas from RMSprop and momentum. The algorithm calculates an exponential moving average of past gradients and squared gradients to adaptively adjust the learning rates for each parameter [Kingma and Ba, 2014].

Furthermore, bias correction is utilized to compensate for the initialization bias of the m and v

Algorithm 1 Adam

Input: Initial parameter vector θ_0
Stochastic objective function $f(\theta_0)$
Learning rate $\alpha = 0.001$
Exponential decay rates for the moment estimates $\beta_1 = 0.9$, $\beta_2 = 0.999$
Numerical term $\epsilon = 10^{-8}$

$m_0 \leftarrow 0$ ▷ Initialize 1^{st} moment vector
$v_0 \leftarrow 0$ ▷ Initialize 2^{st} moment vector
$t \leftarrow 0$ ▷ Initialize timestep
while θ_t not converged **do**
  $t \leftarrow t + 1$
  $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ ▷ Get gradients with respect to stochastic objective at timestep t
  $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ ▷ Update biased first moment estimate
  $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ ▷ Update biased second raw moment estimate
  $\widehat{m_t} \leftarrow \frac{m_t}{1-\beta_1^t}$ ▷ Compute bias-corrected first moment estimate
  $\widehat{v_t} \leftarrow \frac{v_t}{1-\beta_2^t}$ ▷ Compute bias-corrected second raw moment estimate
  $\theta_t \leftarrow \theta_{t-1} - \frac{\alpha \cdot \widehat{m_t}}{\sqrt{\widehat{v_t}}+\epsilon}$ ▷ Update parameters
end while
return θ_t ▷ Return updated parameters

estimates. Particularly, in the early stages of training, the moving averages are biased towards zero. The exact functionality of Adam is stated in Algorithm 1.

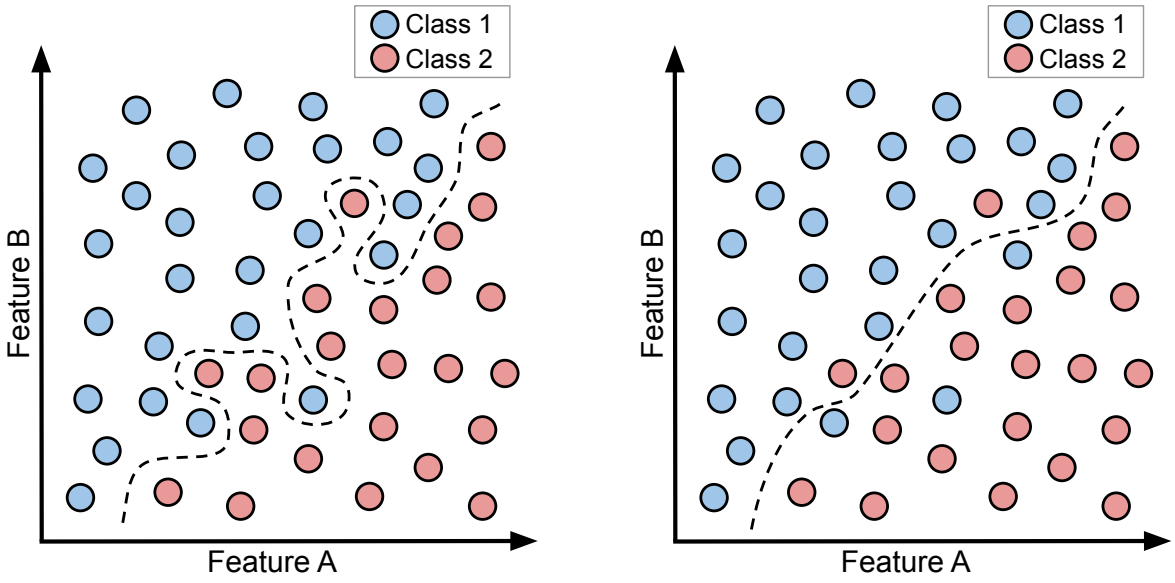
In essence, Adam is conceived to offer a dynamic and effective optimization method that uses momentum, adaptive learning rates and bias corrections. This combination aims to increase the speed of convergence and overcome the complexity of training deep neural networks. By adaptively adjusting learning rates based on historical gradients, Adam deftly navigates the intricacies of different datasets and network architectures, making it an efficient optimization strategy for deep learning endeavors.

2.6 Regularization

Regularization is an essential component of machine learning and neural network training to improve model generalization and prevent overfitting. Rather than learning the fundamental concepts in the data, the model might simply memorize the entire training set. Neural networks are highly prone to this phenomenon due to the large number of trainable parameters (degrees of freedom) they possess [Ying, 2019].

If a model fits the training data too precisely, it might not reliably forecast future data, which means the model is overfitted. In other words, overfitting is a model's inability to generalize [Pitt and Myung, 2002, Lehmann, 2018]. Overfitting can be detected by observing a significant disparity between the model's performance on the training data versus its performance on the validation or test data.

This occurs when not only essential patterns in the data are modeled, but also noise (Figure 2.4). To distinguish between real signals and noise, data is needed. The amount of available data is in direct relationship to the complexity of patterns that can be potentially modeled and distinguished from noise (Nyquist-Shannon sampling theorem) [Shannon, 1949].



(a) Overfitting model. (b) Regularized model.

Figure 2.4: Overfitting and regularization. The dashed black lines represent the decision boundaries of the models. In Figure (a), an overfitted model is depicted, classifying every data point perfectly, resulting in a complex decision boundary. In Figure (b), a regularized model is shown, emphasizing generalization rather than fitting to the noise within the training data.

Regularization techniques address this issue by introducing additional constraints or penalties to the model training process, promoting simpler and more generalized models. In traditional machine learning models, the reduction of fit-able parameters was a widely used method to prevent overfitting. In deep learning, the trend is towards keeping the number of parameters high and enforcing regularization with other approaches. However, creating intermediate information bottlenecks is still frequently utilized (e.g. autoencoders) to prevent overadaptation [Hinton and Salakhutdinov, 2006]. The following sections present commonly used methods to counteract overfitting in neural networks.

2.6.1 L1- and L2-regularization

L2-regularization, also known as L2-norm or ridge regularization, is a long prevalent technique used in machine learning to improve model generalization and prevent overfitting [Kukačka et al., 2017]. L2-regularization operates by imposing a penalty on the size of the model parameters, thus favoring smaller coefficients.

This is achieved by introducing a regularization term into the loss function. For a weight matrix W, the regularization part of the loss is defined as the squared sum of all its elements

$$L_{L2} = \lambda \sum_{ij} W_{ij}^2 = \lambda \|W\|_2. \tag{2.10}$$

The hyperparameter λ controls the strength of the regularization. L2-regularization is not exclusive to neural networks, but is applicable to any parametric model. L2-regularization is also commonly known as weight decay because it encourages the parameters to move towards zero

in order to reduce the regularization term [Ng, 2004].

By analogy, the closely related L1-regularization uses the absolute distances instead of squared distance weightings

$$L_{L1} = \lambda \sum_{ij} |W_{ij}| = \lambda \|W\|_1. \tag{2.11}$$

Both variants can be applied simultaneously to the same loss function (elastic net regularization) [Zou and Hastie, 2005].

In essence, L1- and L2-regularization are promoting simpler models by reducing the influence of individual features.

2.6.2 Dropout

In contrast to L2-regularization, Dropout is a technique specifically crafted for neural network regularization [Srivastava et al., 2014]. The regularization does not originate from a penalty, but indirectly from altering the network topology. Specifically, the network interconnections are changed during each training cycle. Applying Dropout to a network layer means that each node is removed from the network with a probability d.

For instance, if $d = 0.5$, approximately half of the neurons will be omitted. This removal implies that all network connections associated with the dropped neuron are ignored. As a result, the network topology changes with each new training step (i.e. for each individual training sample). This process is illustrated in Figure 2.5.

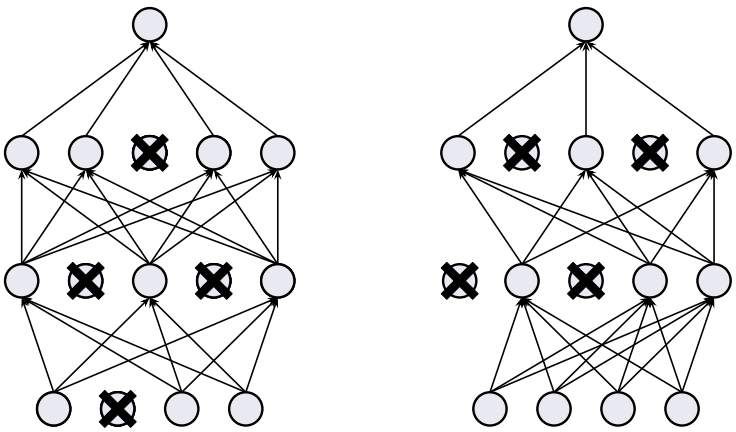


Figure 2.5: Dropout. Neurons excluded at random during each training step, resulting in different network topologies.

Commonly used dropout rates range from 0.1 to 0.5 [Park and Kwak, 2017]. A dropout rate that is too high may lead to underfitting, while a rate that is too low may not effectively prevent overfitting. The optimal dropout rate for a specific use case can only be determined through experimentation and validation.

However, the reduced number of neurons during training comes with implications. Since only a reduced number of neurons is available during training, but the full network is used during inference and testing, the output amplitude has to be adjusted appropriately. To account for the dropout, it is essential that the weights are multiplied by $\frac{1}{(1-d)}$ to balance the reduced number of neurons during training [Goodfellow et al., 2016].

By applying Dropout, the network is forced to generate redundant feature representations, since it prevents the network from relying only on specific neurons to detect essential features. This is a consequence of the network's prediction failing whenever essential neurons drop out. The network learns to not depend on singular neurons by assigning additional neurons to compensate for the deficiency. The network's non-dependability towards specific neurons encourages robustness and generalization [Lehmann, 2018].

Dropout also addresses the challenge of co-adaptation. Co-adaptation occurs when neurons across different layers show strongly correlated behavior. Certain neurons become highly dependent on the presence of other neurons. This dependency can result in poor generalization to new data, as the network becomes too specialized in capturing patterns present in the training data but fails to generalize well to unseen data. Dropout forces neurons to detect features that are universally helpful without excessively relying on specific other neurons [Hinton et al., 2012b].

One negative aspect of Dropout is the increased length of the training process. In every training step, a different sub-network of the full neural network is trained. The resulting gradient updates are solely addressing the sparse network layout. Generally, more sparse gradient updates than full updates are required to achieve comparable performance. The sparse gradient updates can be considered noisy updates in the context of the full network.

Despite this, Dropout remains a leading technique for regularizing neural networks [Labach et al., 2019]. In fact, noisy gradients are not necessarily a disadvantage, as they have been proven to promote generalization in many scenarios [Neelakantan et al., 2015].

2.6.3 Batch normalization

Network parameter updates are usually not calculated for individual training instances, but for multiple instances (batches) simultaneously. Training in large batches benefits generalization, as gradient updates are calculated over a large sample size, reducing the focus on individual entities [Keskar et al., 2016, He et al., 2019].

Whitened input data is known to speed up the convergence of neural networks [LeCun et al., 1998]. Whitening data refers to transforming the data to have a mean of zero, unit variance and to ensure decorrelation. Evolving this concept further, batch normalization normalizes the activation of a neural network layer [Ioffe and Szegedy, 2015].

Batch normalization comprises two consecutive processing parts: a normalization step and a scaling/shifting procedure. During the normalization step, the batch of data x is normalized by the batch mean μ_B and the batch variance σ_B

$$\hat{x}_i = \frac{x_i - \mu_B}{\sigma_B}. \tag{2.12}$$

The resulting batch $\hat{x}_i$ possesses zero mean and unit variance. The batch mean μ_B and the batch variance σ_B are calculated as

$$\mu_B = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \sigma_B = \sqrt{\epsilon + \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_B)^2}. \tag{2.13}$$

The small positive constant ϵ is added to avoid division by zero in Equation 2.12. During inference, the mean and variance are not calculated from the batches, but are estimated from the entire

training set or accumulated during training. This ensures that the normalization is consistent and independent of the batch size.

One challenge with batch normalization is that the imposed unit variance and zero mean on activations $\hat{x}_i$ can limit the expressiveness of the network's neurons. To overcome this limitation, trainable parameters γ and β are introduced, enabling arbitrary mean and standard deviation values [Ioffe and Szegedy, 2015]. The final output of the batch normalization x_B is given as

$$x_{B,i} = \gamma \hat{x}_i + \beta. \tag{2.14}$$

The parameters γ and β are trained analogously to network weights and biases by error backpropagation.

Batch normalization provides regularization effects and leads to faster and more stable training. As specific activation magnitudes are not important due to normalization, but rather the relationships between activations within a batch, batch normalization inherently acts as a regularizer, when applied to neural networks. Incorporating batch normalization layers in a neural network can lessen the need for additional regularization or even make it completely redundant [Lehmann, 2018].

Despite the evident benefits of batch normalization, the underlying mechanisms still remain a topic of discussion. It was previously believed to mitigate internal covariate shift, a phenomenon where parameter initialization and changes in layer input distributions influence the learning rate. However, recent arguments suggest that batch normalization does not actually reduce internal covariate shift but rather smooths the objective function and thereby enhancing performance [Santurkar et al., 2018].

Batch normalization has become a standard component in the design of modern neural networks due to its ability to accelerate training, improve stability, make the network less sensitive to weight initialization and offer a form of regularization. Furthermore, normalizing the activations addresses the vanishing/exploding gradient problem, making batch normalization instrumental for enabling the successful training of deeper neural network architectures [Goodfellow et al., 2016].

2.6.4 Data augmentation

In deep learning, the quality and quantity of data play a pivotal role in model performance. Data augmentation is a technique that addresses both aspects by artificially expanding the dataset through various transformations, while preserving its original meaning [Shorten and Khoshgoftaar, 2019, Wong et al., 2016, Perez and Wang, 2017].

Especially in medical imaging, data scarcity is a common challenge for machine learning projects as obtaining labeled data is expensive and time-consuming [Rajpurkar et al., 2022]. Limited data can lead to overfitting, where the models perform well on training data but fail when generalizing to unseen examples. Data augmentation mitigates these issues by artificially generating additional training samples and thus improving model robustness and generalization.

There are several techniques for data augmentation. In the following, the most relevant options are summarized.

Geometric transformations involve altering the spatial characteristics of the data without changing its semantic content. Common transformations include rotation, translation,

scaling, cropping, flipping and distortion. Slight rotational or translational transformations in MRI images simulate uncertainties in the patient's position and angle. This likely does not change the presence of a heart disease and provides additional variations for the model to learn from [Zoph et al., 2020].

Color augmentation involves modifying the color space of images. This can include changes in brightness, contrast, saturation, and hue. By adjusting the brightness level of images can simulate different lighting conditions, making the model more robust to varying environmental factors.

Noise injection introduces random perturbations to the data, mimicking real-world noise. Gaussian noise, salt and pepper noise, and Poisson noise are common types of noise added to images or numerical data. By exposing the model to augmented inputs, it learns to focus on the most informative features and becomes more tolerant to noise during inference [Maharana et al., 2022, Akbiyik, 2023].

Synthetic data, resembling the original distribution, can be created by generative models such as generative adversarial networks (GANs) or variational autoencoders (VAEs) [Goodfellow et al., 2020, Kingma and Welling, 2013]. These models learn the underlying data manifold and can produce new samples with realistic variations. Augmenting the dataset with synthetic samples from generative models enriches the training data and thereby positively influences model performance [Antoniou et al., 2017].

In summary, data augmentation is a powerful technique for improving the performance and generalization of machine learning models, especially in situations where labeled data is scarce. Furthermore, by expanding the training dataset through various transformations, data augmentation enables models to learn robust representations that more accurately capture the underlying data distribution.

2.7 Convolutional Neural Networks

Convolutional neural networks (CNNs) are commonly used in computer vision. Their ability to detect complex visual patterns is yet unmatched by other approaches [Voulodimos et al., 2018, Szeliski, 2022]. It is remarkable that convolutional layers empower machines with the ability to see [Forsyth and Ponce, 2002]. In order to understand the functionality behind convolutional neural networks, we have to inquire: How does human vision work?

Although the human brain is an incredibly complex structure and is not understood in its entirety by the scientific world, the discovery of simple and complex cells in the visual cortex by Hubel and Wiesel in 1962 gave important glimpses on its optical functionality [Hubel and Wiesel, 1962, Zeki, 1993]. The cells in the visual cortex are organized into receptive fields. These biological receptive fields consist of locally connected photoreceptor cells. The photoreceptor cells are interconnected in such a way that they can recognize contrasts and edges [Lehmann, 2018].

Convolutional neural networks gained prominence after their introduction by LeCun et al. in 1989, due to their effectiveness in handwritten digit classification and face detection [LeCun et al., 1989]. However, in 1980, Fukushima had already proposed the

idea of using convolutional layers in neural networks that resembled contemporary concepts [Fukushima, 1980].

2.7.1 Convolutional layer

In analogy to the visual cortex of mammals, a convolutional layer can also detect simplistic patterns in image data. In the context of computer vision, the receptive field is conventionally referred to as a filter or kernel. The basic functionality of convolutional layers is that a filter is multiplied by a fraction of the input image, resulting in a scalar matrix product (Figure 2.6). The sliding window is then shifted across the image's dimensions and the matrix product is calculated for each position.

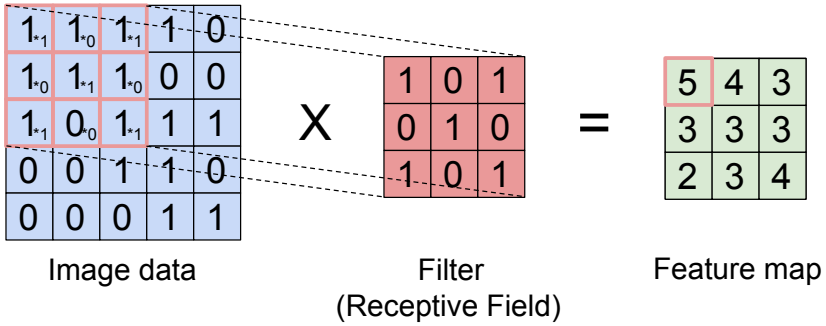


Figure 2.6: Convolutional layer functionality.

The resulting values are called the feature map [Goodfellow et al., 2016]. In areas where the filter matches the image window well, the corresponding feature map entry displays a high value. Filters can be used to detect distinctive patterns such as edges and contrast. This process is shown in Figure 2.7.

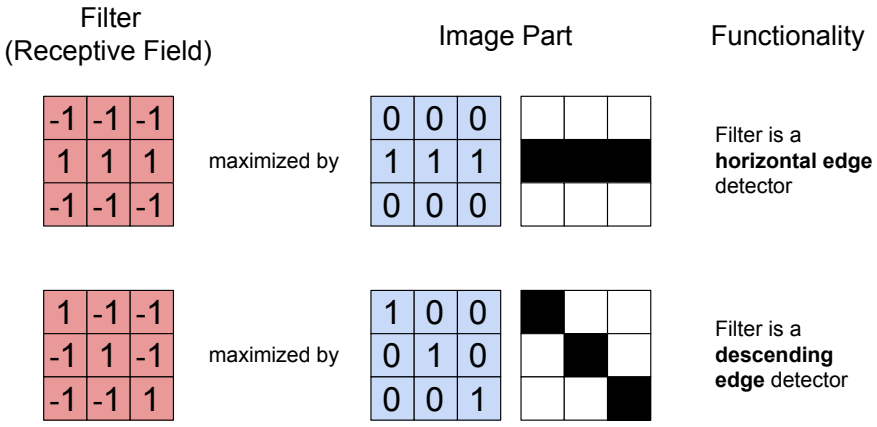


Figure 2.7: Principle of filters functioning as feature detectors.

The feature map is therefore an overview of where certain patterns can be found in the original image. The filter weights are trained via error backpropagation similar to the neuron weights of fully-connected layers. The number of filters per layer are pre-chosen hyperparameters and determine how many feature maps are generated.

The initial layer of a CNN learns a diverse set of filters that are selective for different frequencies, orientations and color blobs [Krizhevsky et al., 2012]. By adding a second convolutional layer

on top, the network now has the ability to recombine learned simplistic patterns and form filters which are selective for more complex structures (Figure 2.8).

The additional layers use a mixture of the previous patterns to generate even higher level features, producing a feature hierarchy. Neural networks begin by detecting simple edges and geometric patterns in the first convolutional layer, which are subsequently combined into higher-level features like circles and other geometric shapes.

As more layers are utilized, these progressively combined features enable the network to recognize increasingly complex shapes, such as human faces or various animal species. With just three layers, it is possible to detect human faces [Zeiler and Fergus, 2013]. Considering that modern neural networks often have more than 100 layers, the attainable complexity goes far beyond what is comprehensible for humans.

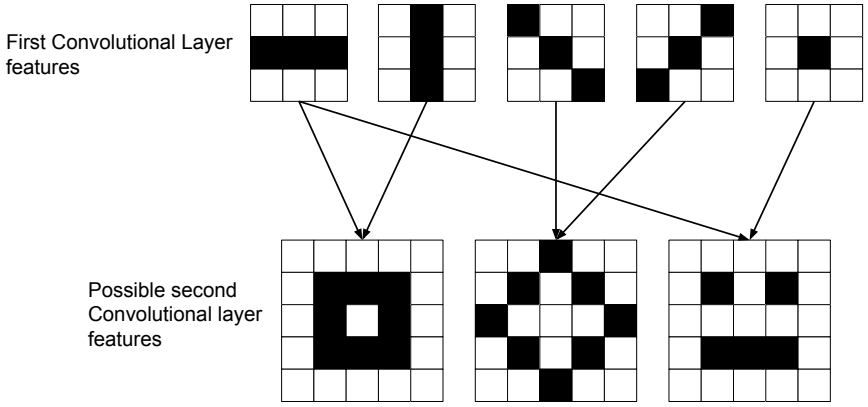


Figure 2.8: Combining features by stacking convolutional layers.

The generic nature of convolutional features, especially in the early network layers, results in high reusability. Task-specific networks can be reused for other tasks. The pre-trained model is retrained to solve a novel task (transfer learning). Only the task-specific features in the back layers of the network have to be adjusted; other parts can remain unmodified [Tan et al., 2018, Lehmann, 2018].

Although the architecture of convolutional neural networks is highly mathematically idealized compared to the visual cortex, it is intriguing to see how feature generation could work in the human brain on the basis of deep learning.

2.7.2 Pooling layer

Alongside their finding of receptive fields, Hubel and Wiesel observed another important aspect in the visual cortex, which motivated an additional core concept of CNNs: the spatial invariance of complex cells. In other words, complex receptive fields are not reliant on the precise location of an object [Hubel and Wiesel, 1962]. This characteristic can be gained in CNNs through the use of pooling layers. Pooling operations are commonly known as sub-sampling or down-sampling. The maximum-pooling technique is a frequently utilized variant employed in neural networks to obtain spatial invariance [Weng et al., 1992].

The max-pooling operation involves the kernel shifting over the data and outputting the maximum value found within the area covered by the pooling window (Figure 2.9). Max-pooling can enhance a model's prediction performance by improving generalization and robustness

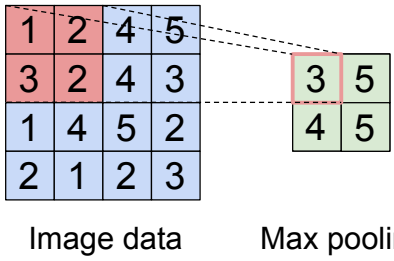


Figure 2.9: Max pooling operation.

through spatial invariance. The concept of spatial invariance can also be interpreted as a decrease in variance within the data [Boureau et al., 2010]. Another major advantage of pooling is the reduction of feature map sizes, which speeds up both training and inference by decreasing computational complexity.

Average-pooling is another popular pooling variant, where instead of returning the maximum value, it calculates the average value of the data within the pooling window.

Taking the idea of local invariance to the extreme, global pooling operations discard all spatial information. Global pooling sets the kernel size equal to the total image size. Only a single numeric value per feature map is extracted [Lin et al., 2013]. This follows the intuition that the key aspect is on whether a feature exists within the image, along with additional importance placed on the frequency and precision of how well the filter pattern is matched.

2.8 Residual Neural Networks

Residual neural networks (ResNets) represent a significant advancement in the design of deep neural network architectures. Introduced by He et al. in 2015, ResNets address key challenges associated with training very deep networks, such as vanishing gradients and degradation of performance [He et al., 2015]. In the same year, Ronneberger et al. introduced the U-net, which heavily relies on residual connections and is still the state-of-the-art architecture for segmentation tasks such as heart segmentation or cancer detection [Ronneberger et al., 2015, Isensee et al., 2021, Liu et al., 2019]. Residual connections are nowadays an essential component of deep learning and can be found in various benchmark models [Tan and Le, 2019, Dai et al., 2021].

The key characteristic of residual networks is the presence of shortcuts, or residual connections, that bypass specific layers. This means that certain connections in the network skip over layers and deliver activations directly to deeper layers. Thus, during backpropagation, gradients are also transmitted through these direct shortcuts.

Without residual connections, significantly increasing the depth of a network is not advantageous. Beyond a certain point, adding more layers can actually diminish the model's performance. This decline is not attributable to overfitting but rather reflects that deeper networks become more challenging to optimize. In theory, deeper networks should perform at least as well as their shallower counterparts, as additional intermediate layers could just output the activation from the preceding layer (i.e. returning their identity) [He et al., 2015].

By incorporating skip connections, ResNets help mitigate the vanishing gradient problem that often occurs in very deep networks. The shortcut connections provide alternate paths for

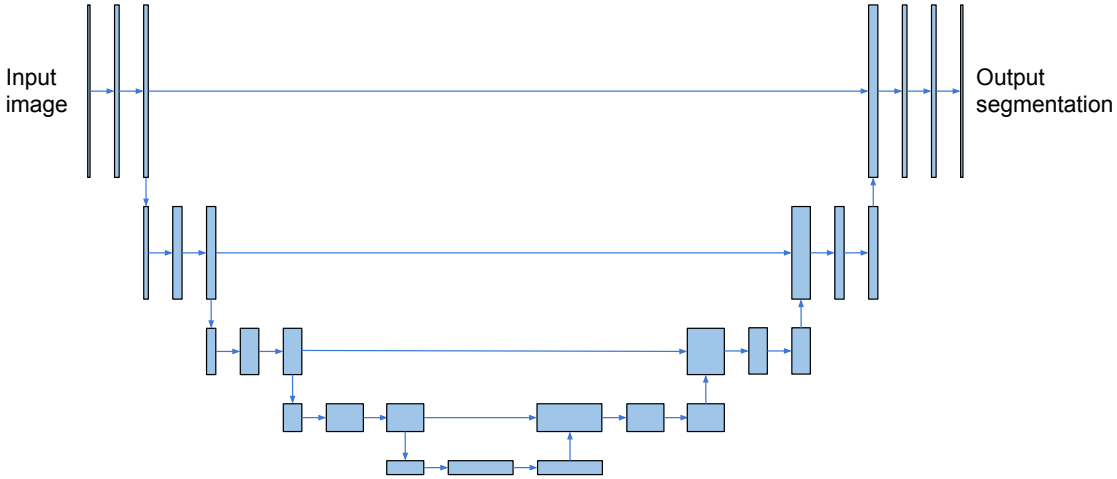


Figure 2.10: U-net. Each box symbolizes a convolutional layer. Box height denotes the feature map resolution. Width indicates the number of feature maps. Each arrow represents a connection. An upward or downward arrow denotes up- or down-sampling by pooling operations.

gradients to flow backward through the network, making it easier to propagate gradients during training. This significantly improves the training of deeper networks [Szegedy et al., 2017]. In Figure 2.10, the various residual connections of a U-net are shown.

2.9 Universal approximation theorem for neural networks

Universal approximation theorems state the limits of what neural networks can theoretically learn [Tikk et al., 2003]. Theoretical foundations of neural network properties have been investigated extensively [Mhaskar, 1996, Dreiseitl and Ohno-Machado, 2002]. In 1989, George Cybenko first proved that neural networks with one hidden layer using ReLU activation and arbitrary width (number of hidden layer neurons) are universal function approximators [Cybenko, 1989]. Hence, they are able to approximate any continuous mathematical function with arbitrary precision. The theorem was extended by Leshno et al. for any non-linear activation function by showing that the universal approximation property is equivalent to the utilization of a non-polynomial activation function [Leshno et al., 1993, Pinkus, 1999]. The resulting doctrine is stated in Theorem 2.9.1

Theorem 2.9.1 (Universal approximation theorem)**.**
Let $C(X, \mathbb{R}^m)$ denote the set of continuous functions from a subset X of a Euclidean $\mathbb{R}^n$ space to a Euclidean space $\mathbb{R}^m$. Let $\sigma \in C(\mathbb{R}, \mathbb{R})$. Note that $(\sigma \circ x)_i = \sigma(x_i)$, so $\sigma \circ x$ denotes σ applied to each component of x.
Then σ is not polynomial if and only if for every $n \in \mathbb{N}$, $m \in \mathbb{N}$, compact $K \subseteq \mathbb{R}^n$, $f \in C(K, \mathbb{R}^m), \varepsilon > 0$ there exist $k \in \mathbb{N}, A \in \mathbb{R}^{k \times n}, b \in \mathbb{R}^k, C \in \mathbb{R}^{m \times k}$ such that

$$\sup_{x \in K} \|f(x) - g(x)\| < \varepsilon$$

where $g(x) = C \cdot (\sigma \circ (A \cdot x + b))$.

Similar to the arbitrary width case, there exist proofs for the arbitrary depth (number of layers)

case [Gripenberg, 2003, Yarotsky, 2017]. In recent years, theorems were also extended for convolutional and residual neural networks [Zhou, 2020, Tabuada and Gharesifard, 2020].

Many different universal approximation theorems exist, implying that neural networks can, in theory, represent a large variety of mathematical functions with appropriate neuron weights. If the chosen network architecture is too small, the inherent representation power may not be sufficient to model the objective function properly. However, the practical value is limited since the theorems do not provide a recipe for distinguishing noise from real signals in the data, which is a major challenge in deep learning. Instead, backpropagation and gradient descent are used for optimization, whereby convergence often results in local minima rather than the optimal solution [Shang and Wah, 1996, Agarwal et al., 2017].

Furthermore, the existence of an objective function f mapping the input information to the desired outcome is not given. Machine learning applications can fail, if the desired objective information can not be derived from the neural network input parameters.

2.10 Attention mapping

Neural networks are often described as black boxes [Castelvecchi, 2016]. Traditionally, decision trees or rule-based learning are perceived to be more explainable than deep learning models [Rokach and Maimon, 2005, Fürnkranz et al., 2012]. However, this is not a problem specific to neural networks, but rather a universal complexity and information problem commonly affecting machine learning algorithms. It may be just as difficult for humans to understand a very large decision tree as it is to comprehend the characteristic matrix multiplication of a neural network, since the amount of parallel processable information is quite limited in humans. These observations lead to the following deliberation:

It may be a paradox in itself to ask an artificial neural network for an explanation, since the complexity of the neural network is potentially far beyond human cognition. The existence of an elucidation understandable to humans is therefore not trivial.

Irrespective of these philosophical considerations, many approaches have been developed to open the black box. A prominent approach in computer vision for generating explanations is attention mapping. The purpose of attention mapping is to mark regions in the input image that were most relevant for the outcome prediction. Attention mapping is mostly done via gradient tracking or occlusion [Smilkov et al., 2017, Shrikumar et al., 2016, Montavon et al., 2018, Zeiler and Fergus, 2013].

In gradient tracking, the gradients of the target class score with respect to the feature maps are calculated. The gradient magnitudes are then projected into a heatmap, indicating parts of the image are most influential in the model's prediction [Selvaraju et al., 2016].

The process of generating saliency (attention) maps by occlusion is highlighted in Figure 2.11. The model prediction for a specific input sequence is compared to the model output for an occluded version of the input. The output difference can then be mapped to the occluded image part. After each part of the image was occluded once or several times, a pixel-precise heatmap can be generated.

Best practice is to replace occluded pixels with the mean pixel value across the input. However,

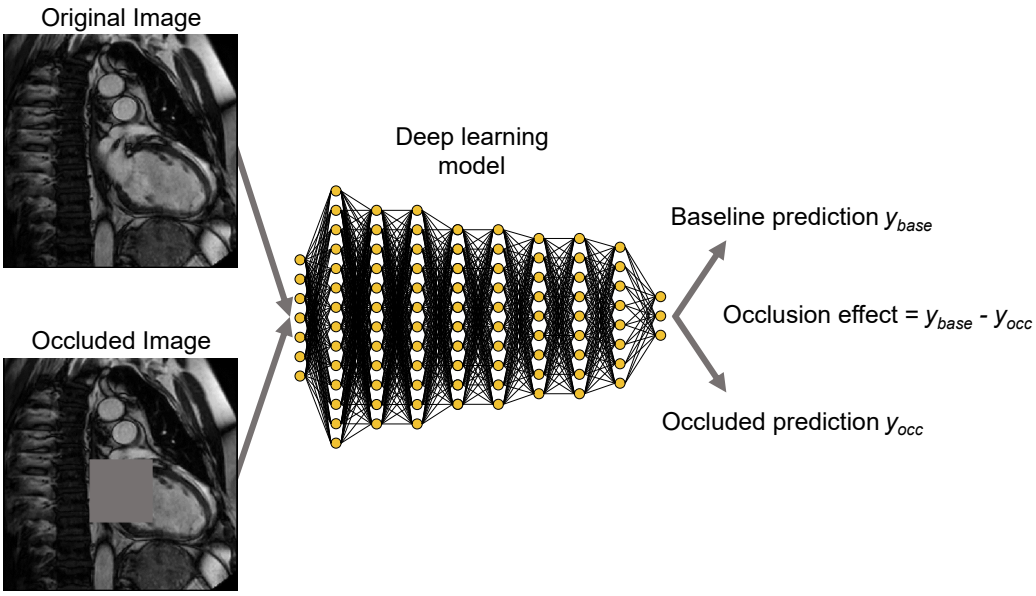


Figure 2.11: Attention mapping by occlusion.

the optimum strategy depends on the specific use case. It can be beneficial to pad around the occlusion kernel with interpolation techniques, since convolutional neural networks are strongly influenced by edges [Yosinski et al., 2015]. The decision on the kernel size is also crucial. The occlusion window should be invasive enough to break the patterns recognized by the convolutions, but not so extensive that it changes the overall semantic content of the input. Combining results from multiple kernel sizes can further increase the robustness and decisiveness of the generated attention localizations.

3 Automated Analysis of Left Ventricular Pressure Curves

In the years between 2006 and 2024 a total of 183,772 left-ventricular pressure curves were measured in the cathlabs of the University Hospital of Heidelberg. Pressure curves and electrocardiogram (ECG) traces were extracted from the hemodynamics database (Ethics vote No S-158/2021) followed by a conversion (decoding) with the help of a self-developed script. Due to the large volume of data, automated processing was essential.

In this chapter, the development and application of an algorithm for the automated evaluation of left-ventricular pressure curves is showcased. The results and structured information, which originated from the automated analysis, built the foundation for the artificial intelligence models developed in Chapter 4, 5 and 6 and enabled comprehensive insights into cardiac hemodynamics.

3.1 Cardiac catheterization

Inter-cardiac pressure is measured via cardiac catheterization. A catheter is inserted into a blood vessel at the patient's arm or groin and pushed through the artery or vein into the heart chamber [Baim and Grossman, 2006, Sorajja et al., 2020]. To access the intercardiac pressure,

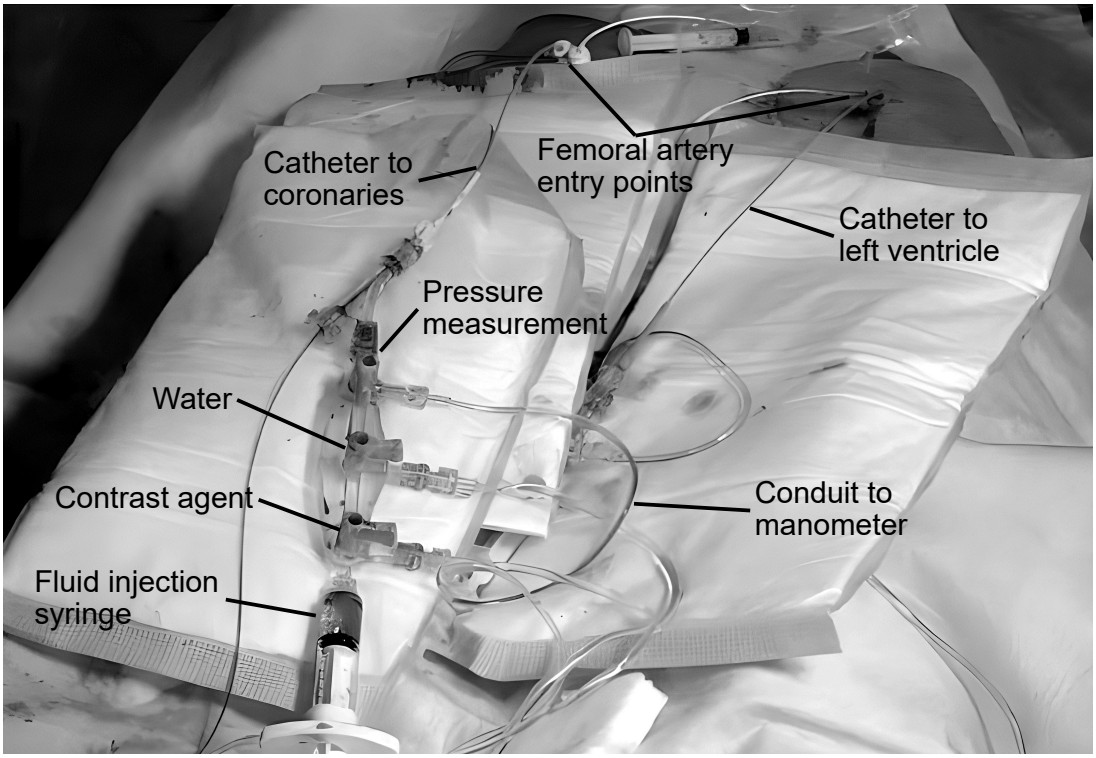


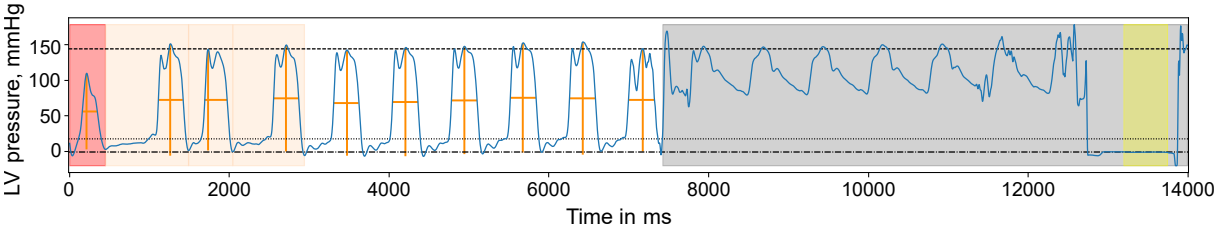
Figure 3.1: Showcase of a cardiac catheterization procedure.

the fluid-filled catheter is used to transmit the pressure signal outside the patient's body to pressure transducers (Figure 3.1). The unit used for quantifying the pressure is traditionally "millimetre of mercury" (mmHg), defined as the extra pressure generated by a column of mercury one millimetre high. The pressure of one millimetre of mercury corresponds to approximately $133.32 \frac{kg}{m \cdot s^2}$ (Pascal) [Taylor, 2001].

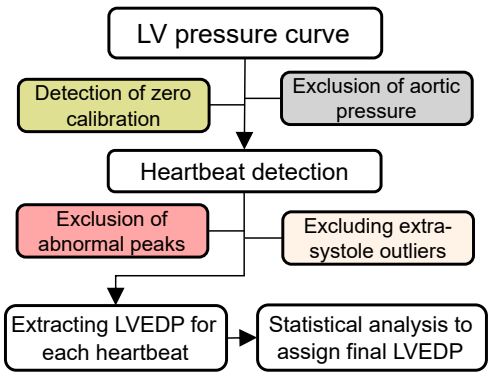
Usually, the pressure is recorded at least twice per patient, once before and once after the ventriculography (visualization of the ventricle). During ventriculography, contrast agent is injected. If capable, the patients hold breath in respiratory center position during both measurements in order to reduce pressure fluctuations caused by intrathoracic pressure changes. The primary goal of the automated analysis is to obtain the left-ventricular end-diastolic pressure (LVEDP) from the curves. The LVEDP is the blood pressure inside the left ventricle at the end of the filling phase of the heart (diastole), in which the volume of the left ventricle is maximized [Taylor, 2018]. Hence, the LVEDP is also referred to as the filling pressure. The extraction of the LVEDP is a non-trivial task, already due to the unstructuredness of the curves. Recorded LV pressure curves may include a multitude of artifacts from the measurement equipment.

3.2 Algorithm description

The left-ventricular pressure curve is a complicated entity. The final algorithm is made up of a multitude of components, each of which solves small subtasks that together constitute the superordinate function. The application was developed according to the "divide-and-conquer" paradigm [Cormen et al., 2022]. In this section, a detailed description of the algorithm components is given. The full algorithm comprised over 2,300 lines of Python code.



(a) Automatically analyzed pressure curve.



(b) Algorithm flow diagram.

Figure 3.2: Left-ventricular pressure curve analysis algorithm example and overview.

An illustration of a fully analyzed pressure curve is shown in Figure 3.2a. An overview of the algorithm functionality is presented in Figure 3.2b. The color encoding of areas marked in the

curve matches the colorization of components in the flow diagram. Next to the main functionality of extracting the LVEDP, the analysis gives us structured access to a multitude of other variables obtained from the pressure curves.

3.2.1 Categorization of curve sections

During the left heart catheterization, the catheter is either in the left ventricle or is retracted into the aorta. Furthermore, the measurement can be disconnected from the catheter. Then, usually, outside pressure is displayed. To distinguish intervals showing aortic pressure from ventricular pressure, the difference in minimum pressure is utilized. Whereas the normal aortic pressure is between 60 and 140 mmHg, the minimum left-ventricular pressure during a heart cycle is usually around zero mmHg [Corsini et al., 2022]. Hence, intervals showing a consistent pressure of greater than 40 mmHg for more than one second can be classified as aortic pressure and were excluded from further analysis.

Next, intervals of extreme values were excluded. The typical left-ventricular pressure ranges from -5 to 140 mmHg. Outliers are defined as values lower than -30 or greater than 300 mmHg indicating artifacts from the measurement device. These thresholds were selected conservatively due to the diverse patient spectrum, including patients with severe cardiological impairments.

Another source of expected artifacts is the disconnection of the catheter from the pressure conductor or the catheter pullback from the ventricle into the aorta. In both cases, the pressure curve may show strong fluctuations. Accordingly, intervals with more than four oscillations of ±10 mmHg within 500 ms were excluded as well.

3.2.2 Detection of cardiac cycles and outlier detection

After merging and excluding all previously defined intervals, heartbeats were detected by recognizing systolic peaks with prominence (freestanding height) greater than 35 mmHg and a width of more than 70 ms. The peak width refers to the respective width at 50 % height. It might happen that a peak is isolated, i.e. that no other consecutive peak is found in the same interval, usually because the peak was followed or preceded by an excluded interval. Isolated peaks were as well excluded.

In the following, a multitude of statistical tests were conducted to exclude various outlier heartbeats. The implemented outlier detection is based on Median Absolute Deviation (MAD) [Leys et al., 2013]. However, the MAD score was supplemented by an additional condition to ensure that only empirically relevant deviations were classified as outliers. The absolute difference between median and potential outlier has to be above a defined threshold in addition to the z-score requirement from MAD [Falk et al., 2012]. The full criterion is illustrated in Algorithm 2. The tilde operator ($\sim$) denotes the median.

Next, extrasystoles were detected by comparing cardiac cycle lengths. Extrasystoles are premature heartbeats that occur when the cardiac electrical signals cause the heart to contract earlier than normal, potentially disrupting the normal cardiac rhythm [Greten and Andrassy, 2010]. The reference point for obtaining cardiac cycle length was the systolic peak. The cycle-specific heart rate (beats per minute) can be calculated from the length of the cardiac cycle. A predefined exclusion cutoff was not suitable, since cardiac cycle length is largely defined by the patient's

Algorithm 2 Outlier detection

Input: List of numbers $A = [a_1, a_2, ..., a_n]$
Z-score threshold z_t
Absolute difference threshold d_t
Output: Boolean mask indicating outliers
$D \leftarrow |A - \widetilde{A}|$ ▷ List of absolute differences to the median
$Z \leftarrow 0.6745 \frac{D}{\widetilde{D}}$ ▷ List of modified z-scores
$Z_b \leftarrow Z > z_t$ ▷ Boolean list z-score outliers
$D_b \leftarrow D > d_t$ ▷ Boolean list absolute difference outliers
return $Z_b \wedge D_b$ ▷ Conjunction list of z-score and absolute difference outliers

baseline heart rate. Instead, heartbeats were excluded in an individualized iterative process. Starting with the list of cardiac cycle lengths, very slow heartbeats, indicating mostly interruptions in the measuring process, were excluded.

Thereafter, heartbeats with high heart frequency were eliminated by comparing them to the 25 % list percentile. The comparison to the percentile instead of the cycle with the lowest heart frequency adds another layer of protection against outliers. The criterion for exclusion was that the reference heart frequency divided by the potential outlier heart frequency is smaller than 0.7. If there are only three heartbeats remaining, the removal process stops. The described algorithm for the exclusion of extrasystolic heartbeats is summarized in Algorithm 3. Once the selection of heartbeats has been completed, the patient's total heart rate is calculated based on the remaining heartbeats.

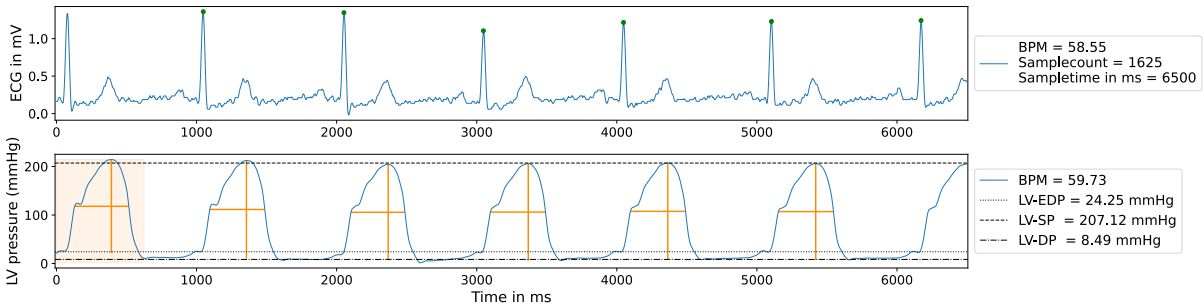
Algorithm 3 Exclusion of extrasystolic heartbeats

Input: List of cardiac cycle lengths in bpm $A = [a_1, a_2, ..., a_n]$
Output: List of cardiac cycle lengths outliers
while Length A > 3 **do**
  if min(A) < 30 bpm **then** ▷ Exclude slow heartbeats
    $A \leftarrow A$ without lowest value
  else if $\frac{25\,\%\text{ percentile of } A}{\max(A)} < 0.7$ **then** ▷ Exclude extrasystolic outliers
    $A \leftarrow A$ without highest value
  else
    return A
  end if
end while
return A

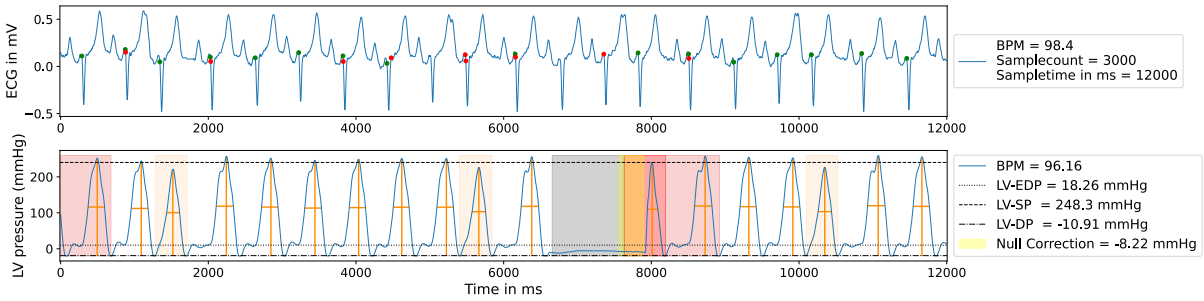
At this stage of the curve analysis, the algorithm is utilized to detect outlier peaks in terms of peak width and peak prominence. The used z-score threshold z_t was 3.5 in both cases. The absolute difference threshold d_t was 75 ms for peak widths and 15 mmHg for prominences. Outliers of that kind indicate highly irregular or only partially recorded cardiac cycles.

3.2.3 Automated curve calibration

During the ventricular pressure measurement, the physician normally records a calibration interval. The calibration is a part of the pressure curve, where the measurement device is not measuring the ventricular pressure, but the environment pressure (1013.25 hPa = 1 atm). The



(a) Missing calibration interval



(b) Calibration interval between two heartbeats

Figure 3.3: Automatically annotated left ventricular curves with ECG traces.

next step in the automated analysis is the detection of the calibration interval. An example of a detected calibration interval (yellow color) is shown in Fig. 3.2a.

Detecting the calibration interval is not trivial as it can be in any part of the curve or not present at all (Figure 3.3a). The basic approach is that the calibration is an interval of relatively constant pressure and should be around zero mmHg. However, these simple rules are not sufficient to solve the problem adequately. In Figure 3.3b, the natural interval between two heartbeats is very shallow and can easily be mistaken for the calibration. In this case, the ECG trace can be utilized to check if a heartbeat indicated by the electric stimulus (QRS complex) in the ECG is missing in the pressure curve [Josephson, 2008]. Also, the calibration interval itself may fluctuate by several millimetres of mercury. Especially, unclean calibration intervals can be difficult to distinguish from natural pressure signals. Therefore, several additional characteristics of calibration intervals are used for effective differentiation. In the following, the key criteria are summarized.

Sufficiently long interval of relatively constant pressure. If the candidate interval is short, then the threshold for pressure variation must be chosen strictly. Vice versa, if the interval is longer, the variation threshold can be chosen less strictly.

High derivatives at interval borders. When changing the hydraulic manometer to measure the environmental pressure, the pressure abruptly changes to zero. Analysis showed that the derivatives measured in such an incident are usually distinguishable higher than derivatives measured during normal heart function.

Consideration of adjacent peaks and ECG. If the potential calibration interval is in between two heartbeats, it can be easily confused with a natural shallow interval during the diastole. In such a case, the ECG can be utilized to detect heartbeats that are missing

in the pressure curve.

Based on these criteria, a set of rules is comprised. An example of a rule pattern is: minimum 700 ms interval length, maximum difference in pressure during that interval 0.7 mmHg, interval is not between peaks (i.e. at the start or end of the curve). If the candidate interval is only 500 ms long, then the maximum pressure difference should be more strict (e.g. 0.5 mmHg) to preserve the same level of confidence. If additional information is included in the rule (e.g. high derivatives and ECG consideration), other conditions can be chosen less strictly.

The algorithm matches each pattern described in those rules to the curve. If a rule is successfully matched, the iteration breaks and the mean value of the found calibration interval is returned. If multiple curve sequences are matched to the same rule, outlier detection is performed and the arithmetic mean is calculated over all non-outlier calibration candidates. This step is necessary since sometimes multiple calibration intervals are recorded by the physician. Additionally, a single calibration interval can have artifacts. In that case, the interval part before and after the artifact (e.g. fluctuation in pressure) is matched separately to the pattern. This iterative and outlier-identifying approach not only ensures the robust determination of an existing calibration interval, but also ensures that only the calibration that carries the highest confidence is selected.

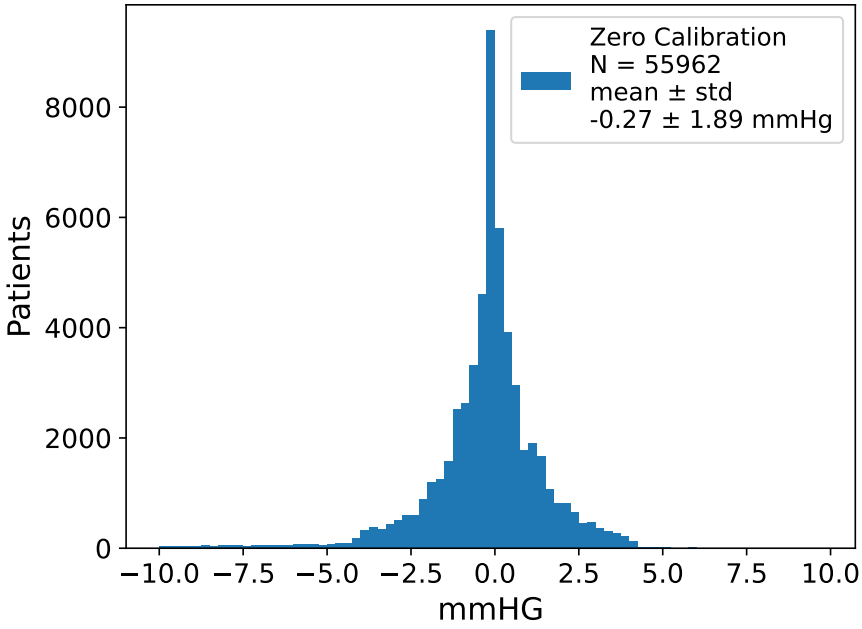


Figure 3.4: Distribution of automatically detected calibration intervals. The offset needed to calibrate the LV curve correctly is presented.

The LV curve can now be adjusted by the calculated offset in order to calibrate the pressure scale to environment pressure. In Figure 3.4, the distribution of found calibration intervals is shown. The mean offset value was -0.27 mmHg with a standard deviation of 1.89 mmHg. The occurrence of calibration offsets greater than 4 mmHg was more rare than offsets lower than -4 mmHg. This is potentially the case because if the offset is positive, then the diastolic intervals are frequently below zero, which is easily detectable by the human eye. In contrast, a negative offset is more difficult to detect since the curve often does not cross the negative range.

This analysis emphasizes the importance of careful left ventricular pressure curve calibration, as even small fluctuations of a few millimetres of mercury can significantly influence the diagnosis

of diastolic heart failure [Oh et al., 2023].

3.2.4 Feature extraction

The final step of the automated curve analysis is the extraction of features. The primary goal is to extract the LVEDP, however a variety of additional features are extracted to find relevant new biomarkers. In this section all automatically extracted features are described.

Left-ventricular end-diastolic pressure (LV-EDP)
Starting at half height on the left flank of the systolic peak, a linear function is fitted to 20 ms curve intervals. The interval is shifted by a single data point at every step. The fit parameters are the slope a and the bias of the linear function. The slope at 50 % height is usually high since this is the center of the contraction phase of the heart cycle. While the 20 ms window shifts to the left, the slope will decrease at some point. If the slope is lower than $0.14\,\frac{mmHg}{ms}$, the related interval is an LVEDP candidate. Multiple candidate intervals can be found until the algorithm terminates at the end of the heart cycle. Since patients may have a weak heartbeat, a standalone slope threshold is not sufficient. Hence, the algorithm also takes into account the adjacent interval slopes a_{t+1}.

$$\exists x \in [0, 0.2] : a_t < x + 0.3 \text{ and } a_{t+1} - a_t > 2x + 0.3. \tag{3.1}$$

The condition (Equation 3.1) is checked for every time step t, starting with the last time step. If the condition is fulfilled in time step t the LVEDP is assigned as 80 % height of the corresponding interval and does not need to be checked for $t-1, t-2, \cdots$. The consideration of the whole 20 ms interval compared to calculating the slope between only two points, showed an improvement in terms of robustness.

Left-ventricular peak systolic pressure (LV-SP)
The main challenge to obtain the peak systolic pressure is to take into account systolic overshoots. This is done by detecting peaks in the 50 % height interval (left edge to right edge) of the systolic phase. Systolic overshoots are characterized by a small peak width and small prominence-to-width ratio. Subpeaks with higher widths are unlikely overshoots and represent a more accurate manifestation of the actual pressure. Peaks with a high prominence-to-width ratio compared to other subpeaks are excluded. The remaining peaks are fitted by a Gauss function and the height is assigned as the function maximum [Lifshits, 2013]. All subpeak heights are weighted by width to finally assign the peak systolic pressure.

Left-ventricular minimum diastolic pressure (LV-DP)
The minimum diastolic pressure is obtained in a similar way to the peak systolic pressure. Peak detection is executed on the inverted curve from the diastolic phase. Gaussian functions are fitted to smoothen the peak values. However, the LV-DP is less prone to overshoots and usually only one peak is considered relevant.

Heart rate
A heart rate is assigned to each individual heartbeat by measuring the distance between systoles. The heart rate in beats per minute can then be derived from the time interval.

Peak prominence
The peak prominence is measured as the freestanding peak height. Systolic overshoots are treated with the same regularizing method utilized to obtain the systolic pressure.

Peak width
The peak width is measured as the width at 50 % prominence.

Peak integral
The peak integral is measured from the minimum diastolic pressure of the current peak to the next peak. The mean pressure value during that time period is calculated and multiplied by the heart cycle time.

Peak rise time
Peak rise time is defined as the time the curve needs to rise from 25 % of the total prominence to 75 %.

Peak receding time
In analogy, the peak receding time is defined as the time the curve needs to recede from 75 % of the total peak prominence to 25 %.

Rising slope
A straight line is fitted to the left peak edge at 50 % prominence. A 20 ms interval is fitted. The rising slope is defined as the slope of the fitted function.

Receding slope
A straight line is fitted to the right peak edge at 50 % prominence (20 ms interval). The receding slope is defined as the slope parameter of the straight.

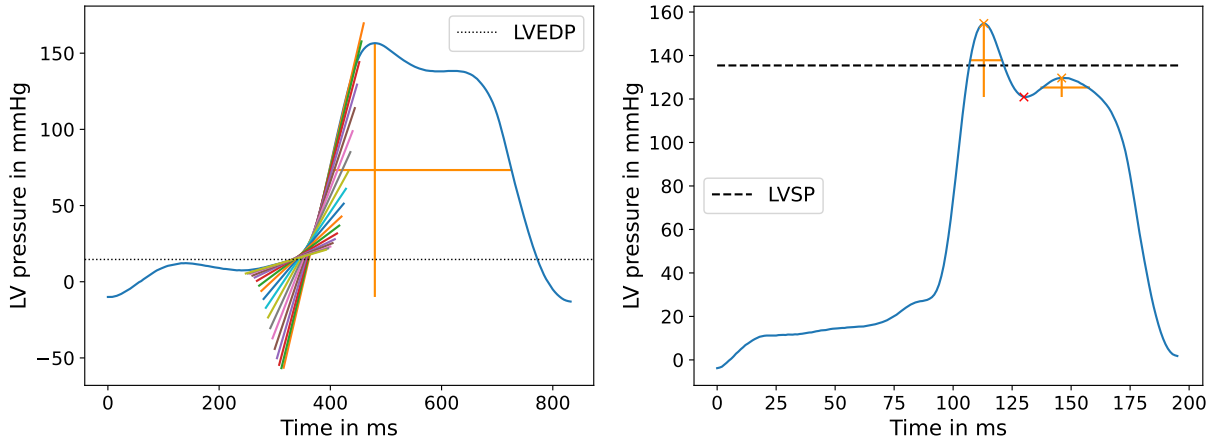


Figure 3.5: Acquisition of LV-EDP by fitting multiple straight lines and acquisition of LV-SP by weighting subpeaks by width.

The extraction of the features is performed for each non-outlier heartbeat individually resulting in a list of values for each feature. Outlier detection with z-score threshold of 1.6 is executed before assigning the final feature value as the arithmetic mean.

3.3 Comparison of automated labeling of LVEDP versus labeling by human experts

Data quality is crucial for the success of machine learning applications [Jain et al., 2020]. The LVEDP is labeled in the cathlab by clinical experts following the inter-ventricular pressure measurement. The dataset was collected over a period of almost 20 years, and accordingly the LVEDP was labeled by over 100 different investigators. Hence, a major occurring problem is the inter-observer variability [Pinedo et al., 2010, Maroules et al., 2018].

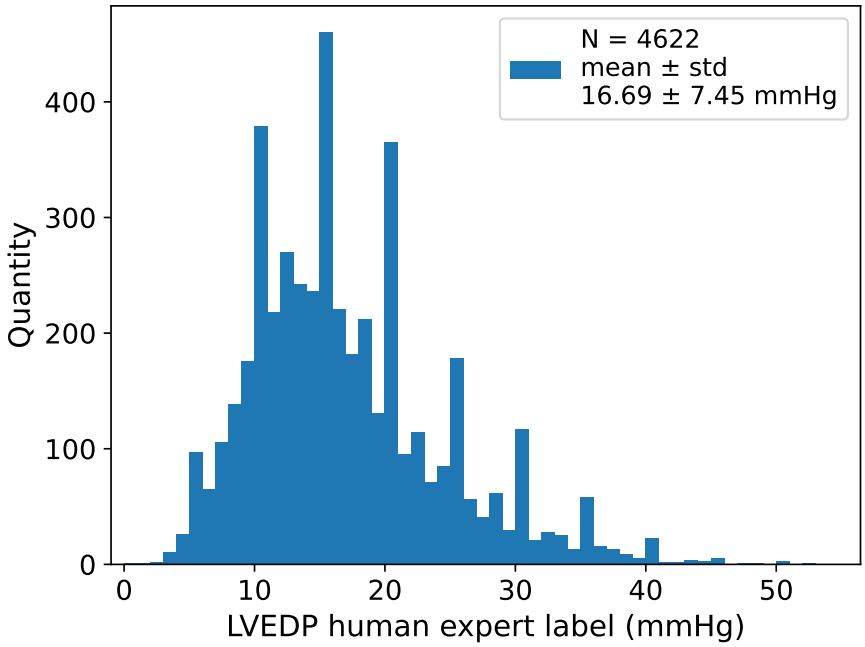


Figure 3.6: LVEDP human expert label distribution.

In contrast, a deterministic algorithm system provides a consistent concept for the artificial intelligence agent to learn [Lehmann, 2018]. In Figure 3.6, a histogram of the expert LVEDP labels is shown. The histogram reveals that the human experts tend to label in steps of 5 mmHg. Hence, labels such as 5, 10, 15, 20 mmHg are much more frequent than others, whereas the automated algorithm labels are continuous without any quantization bias.

Expert and algorithm labels showed a strong Pearson correlation of 0.84 (Fig 3.7). The corresponding coefficient of determination (R^2) was 0.67 [Wright, 1921, Chicco et al., 2021]. The software framework assisting the physicians during the clinical routine only allowed whole numbers as input for LVEDP. This value quantization is evident on the x-axis of the graph. Minor differences in expert versus algorithm labels can be explained by considering calibration, exclusion of extrasystoles, dequantization and exact statistical tests.

Larger differences are often caused by saddle points in the pressure curve during the end-diastolic phase. The left ventricular pressure at the end of the saddle point would be considered correct LVEDP, however in some cases it is hard to distinguish if the filling pressure is still manifested or the systole already started. The difference between measuring the LVEDP after and before the saddle point can easily be 20 mmHg or even more.

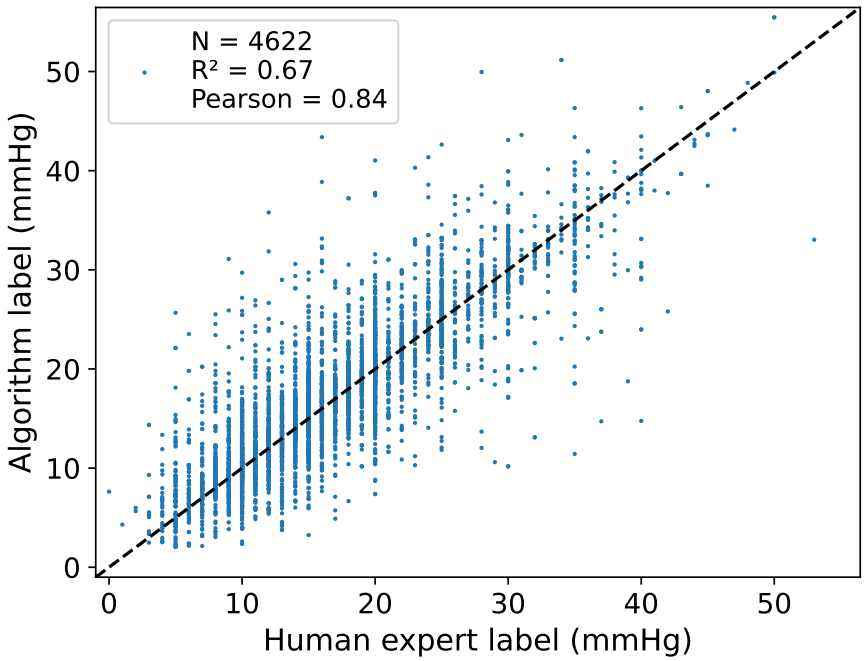


Figure 3.7: LVEDP human expert label versus automated algorithm label.

3.4 Effect of contrast agent application

During the cardiac catheterization, the ventricular pressure is usually measured twice. Once before and once after the ventriculography. For the visualization of the ventricle contrast agent is needed. At the University Hospital of Heidelberg, mainly iodine-based contrast agents such as Iomeprol or Iohexol are utilized [Koç et al., 2019]. The imaging is carried out using X-ray angiography protocols [Van der Meer, 2002].

The automated analysis of pressure curves enables quantification of contrast agent effects on various heart characteristics. In Figure 3.8, a comparison of heart rate, LVEDP and LVSP distributions before (blue) and after (orange) contrast agent is shown. Dashed lines are the accordingly fitted Gaussian distributions. Stated values are mean±std.

The average LVEDP was 17.60 mmHg (CI95% 17.54-17.66), with a volume-induced increase to 18.64 mmHg (CI95% 18.57-18.71) after ventriculography indicating a significant effect. As many as 20,402 patients (30.1%) had resting LVEDP greater than 20 mmHg, representing candidates with impaired diastolic function [Leistner et al., 2020]. Furthermore, the average patient heart rate accelerates from 73.3 bpm to 74.8 bpm and the peak systolic pressure decreases from 139 mmHg to 137 mmHg. All observed trends are in concordance with the toxic influence of contrast agents on heart functionality [Hasebroock and Serkova, 2009].

3.5 Assessment of biological variation of the LVEDP

The LVEDP was computed for each cardiac cycle within the period of time where patients have been instructed to stop breathing during cardiac catheterization. Outliers in terms of the pressure curve's width or height were removed beforehand and assessment was undertaken in the relatively stable situation of at least three consecutive heartbeats. The difference between

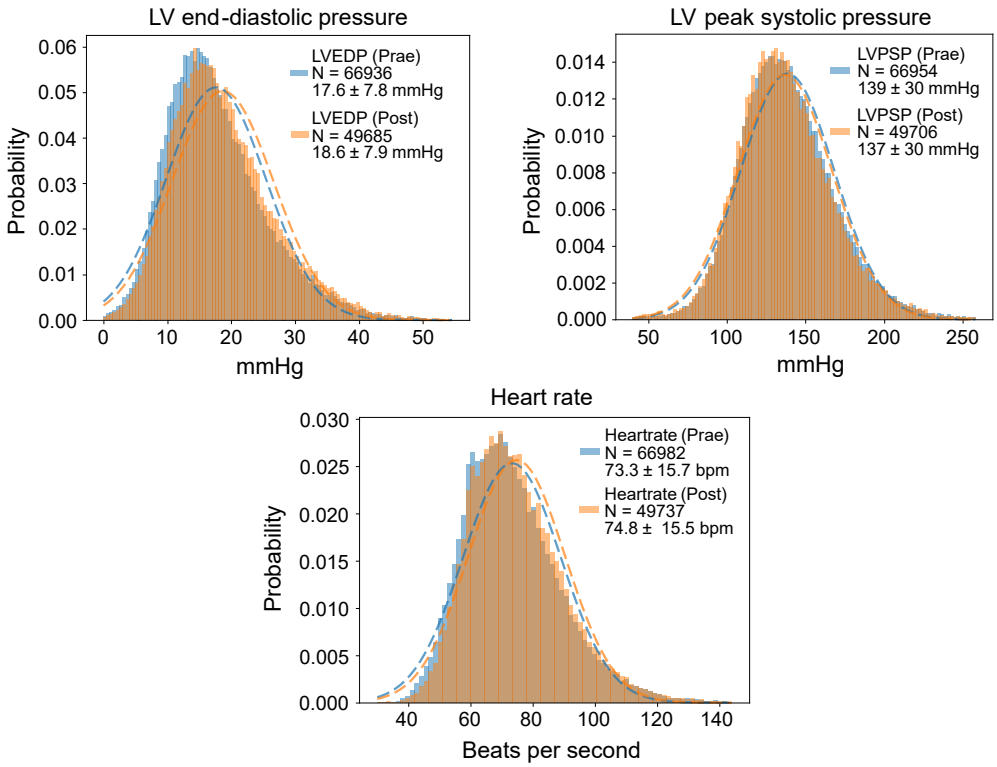


Figure 3.8: Heartrate, left ventricular end-diastolic pressure and left ventricular peak systolic pressure distributions before (Prae) and after (Post) ventriculography.

minimum and maximum LVEDP in the pre-determined interval was calculated to assess LVEDP's biovariability.

Even when using direct invasive measurements of LVEDP in stable patients, considerable variation within one individual can be seen and LVEDP may significantly change with each cardiac cycle. To underline this fact, the heartbeat-specific LVEDP-variability in cardiac catheterization curves was calculated over more than 183,000 processed LV pressure curves. The beat-to-beat variation per patient (difference between minimum and maximum LVEDP) was on average 7.91 mmHg with a standard deviation of 5.48 mmHg.

These results indicate that the LVEDP is a volatile parameter underlying considerable background noise. Hence, LVEDP prediction models (as developed in Chapter 5) are particularly difficult to create, as they should be competent to deal with noisy data. Furthermore, the quantized noise levels introduce theoretical performance limits for those prediction models.

4 Prediction of Clinical Phenotypes from Left-Ventricular Pressure Curves using Artificial Neural Networks

The detailed automated curve analysis enables various scientific opportunities. This chapter investigates the hidden information that can be derived from left ventricular pressure curves using artificial neural networks. Neural networks are known to be able to derive information, which is not accessible for humans from various data sources. Popular examples are the sex and cardiovascular risk factor prediction from retinal fundus photographs or detection of mRNA splicing sites via deep learning [Poplin et al., 2018, Jaganathan et al., 2019].

The investigated phenotypes are patient age, sex and the presence of a coronary artery disease (CAD). Coronary artery disease is a condition where plaque buildup narrows the arteries supplying blood to the heart, leading to reduced blood flow and heart insufficiency [McCullough, 2007]. CAD is diagnosed with a computerized tomography coronary angiogram. Thereby, a contrast agent is injected into the coronary arteries [Kariyanna et al., 2020]. If the presence or absence of CAD can be predicted alone from the left ventricular pressure, such a system could function as a gatekeeper for the coronary angiogram. If the angiogram is only performed on patients with positive CAD indication from the pressure curve, the amount of procedures could potentially be greatly reduced. The predictions of sex and age do not have such immediate applications. However, the AI-based analysis of the overall information content could benefit future research and the general comprehension of pressure curves.

4.1 Correlation of automatically extracted features with phenotypes

Before utilizing machine learning techniques to find complex patterns, the stand-alone correlations of the derived features by the pressure curve algorithm with the mentioned phenotypes are investigated.

In Table 4.1, the correlations of extracted features with sex, age and coronary artery disease are shown. The binary classification tasks are evaluated in terms of receiver operating characteristic area under the curve (ROC AUC) [Fawcett, 2006]. The correlation between features and patient age is quantified using the Pearson correlation coefficient [Lee Rodgers and Nicewander, 1988]. All observed ROC AUC values were smaller than 0.6 for the sex and CAD, indicating a low discrimination capability [Yang and Berdine, 2017]. Also, all the Pearson correlations with age were smaller than 0.3, displaying a weak relationship [Akoglu, 2018].

The peak systolic pressure showed the highest AUC with sex. Women had a higher peak systolic pressure at rest than men. Additionally, the peak systolic pressure also had the second-highest correlation with patient age. Surprisingly, the peak receding time has the highest

Phenotype Metric	Sex ROC AUC	Age Pearson	CAD ROC AUC
Peak systolic pressure	0.586	0.242	0.541
Peak integral	0.583	0.14	0.507
Peak prominence	0.58	0.162	0.518
Peak rise time	0.571	0.231	0.51
Receding slope	0.569	0.051	0.535
Peak width	0.562	0.042	0.513
End-diastolic pressure	0.551	0.017	0.56
Heart rate	0.541	-0.072	0.508
Minimum diastolic pressure	0.538	0.109	0.553
Rising slope	0.505	-0.054	0.508
Peak receding time	0.504	0.283	0.565

Table 4.1: Correlation of automatically extracted features with patient phenotypes.

correlation with age and the highest discernment with CAD, although carrying the lowest amount of information about the patient's sex. In particular, in patients with CAD, the pressure curve needed more time to recede from 75 % to 25 % pressure after the ventricle contraction.

4.2 Prediction of age, sex and coronary artery disease from left ventricular pressure curves

In the previous section, only weak correlations with phenotypes were found. Now, the pattern matching capabilities of AI are leveraged to substantially increase the performance. Two different approaches were investigated. First, the extracted features are used as inputs for the artificial neural network. Secondly, 1D-CNNs were utilized to predict the phenotypes directly from the pressure curves.

Using the pressure curves directly as input for the CNN was not sufficient. Standardization greatly increased performance. Artifacts from the measurement device and aortic pressure intervals introduce disorder to the curve. However, the pressure curve algorithm (Chapter 3) was used to find intervals of "clean" ventricular pressure. The evaluation showed that most examinations had such intervals with a duration of more than two seconds. The input for the CNNs was accordingly defined as two seconds of clean ventricular pressure intervals. The time series input was further standardized so that it always starts at the point of minimum diastolic pressure.

The models were trained and benchmarked using 5-fold cross-validation. For each phenotype, several neural network architectures were evaluated with varying numbers of network layers and layer characteristics (activation function, batch normalization, number of filters/neurons, Dropout). The performances from the best respective models were reported.

The best feature-based neural networks achieved an ROC AUC of 0.71 predicting sex, a Pearson correlation of 0.47 with patient age and an ROC AUC of 0.66 classifying coronary artery disease (Table 4.5). This is already an improvement in comparison to the individual feature correlations (Section 4.1).

However, if the AI models have even more freedom to find arbitrary patterns, as realized with

the preprocessed input curves instead of the feature-based approach, all performance indicators increase further. The trained CNN was able to predict sex with an AUC of 0.78 (CI95% 0.77-0.78) and patient's age with a mean absolute error of 7.86 years (CI95% 7.77-7.95) and a Pearson correlation of 0.57 (CI95% 0.56-0.57) (Table 4.2). The accuracy when giving the model a margin

Prediction task	Number of samples	ROC AUC	Accuracy
Sex	33,405	0.78 (CI95% 0.77-0.78)	0.71 (CI95% 0.70-0.72)
CAD	24,005	0.70 (CI95% 0.70-0.71)	0.65 (CI95% 0.64-0.66)
Age	33,461	**Pearson correlation** 0.57 (CI95% 0.56-0.57) **Mean absolute error [years]** 7.86 (CI95% 7.77-7.95)	**Age ± 10 years accuracy** 0.71 (CI95% 0.70-0.71) **Age ± 5 years accuracy** 0.41 (CI95% 0.40-0.41)

Table 4.2: Predicting phenotypes from LV pressure curves utilizing convolutional neural networks.

of ±10 years on the age prediction was 0.71 (CI95% 0.70-0.71). If the margin was only ±5 years, the according accuracy was 0.41 (CI95% 0.40-0.41). The scatterplot of predicted versus actual patient age is presented in Figure 4.1. A more clinically relevant example was the predictive capability for the presence of coronary artery disease (AUC=0.70; CI95% 0.70-0.71), defined as ≥50% stenosis of a major coronary artery [Libby and Theroux, 2005].

Attention mapping with Grad-CAM showed that the systolic peak was most relevant for the sex prediction. This is concordant with the finding that the automatically extracted peak systolic pressure showed the strongest correlation with sex [Selvaraju et al., 2017]. For both other phenotypes no significant trends were discovered.

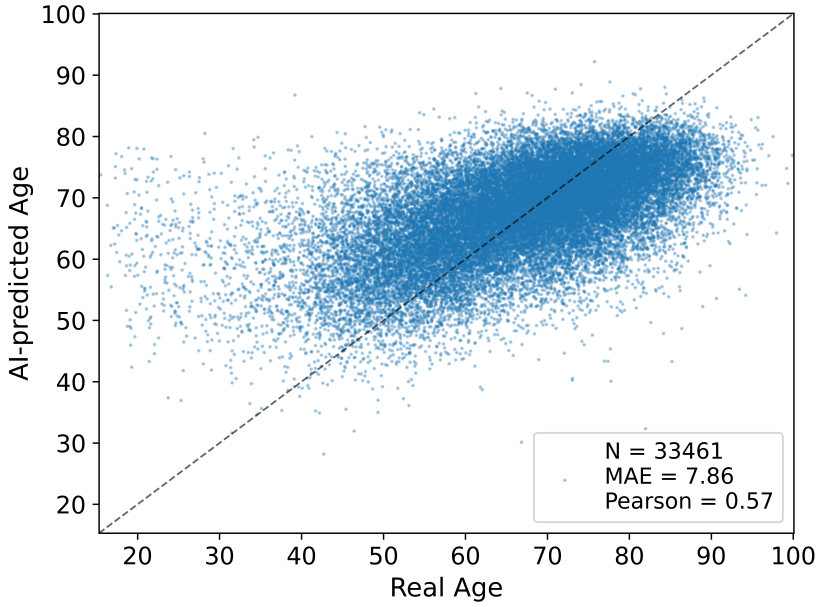


Figure 4.1: AI-predicted age versus actual patient age.

The circumstances of the age predictions were further investigated. Therefore, the cohort was divided into two separate groups. The first group included patients in which the AI age prediction was more than 10 years older than the actual patient age. Whereas the second group comprised patients with a predicted age of at least 10 years younger than the actual age. The

group of patients categorized by the AI as at least 10 years older had an actual mean age of 45.8 years. The group labeled at least 10 years younger had a mean age of 76.5 years. Due to the large age discrepancy, the patient groups could not be compared directly. To ensure a fair comparison, cohorts of the same age were sampled from the dataset according to the observed age distribution (Table 4.3).

For those patients, comprehensive information on cardiovascular health was collected from the electronic health records. The cohort of patients with at least 10 years older predicted age had significantly higher NT-proBNP levels than the according age-matched cohort [Hall, 2005]. The left-ventricular ejection fraction (LV-EF) was significantly lowered and the occurrence of dilated cardiomyopathy (DCM) was elevated. All these factors indicate an overall deteriorated cardiovascular health compared to the age-matched cohort.

Variable	AI predicted >10 years older than actual age	Age match	p-value	AI predicted >10 years younger than actual age	Age match	p-value
Age (years)	45.8±12.6	45.8±12.6	1.00	76.5±7.1	76.5±7.1	1.00
Sex (%male)	71.3	74.3	0.225	77.7	74.6	0.286
NT-proBNP mean±std (ng/l)	1982±6300	1410±4024	**0.0489**	1801±3513	2937±7273	**0.0045**
NT-proBNP median (ng/l)	437	348	-	454	750	-
LV-EF mean±std (%)	51.8±13.9	54.0±12.1	**0.0021**	55.6±11.2	54.1±12.9	0.0692
Atrial fibrillation (%)	9.4	7.8	0.3183	24.8	27.2	0.419
DCM (%)	24.6	19.6	**0.0294**	6.8	8.0	0.4993
hsTnT mean±std (ng/l)	274±553	293±583	0.5426	232±560	155±409	**0.0188**
hsTnT median (ng/l)	31.0	28.0	-	29.0	28.0	-
CAD (%)	66.1	63.8	0.4020	95.4	95.3	0.9630
Arterial hypertension (%)	60.8	58.1	0.3219	83.8	87.0	0.1933
Heart attack (%)	17.6	19.1	0.4921	25.6	27.4	0.4951

Table 4.3: Comparing patient characteristics and cardiovascular health between AI-predicted age groups.

Following the same pattern, the group of patients with an age predicted at least 10 years younger had improved cardiovascular health compared to patients from the same age distribution. This was manifested in significantly lowered Nt-proBNP levels and improved ejection fraction. The measured hsTnT levels were higher in the group of patients with younger AI-predicted age, indicting the contrary [Gaggin and Januzzi Jr, 2013]. However, the hsTnT median was qualitatively equally leveled in both groups.

In summary, the results suggest that the AI age prediction is influenced by the cardiac health status of the patient, and thus, the model potentially categorizes patients with a higher biological age as older [Jylhävä et al., 2017].

4.3 Phenotype prediction based on Fourier transformed pressure curves

Fourier transformation is a mathematical procedure that represents a function as a combination of sinusoidal functions with varying frequencies, phases and amplitudes. Fourier transformation allows to move between the time and frequency domain, providing a powerful way to analyze signals and functions [Bochner and Chandrasekharan, 1949]. By expressing complex functions as a sum of simpler trigonometric functions, it enables the understanding of intricate mathematical

structures in terms of their frequency components [Kaiser and Hudgins, 1994]. The Fourier transform $\hat{f}$ of an integrable function $f : \mathbb{R} \to \mathbb{C}$ is defined as

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} dx. \tag{4.1}$$

The Discrete Fourier Transform (DFT) is used for sampled, discrete data points. It transforms a sequence of discrete numbers $x = (x_0, \dots, x_{N-1}) \in \mathbb{C}$ into the Fourier-transformed sequence of complex numbers $X = (X_0, \dots, X_{N-1})$, which is defined by

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn}. \tag{4.2}$$

The DFT is usually implemented utilizing the Fast Fourier Transform (FFT) algorithm [Brigham, 1988]. Since characteristics of the LV pressure curve are strongly influenced by the patient's heart rate, a frequency-based analysis is a natural choice.

The preprocessed clean intervals had a length of two seconds. The sampling rate from the manometer was 250 Hz, equaling one measurement every 4 ms. Hence, each curve consisted of 500 data points. If the DFT is calculated for purely real input, the output is Hermitian-symmetric [Cooley and Tukey, 1965, Press et al., 1986]. In other words, the negative frequency terms are merely the complex conjugates of the corresponding positive frequency terms ($X_{N-k} = \overline{X_k}$), so that the negative frequency terms are redundant and do not provide additional information. Accordingly, the full FFT spectrum is characterized by only 251 coefficients ($\frac{N}{2} + 1$).

4.3.1 Model evaluation on reconstructed curves from partial FFT spectra

An elegant method to estimate the information content of specific DFT coefficients regarding the prediction task is to evaluate the existing model on reconstructed curves. The convolutional neural network for the sex prediction task, using LV pressure curves as the input, achieved an ROC AUC of 0.7765. This benchmark was obtained by evaluating the network on the test set.

However, it is possible to reconstruct the pressure curves based on partial spectra. In Figure 4.2, curve reconstructions with varying numbers of included coefficients are shown. The first 10 Fourier coefficients already provide a qualitatively accurate reconstruction. A reconstruction with less than half the spectrum (70 coefficients) looks nearly indistinguishable from the original time series.

The unchanged model was evaluated on the reconstructed test set curves. The reconstruction with 100 or more coefficients produced the exact same performance. The reconstruction with the first 70 coefficients even resulted in a minor performance increase. This highlights the applicability of low-pass filters for noise reduction [Gasquet and Witomski, 2013]. The reduction of noise is another layer of data standardization that can lead to improved results in machine learning tasks. By using 30 coefficients, the ROC AUC still only drops by 1.96 %. With only 5 coefficients the ROC AUC is still above 0.6. This may be partially explained by the fact that the peak systolic pressure (highest feature correlation, Section 4.1) is already manifested by a low amount of Fourier coefficients. However, it is intriguing that the neural network is capable of deriving the sex from highly deprived curves considerably better than baseline (random guessing) and any previously observed feature correlation.

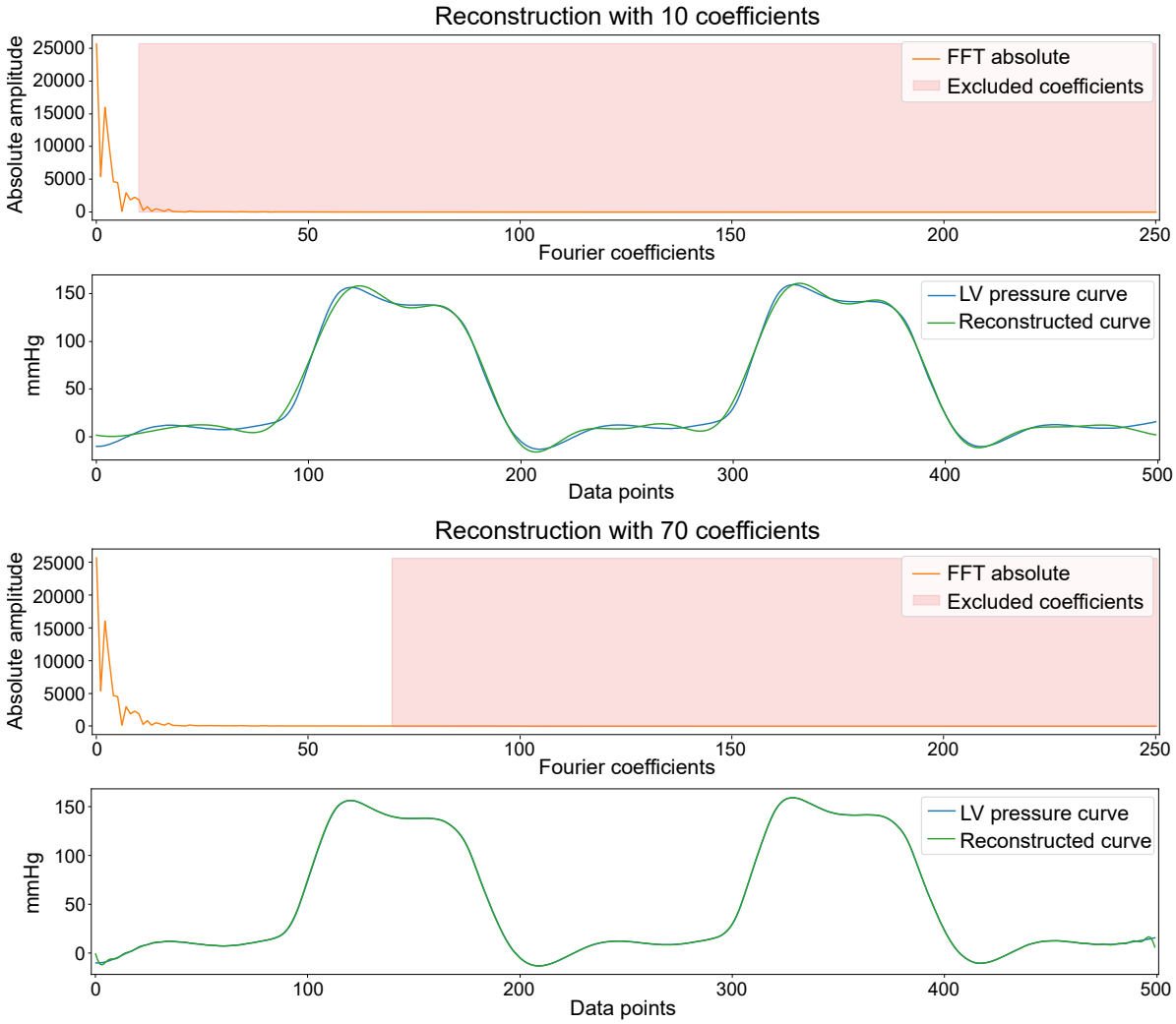


Figure 4.2: LV pressure curve reconstruction using partial FFT spectra.

The reconstruction approach could also function as a potential data augmentation technique in neural network training pipelines.

In conclusion, this experiment indicates that the Fourier coefficients at the end of the spectrum contribute only minor information. This enables the option to discard parts of the FFT spectrum and reduce the number of parameters (problem size) considerably. Subsequent machine learning ventures could benefit from the reduced problem size in terms of training/inference time and energy cost [García-Martín et al., 2019].

It is still imaginable that newly trained machine learning models could utilize the information contained in these coefficients. However, the experiments conducted in this section with normal time series input constitute lower bounds on performance for FFT models trained on partial spectra, since the information to match the CNN performance is potentially available.

4.3.2 Scaling methods for FFT spectra

Normalization of input data is an important component for deep learning applications, with a positive impact on benchmark performance [Jayalakshmi and Santhakumaran, 2011]. Often, input parameters are preprocessed to be in the range $[0, 1]$. Dealing with FFT spectra is particularly

FFT reconstruction with	Sex ROC AUC
251 coefficients	0.7765
200 coefficients	0.7765
100 coefficients	0.7765
70 coefficients	0.7769
50 coefficients	0.7734
40 coefficients	0.7634
30 coefficients	0.7613
20 coefficients	0.7383
10 coefficients	0.6470
5 coefficients	0.6059
Baseline CNN	0.7765

Table 4.4: Re-evaluation of the best CNN model using reconstructed curves based on partial FFT spectra for sex prediction.

difficult, since the Fourier coefficients deviate by orders of magnitude. Commonly, the first Fourier coefficient of left-ventricular pressure curves is easily greater than 30,000. In contrast, other coefficients may be smaller than one. Ordinary Min-Max normalization is potentially suboptimal, since smaller coefficients encoding information are divided by a large number [Larose, 2005]. This results in values very close to zero, which can not be properly distinguished by the artificial neural network

In order to solve the described problem, more sophisticated normalization techniques are needed. For left ventricular pressure curves, the first Fourier coefficient X_0 is the largest occurring value in the FFT spectrum (Figure 4.2), representing the sum over all data points values. The second peak in the FFT corresponds to the patient's heart rate, which is the dominant frequency in the spectrum. The remaining peaks are multiples of the heart frequency. The observed magnitudes of the peaks are more likely to follow an exponential rather than a linear relationship. Therefore, applying an exponential scaling function is at hand. In a series of experiments, multiple normalization methods were evaluated and compared.

Logarithmic scaling

To deal with the large difference in coefficients, the logarithm of each coefficient is calculated.

Exponential scaling

An exponential function $f(t) = a \cdot e^{-bt}$ is fitted via least-squares method [Kelley, 1999] to every individual curve. The medians of fitting parameters $\tilde{a}, \tilde{b}$ are calculated across the whole dataset. The final scaling function is given by inserting the parameter medians into the exponential function and dividing every FFT coefficient by the respective function value $X_{t,norm} = \frac{X_t}{\tilde{a} \cdot e^{-\tilde{b}t}}$ at position t.

Channel-wise normalization

In the two previous methods, a normalization function is applied to all coefficients. Therefore, the difference in magnitude of the first FFT coefficients compared to higher coefficients is tackled. However, early coefficients in the spectrum, not corresponding to the heart rate

frequency, may still be divided by a large number. In this method, the median value for each coefficient $\widetilde{X_t}$ is calculated across the whole dataset. The transformation is then given by $X_{t,norm} = \frac{X_t}{\widetilde{X_t}}$.

No normalization
Control experiments are also conducted with unprocessed FFT coefficients as neural network input (Min-Max scaling is still applied).

After the application of every described method, Min-Max scaling was additionally applied to ensure the final coefficients to be in the range $[0, 1]$.

4.3.3 Neural network architectures for FFT data

The utilization of convolutional layers is a leading paradigm in computer vision and time series analysis [Zhao et al., 2017]. In computer vision, the pixels are locally related, whereas in time series, the data points are temporally related. In a locally related setting multiple stacked convolutions can describe patterns that match complex objects or structures. In a temporal setting, the convolutions can detect specific changes or motifs over time.

In the FFT spectra, the data points are related in frequency space. In frequency space, the interpretation of convolutions is not as trivial. What does it mean to combine coefficients corresponding to adjacent frequencies? The modeling of features, which showed the highest correlations, such as peak systolic pressure (Section 4.1), can not be modeled as easily as in time-space.

To investigate the applicability of convolutional neural networks to FFTs, the experiments are conducted with networks based on convolutional layers and networks consisting of only fully-connected layers. Again, many different architectures are calculated for each scenario with varying numbers of layers, numbers of neurons / filter sizes / numbers of filters, residual connections, activation functions, loss functions and application of different regularization techniques.

4.3.4 Model evaluation predicting phenotypes from FFT spectra

A comprehensive analysis was conducted predicting patient age, sex or the presence of coronary artery disease based on FFT spectra from LV pressure curves. The FFT models were developed with the same training scheme as the CNNs operating on the time series data (Section 4.2). Different neural network archetypes and the already introduced various scaling methods were evaluated (Table 4.5). The best result was reported. The inputs for the models were the first 70 coefficients of the FFT spectrum, which is considerably less than the 500 data points used in the time series model. In terms of normalization, logarithmic scaling was found to be inferior to the three other scaling schemes. Overall, channel-wise normalization yielded the best results. However, there was no qualitative difference to exponential scaling and no normalization. In terms of architecture, the dense network architectures performed slightly better than the convolutional FFT networks. This indicates, that building features based on neighboring frequency proportions was not beneficial for the investigated prediction tasks.

Model Type	Sex		Age		CAD	
	ROC AUC	Accuracy	MAE	Pearson	ROC AUC	Accuracy
FFT Dense No normalization	0.74	0.68	8.70	0.45	0.68	0.63
FFT Dense Logarithmic scaling	0.71	0.65	8.98	0.39	0.66	0.62
FFT Dense Exponential scaling	0.74	0.68	8.72	0.44	0.68	0.63
FFT Dense Channel-wise normalization	0.74	0.68	8.69	0.45	0.68	0.64
FFT CNN No normalization	0.73	0.67	9.04	0.39	0.67	0.62
FFT CNN Logarithmic scaling	0.71	0.66	9.11	0.39	0.65	0.61
FFT CNN Exponential scaling	0.73	0.67	8.97	0.40	0.67	0.62
FFT CNN Channel-wise normalization	0.73	0.67	8.89	0.41	0.67	0.62
Feature-based NN	0.71	0.65	8.61	0.47	0.66	0.61
Baseline CNN	0.78	0.71	7.86	0.57	0.70	0.65

Table 4.5: Evaluation of different architectures and scaling methods to predict sex, age and diagnosis of a coronary artery disease from left ventricular pressure curves.

Also, in Table 4.5, a comparison of the networks operating with FFT spectra as input and the CNNs trained on the time series data is shown. The FFT approach resulted in a considerable performance decrease compared to the baseline CNN. However, it was still superior to the feature-based approach in terms of sex and CAD prediction.

Although the cardiac cycle is a highly recurrent and frequency-dependent entity, the use of FFT spectra was found to be inferior to normal CNN approaches. Moreover, the performance loss of the FFT compared to the normal time series approach is likely not a normalization problem. These findings may be strongly influenced by the actual prediction task. For other tasks, frequency-based features may be more expressive. Also, the combination of frequency- and time series-based features could be a promising approach, but it would increase the problem size.

5 Predicting Ventricular Pressure from Cardiac MRI (AI-LVEDP)

Intercardiac pressure provides important information on the heart's diastolic properties, is instrumental in deciding on cardiac surgeries and interventions such as valve replacements, and is predictive for negative clinical events [Aalaei-Andabili and Bavry, 2019, Brienesse et al., 2018]. To obtain the LVEDP, the pressure in the left heart chamber is measured invasively during cardiac catheterization. Thereby, a catheter is inserted through the patient's groin or arm and guided through the venous system to the heart. Although the procedure is routinely done and associated with low number of complications, the intervention is not completely risk-free and potentially induces stress and discomfort for the patient [Al-Hijji et al., 2019].

In this chapter, it is investigated to non-invasively predict left ventricular cardiac filling pressure from cardiac MRI. This developed AI model is referred to as AI-LVEDP.

5.1 Core cohort dataset

The retrospective study cohort consisted of 66,936 patients who underwent cardiac catheterization with more than 183,000 individual left-ventricular pressure measurements at the University Hospital of Heidelberg (Figure 1.1). Of these patients, 11,699 cardiac MRI examinations were recorded. The examinations included diagnostic and therapeutic as well as elective and emergency cases and were extracted from the University Clinic's IT infrastructure in accordance with the ethics vote (No S-158/2021). Data from the cMR examinations was extracted from the central PACS servers.

When deriving the cardiac filling pressure from the MRI, the most important prerequisite is that the LVEDP is the same or at least comparable at the time of the cardiac catheterization and at the time of the cardiac MRI. Strict filtering procedures were enforced to reduce the likelihood of significant changes to the filling pressures. The first precaution was to only select patients with both cMRI and cardiac catheterization performed within a timeframe of 30 days. 4,708 cMRI/cardiac catheterization couplets could be identified with both procedures within 30-day window. The mean time difference was 4.6±7.1 days. However, 55.5% of patients had both examinations within one day.

Next, diagnosis and summary reports of the cardiac catheterization were extracted from the hemodynamics/cathlab database. Based on this information, all emergency examinations (n=1,646), surgeries (n=183) and coronary interventions (n=2,019) were excluded from the dataset as these can be associated with higher fluctuations in filling pressures. Emergencies included (non-) ST-elevation myocardial infarction, ventricular fibrillation and Tako-Tsubo syndrome [Sinning et al., 2010]. The final core cohort dataset consisted of 1,437 examination pairs

Variable	Value
Number of examination pairs	1,437
Number of patients	1,400
Sex (% male)	69.1
Age (years) mean ± std	60.1 ± 15.7
Weight (kg) mean ± std	80.0 ± 15.9
NYHA (%)	
I	32.6
II	36.4
III	19.9
IV	11.1
Time between cardiac catheterization and cMRI	
Days between examinations mean ± std	4.6 ± 7.1
Cardiac catheterization	
LV-EDP (mmHg) mean ± std	16.8 ± 7.7
Cardiac MRI examination	
LV-EF (%) mean ± std	53.2 ± 14.5
LV mass (g) mean ± std	117.3 ± 45.9
LV-EDD (mm) mean ± std	52.8 ± 8.8
LV-ESD (mm) mean ± std	31.3 ± 12.8
LA-diameter (mm) mean ± std	38.3 ± 8.1

Table 5.1: Clinical characteristics of core cohort patients.

from 1,400 patients. 37 patients had two cMRI/cardiac catheterization couplets, all at least one year apart.

The recorded pressure curves were automatically analyzed by the in the scope of this dissertation developed algorithm described in Chapter 3. The automatically extracted LVEDP value is used as the ground truth for the regression task. Therefore, the ground truth is completely free of observer bias.

The core cohort is the development cohort for predicting LVEDP from cMRI (AI-LVEDP). The characterization of training data is essential to estimate potential limitations of AI systems [Johnson et al., 2021]. The mean age of patients in the core cohort was 60.1 years with a standard deviation of 15.7 years. The mean patient weight was 80.0±15.9 kg. 30.4 % experienced no cardiac-related limitations on physical activity (NYHA I), whereas 12.6 % already showed symptoms of heart failure at rest (NYHA IV) [Dolgin et al., 1994]. The automatically extracted LVEDP was on average 16.8 ± 7.7 mmHg. Additional patient characteristics can be found in Table 5.1.

Additionally, electronic health records were investigated to obtain a more detailed disease spectrum (Table 5.2). 4.5 % of patients had a diagnosed hypertrophic cardiomyopathy. The most prevalent cardiomyopathy was dilated cardiomyopathy, with a frequency of 25 %. The majority of patients (73.5 %) suffered from coronary artery disease. The occurrence of atrial fibrillation (14.3 %) is of particular interest, since commonly used LV filling pressure biomarkers from echocardiography can not be calculated during rapid and irregular beating of the atrium [Nagueh et al., 1996]. Furthermore, 14.3 % had diabetes mellitus.

The ethnicity of patients was not individually recorded. However, the expected ethnicity

Condition	Occurrence (%)
Hypertrophic cardiomyopathy	4.5
Restrictive cardiomyopathy	0.2
Arrhythmogenic right ventricular cardiomyopathy	0.1
Other cardiomyopathies	2.2
Dilated cardiomyopathy	25.0
Coronary artery disease	73.5
Atrial fibrillation	14.3
Arterial hypertension	66.7
Diabetes mellitus	13.4
Peripheral artery disease	0.1
Stroke	0.5

Table 5.2: Patient's disease spectrum in the core cohort.

spectrum of the dataset is represented by the southern German population. In the area covered by the University Hospital of Heidelberg, 30.9 % of inhabitants have a migration background. 50 % of those come from other EU states than Germany, predominately Italy and Greece. The majority of patients coming from outside the EU are from Turkey and the former Yugoslavia with corresponding ethnicities[1].

5.2 Model development of AI-LVEDP

The 2-chamber cine MRI view was selected for the input, as this method is frequently used in various cMRI protocols. Preliminary investigations in this study have shown that the 2-chamber view is the most promising compared to the short-axis and 3-/4-chamber view. When multiple 2-chamber MRIs were obtained during the same examination, the latter was used, assuming that the sequence was recorded again due to quality issues. The 2-chamber cine MRI shows the left ventricle and the left atrium of the heart. Also, the area of the mitral valve is clearly visible, which could be important as the area of the mitral valve is used to infer the filling pressure during echocardiography [Dagdelen et al., 2001]. A cine MRI consists of multiple frames obtained over multiple heart cycles. The final cine MRI in the core dataset consists of 35 frames representing one full heart cycle. All 35 images can not be obtained during a single heart cycle because of the limited image acquisition speed from the MRI machine (relaxation time).

The dataset was divided into a training, validation, and test dataset. Before the training phase, the neural network weights were initialized at random (Glorot initialization) [Glorot and Bengio, 2010]. During training, cMRI cine sequences were presented to the neural network. The network predicted the LVEDP and the prediction was compared to the actual value from the invasive hemodynamic measurement. The discrepancy was used to adjust network weight via backpropagation. The training process consisted of up to 500 epochs. The validation dataset was used to detect potential overfitting. Also, if poor model convergence was detected on the validation set, the network weights were reinitialized and the training process was repeated.

[1] https://www.baden-wuerttemberg.de/en/our-state/the-state-and-its-people, 15:42 09.03.2024

In addition to the cMRI, the AI-LVEDP received the patient's heart rate during the MRI examination, the patient's sex, age, and weight as additional input. All parameters are commonly obtained before or during the MRI examination. Especially heart rate and patient's weight are essential for the correct calibration of the cMRI procedure. The neural network architecture ensured a non-linear combination of the patient information with the cMRI features by leaving at least one hidden layer after feature concatenation. The used optimizer was Adam [Kingma and Ba, 2014]. To improve the expressiveness of the results, 5-fold cross-validation was used for performance estimation. The results were averaged over those five cross-validation models. For model development Python, Tensorflow, Pytorch and Keras were used [Abadi et al., 2016, Paszke et al., 2019, Chollet, 2018].

5.2.1 Data preprocessing

Based on each individual patient, the investigators and software plan the cMRI protocol. Each sequence might differ in terms of physical image size and resolution. The cMRI images were preprocessed so that the size of the cutout was physically 29.5x29.5 cm and the image dimension was 200x200 pixels. Hence, the physical scale on all cMRIs was consistent throughout the dataset. The final image resolution was adjusted using bi-cubic or Lanczos interpolation [Keys, 1981, Duchon, 1979]. In case that the whole obtained cutout was smaller than the required size, the image limits were padded with black pixels.

The greyscale values of the cMRI are coded in 8-bit. It was observed that sometimes, outlier pixels appeared very bright compared to the rest of the MRI. This made the overall MRI appear much darker and potentially inhibited optimal machine learning performance. In order to address this observation, each pixel was normalized to the 99 % percentile brightest pixel in each individual MRI sequence. Pixels brighter than the 99 % threshold were clipped at one. Hence, the resulting pixel values were all in the range $[0, 1]$.

Each sequence resembled a complete cardiac cycle. The standard cMRI protocol for 2-chamber cine MRIs obtained 35 frames. These frames were obtained throughout multiple heart cycles due to limits in frame acquisition speed. The start of the MRI sequences was gated by ECG (95% of cases) or pulse-oximetry (5%), for the ECG gating the R-wave is used as a trigger. The R-wave usually does not align with the heart-cycle-induced increase in oxygen saturation in the blood detected during the pulse-oximetry [Denslow and Buckles, 1993]. To mitigate this discrepancy, time synchronization was done manually for all pulse-oximetry-gated MRI sequences.

5.2.2 Data augmentation

Data augmentation during model training can significantly improve generalization [Mikołajczyk and Grochowski, 2018]. Thereby, the training images were slightly rotated, vertically/horizontally shifted at random for each training epoch. The specific augmentation operations are stated in Table 5.3.

For the random rotation, shift operations and brightness adjustment, a random value is uniformly sampled from the stated interval [Walpole et al., 1993]. In analogy, a random bit is generated to determine whether the instance is horizontally flipped. The horizontal flip is

Operation	Parameter space
Random rotation	$[-20°, 20°]$
Horizontal shift	$[-7\%, 7\%]$
Vertical shift	$[-7\%, 7\%]$
Brightness adjustment	$[-20\%, 20\%]$
Horizontal flip	Flip probability 50%

Table 5.3: Applied data augmentations during neural network training.

particularly important, since the 2CH MRI dataset included sequences recorded from the sagittal and coronal body plane of the patient [Ginat et al., 2011]. The sagittal sequence looks 'taken from the other side' compared to the coronary sequence. The flip operation mediates between those two views, resulting in higher generalization.

To ensure the consistency of these transformations throughout the image stack (time dimension) of the 2CH Cine MRI, each operation is applied with the same random value to each image in the sequence.

5.2.3 AI-LVEDP output calibration

After training; the neural networks had the tendency to predict LVEDP very conservatively. The ground truth spectrum of LVEDP values varied between 1 and 55 mmHg, whereas the neural network predictions only ranged from 7 to 34 mmHg. The model was trained to minimize the discrepancy between ground truth and prediction. Hence, it naturally tries to avoid large deviations by predicting mostly conservative LVEDP values centered around the data mean.

Some loss function tend to penalize large differences more than others, e.g. mean square loss penalizes outliers quadratically, whereas mean absolute loss only penalizes by absolute distance. Furthermore, strictly monotonously rising custom loss functions, which penalize larger variations even less harshly (e.g. logarithmic scaling) can be easily implemented.

However, it was not possible to adjust the prediction spectrum satisfactorily by applying the aforementioned techniques. To address the problem, a post-training calibration was implemented by utilizing the validation dataset. The test dataset remained untouched and was only used for the final performance benchmarking. With the ground truths labels of the validation dataset v_{truth} and the model predictions on the validation data v_{pred}, the transformation of the model output y is given by

$$y_{recal} = \frac{\mathrm{std}(v_{truth})}{\mathrm{std}(v_{pred})} \left(y - \mathrm{mean}(v_{pred}) + \mathrm{mean}(v_{truth}) \right), \tag{5.1}$$

where std denotes the standard deviation and mean the arithmetic mean of the list of values.

The calibration assures that the AI-LVEDP output distribution matches the ground truth LVEDP distribution of validation dataset in terms of mean and standard deviation. Therefore, the calibrated prediction spectrum is comparable to the LVEDP spectrum obtained from the cathlab.

5.3 Model evaluation on core cohort

A 3D-convolutional neural network (AI-LVEDP) was trained based on the core cohort to predict the LVEDP from cMRI. The model output is a single value representing the predicted LVEDP value in mmHg. The final cross-validated AI-LVEDP model could distinguish between elevated LVEDP (>20 mmHg) and normal LVEDP (<14 mmHg) with an accuracy of 0.75 (CI95% 0.70-0.80) and AUC of 0.78 (CI95% 0.75-0.81) based on ROC analysis (Table 5.6). The mean absolute error was 5.27 mmHg (CI95% 5.14-5.41) for individual predictions.

The dichotomization into the elevated versus normal LVEDP group represents relevant diagnosis windows used in clinical practice [Dokainish et al., 2008, Chemla et al., 2009, Posina et al., 2013]. Performance metrics for other thresholds mentioned in the literature are shown in Table 5.7. Because of the large biovariability of LVEDP discussed in Section 3.5 and the fact that most pressure measurement data points cluster around 15 mmHg, it is particularly difficult for AI-LVEDP and other commonly used biomarkers to distinguish between <15 and >15 mmHg.

5.3.1 Model architecture

The architecture of a neural network refers to the network topology, i.e. the arrangement of the neurons and their interconnection. The neural network architecture can have a great impact on the model performance [Bianco et al., 2018]. The final neural network architecture of AI-LVEDP is presented in Figure 5.1.

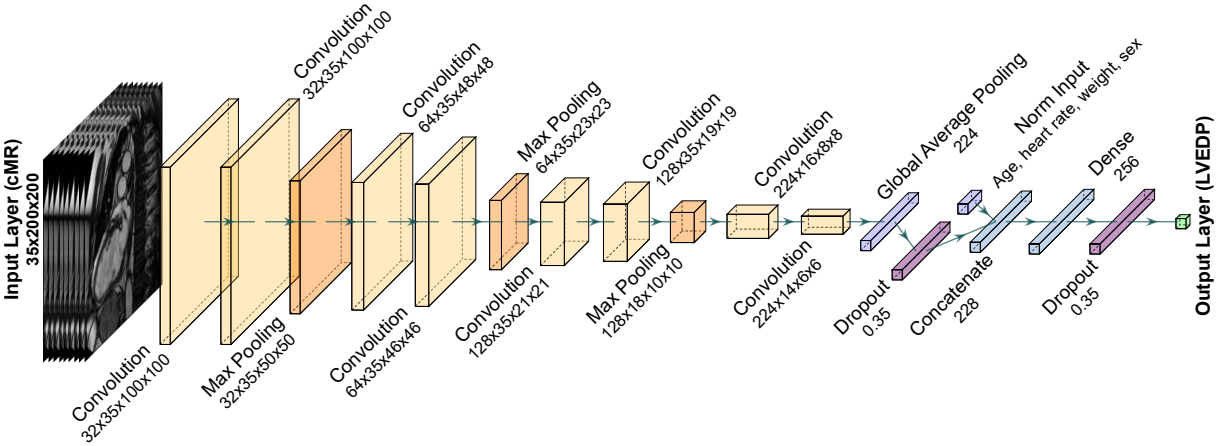


Figure 5.1: AI-LVEDP neural network architecture.

Several paradigms can be recognized by looking at the specified output dimensions of the hidden layers of AI-LVEDP. The first convolutional layers allow the network to build up complex 2D-feature maps. However, the network is not yet able to utilize temporal characteristics between MRI images. It has proven to be advantageous not to model local and temporal patterns simultaneously from the outset. In the rear layers of the network, the convolutions spanned over the time dimension. This enabled the learning of patterns that represent the change in cMRI images over time.

AI-LVEDP received the patient's age, weight, sex and heart rate during the cMRI as additional input. The additional input features were normalized to the range $[0, 1]$ to ensure optimal

information flow. The network architecture ensured that the additional input features could be combined in a non-linear fashion with latent features originating from the cMRI.

Several architectures were explored with varying numbers of layers, neurons and network topology based on the architecture shown in Figure 5.1. Furthermore, commonly used computer vision benchmark models were also explored instead of developing a completely custom topology. Architectures based on InceptionV3 received an ROC AUC of 0.77 (CI95% 0.74-0.80) [Szegedy et al., 2016]. Adapted networks from the EfficientNet family achieved an AUC of 0.78 (CI95% 0.75-0.80) [Tan and Le, 2019]. Likely, the limit of distinguishable pattern complexity was already reached. In order to distinguish more volatile patterns from noise, more data would be needed. However, it is also possible that the signal originating from the MRIs has already been exhausted with regard to the LVEDP. A study that includes significantly more data could be used to estimate whether all the relevant information contained in the cMRIs has been utilized.

The results with EfficientNet and Inception were qualitatively indistinguishable from the benchmark of the final AI-LVEDP model. However, the custom architecture was more resource-efficient in terms of the number of trainable parameters and execution time and was therefore preferred. Moreover, especially Inception was difficult to train. Sometimes the network was converging to a poor local minimum (detected by evaluation on the validation set) requiring reinitialization and retraining. This indicates that AI-LVEDP had advantages in terms of robustness.

The modeling of the time dimension was finally done by 3D-convolutions. Prior trials modeling the cMRI time dimension based on Long short-term memory (LSTM) modules yielded significantly inferior results [Hochreiter and Schmidhuber, 1997]. Experimentation with vision transformers showed more promising results [Dosovitskiy et al., 2020]. This could be an interesting approach for future research.

5.3.2 Analysis of AI-LVEDP predictions and NT-proBNP levels

N-terminal pro b-type natriuretic peptide (NT-proBNP) is a biomarker used extensively in cardiology for several purposes [Mueller et al., 2007]. NT-proBNP is a widely used biomarker for heart failure and is used for risk stratification for patients with acute coronary syndromes (ACS)[McKie and Burnett, 2016, Hall, 2005, Zdravkovic et al., 2013].

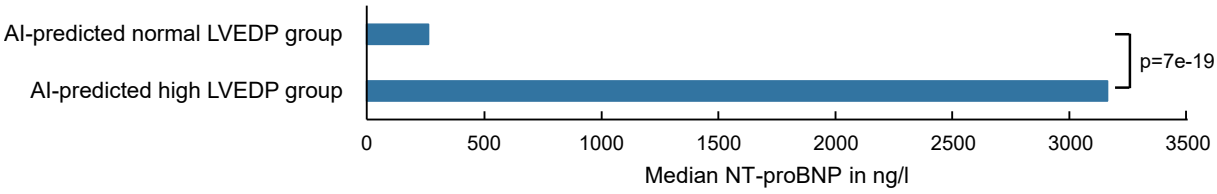


Figure 5.2: NT-proBNP levels in AI-predicted elevated and normal LVEDP groups.

NT-proBNP concentrations are measured via blood tests. Available measurements from patients were compared between the subgroup of individuals with an AI-predicted increased LVEDP and those patients with predictably normal pressure. The median NT-proBNP serum concentration was 263 ng/l in the normal LVEDP cohort and 3,156 ng/l in the increased LVEDP group (Figure 5.2). A t-test showed that the concentration differences were statistically significant (p-value = 7e-19). This further highlights the capabilities of AI-LVEDP dichotomizing patients in biologically and clinically relevant groups.

The Log(NT-proBNP) levels correlated with the ground truth LVEDP with a Pearson correlation of 0.41, which is slightly lower than the AI-LVEDP predictions (Pearson = 0.46).

5.3.3 Subgroup analysis

To test AI models for potential bias in the predictions, benchmarking performance for various patient subgroups is essential [Wang et al., 2007]. For this purpose, the trained AI-LVEDP was evaluated for specific subgroups of the test dataset (e.g. only female patients). The results were accumulated across all cross-validation folds. In Table 5.4, a comprehensive subgroup analysis investigating the impact of various patient characteristics is shown, thus revealing possible model biases.

The classification performance was homogeneous across age groups, sex and diagnosis of atrial fibrillation. The distribution of left ventricular ejection fraction (LV-EF) was as follows: LV-EF>50: n=996, 40≤LV-EF≤50: n=187, LV-EF<40: n=231, showing that the largest fraction of cases was in the normal LV-EF group and patients in this group were considered heart failure patients with preserved ejection fraction (HF-pEF) if they had elevated LVEDP. In the different

Subgroup	**Age (years)**		**Sex**		**LV-EF (%)**			**Afib**	
Cut-off	≥60	<60	male	female	>50	40-49	<40	yes	no
AUC	0.78	0.78	0.79	0.77	0.72	0.78	0.83	0.77	0.78

Subgroup	**Days between MRI and CC**			**NT-proBNP (ng/l)**		**Diuretics changed**	
Cut-off	≤1	>1	>10	>125	≤125	yes	no
AUC	0.77	0.80	0.75	0.78	0.74	0.79	0.76

Table 5.4: AI-LVEDP model performance in different patient subgroups.

LV-EF groups, there was a slightly reduced performance in the normal group (AUC=0.72), which is the group that is also especially challenging for the human experts (AUC=0.64).

Most patients had the cardiac catheterization and the cMRI within one day. The corresponding AUC for this time window was 0.77. All cases outside the one-day window resulted in an AUC of 0.8. In the subgroup of procedures that are more than 10 days apart, the AUC decreases to 0.75. This indicates the introduction of additional noise due to the larger time difference.

Next, differences in NT-proBNP subgroups are investigated. In patients with blood concentrations >125 ng/l the AUC was 0.78. In the reciprocal group with ≤125 ng/l, the performance declined to 0.74. This is in concordance with the trend already observed in the ejection fraction groups that AI-LVEDP tends to benchmark higher in groups with less healthy patients (low EF, high NT-proBNP).

Diuretics play a crucial role in the management of various conditions in cardiology by promoting the excretion of excess sodium and water from the body, thereby reducing blood volume and subsequently lowering blood pressure [Mullens et al., 2019, Felker et al., 2020]. Hence, diuretics likely have a direct effect on cardiac filling pressure.

To understand whether diuretics have any negative impact on the prediction by affecting the filling pressures the two assessments, 2501 doctoral letters (electronic health records) were

manually screened for the use and change of diuretics during the time window, as no structured information on diuretics was available. 42.3 % of patients had any diuretics including aldosterone-antagonists [Garthwaite and McMahon, 2004]. A total of 25 % of patients had a change in the diuretics medication. The AI performance remained robust in the diuretics group (AUC=0.79), irrespective of any changes in the dose or type of diuretics within the time window between the cMRI and hemodynamic assessment. The AUC in the cases in which the diuretic medication did not change or was completely absent was 0.76. This analysis suggests a rather negligible influence of diuretics on LVEDP compared to other factors.

5.4 External validation

External validation is often seen as the ultimate test to conclusively evaluate the safety, reliability and generalizability of a machine learning model [Youssef et al., 2023]. Internal cross-validation is not sufficient [Keevers, 2019]. Often, prediction models are too specialized for certain conditions and protocols in the development cohort and fail to reproduce applicability on reproduction cohorts [Data et al., 2016, Collins and Moons, 2019]. However, some argue that there is no such thing as a validated prediction model, as protocols and patient spectra change over time [Van Calster et al., 2023]. Hence, a careful monitoring of model performance over time is required in order to detect concept drifts and to ensure maximum reliability.

5.4.1 Validation cohort dataset

To access the generalizability of AI-LVEDP, both external validation and observation of model performance over time was conducted. AI-LVEDP was tested in three independent validation cohorts using the same model without retraining.

The external cohorts were derived from two centers, one university hospital (Ulm) and one larger-scale regional hospital (Robert-Bosch Hospital, Stuttgart). Both centers are experienced cardiac imaging centers with in-house cMR machines, allowing for narrow time windows between invasive and non-invasive phenotyping. From an initial set of 1,800 invasive and 2,200 cMR-examinations, matching for the timeframe of a maximum of 30 days and application of exclusion criteria (Section 5.1) resulted in n=113 cMR-hemodynamics couplets from Ulm and n=240 cMR-hemodynamics couplets from Stuttgart.

Furthermore, a current internal validation cohort from the University Hospital of Heidelberg was realized to observe model performance over time. This cohort was temporally separated from the core cohort and comprised n=903 patients from the years 2019-2023, whereas the development cohort included patients from the years 2006-2019.

A comparison of patient characteristics for all cohorts is shown in Table 5.5. Patients in the core cohort and external validation cohorts had an average age close to 60 years. The more recent validation cohort from Heidelberg had an average age of 65.6 years suggesting an aging patient collective. The mean weight was around 80 kg in all centers. The ratio of male patients is increased in all centers, which is consistent with known sex differences in patients with heart failure [Strömberg and Mårtensson, 2003].

The MRI scanner vendors, models and field strengths differed across the three centers, which

Variable	Core cohort	Ulm	Stuttgart	Heidelberg
Number of patients (n)	1,437	113	240	903
Sex (% male)	69.1	65.5	67.5	73.3
Age (years)	60.1 ± 15.7	59.9 ± 17.0	57.4 ± 14.2	65.6 ± 14.4
Weight (kg)	80.0 ± 15.9	81.2 ± 17.4	79.9 ± 15.1	80.9 ± 16.1
NYHA (%)				
I	32.6	37.6	13.5	29.1
II	36.4	23.9	43.9	44.1
III	19.9	33.9	31.6	23.5
IV	11.1	4.6	11.0	3.3
Days between cardiac catheterization and cardiac MRI	4.6 ± 7.1	7.7 ± 8.5	5.1 ± 10.0	3.9 ± 6.6
Cardiac catheterization				
LV-EDP (mmHg)	16.8 ± 7.7	18.9 ± 8.3	20.6 ± 9.0	16.0 ± 7.5
Cardiac MRI examination				
LV-EF (%)	53.2 ± 14.5	56.3 ± 15.4	43.5 ± 18.8	51.8 ± 14.5
LV mass (g)	117.3 ± 45.9	125.2 ± 53.7	-	118.7 ± 41.6
LV-EDD (mm)	52.8 ± 8.8	53.3 ± 9.0	57.1 ± 10.1	51.4 ± 8.4
LV-ESD (mm)	31.3 ± 12.8	36.6 ± 10.5	-	35.8 ± 10.8
Left atrial diameter (mm)	38.3 ± 8.1	44.1 ± 8.1	38.1 ± 8.3	40.3 ± 8.9
Cardiac MRI model, vendor and magnetic field strength	Philips Medical Systems Achieva 1.5T 59.6% of cases Philips Medical Systems Ingenia CX 1.5T 18.4% of cases Philips Medical Systems Ingenia 3T 14.9% of cases Philips Medical Systems Intera Achieva 1.5T 7.1% of cases	Philips Medical Systems Achieva 1.5T 100% of cases	SIEMENS Sonata 1.5T 70% of cases SIEMENS Aera 1.5T 30% of cases	Philips Medical Systems Achieva 1.5T 29.6% of cases Philips Medical Systems Ingenia CX 1.5T 36.2% of cases Philips Medical Systems Ingenia 3T 30.7% of cases Philips Medical Systems Intera Achieva 1.5T 3.5% of cases

Table 5.5: Overview of core cohort characteristics and validation cohorts. Values are stated as mean ± std.

introduced additional heterogeneity and made generalization more difficult. Heidelberg and Ulm used on Philips machines, whereas Stuttgart relied on Siemens scanners. The cMRI protocols in all centers greatly varied, which introduced challenges. The cMRI images used for model training were 2-chamber cine sequences consisting of 35 frames. However, other centers greatly utilized other framerates ranging from 15 to 40 images per heart cycle. To overcome this problem, Real-time intermediate flow estimation RIFE was used for up/downsampling of framerates to match the development cohort [Huang et al., 2020]. LV mass and LV ESD were not routinely assessed during the cMRI in Stuttgart, which further illustrates the differences in the protocols.

The time window distribution for cMRI and hemodynamics is shown in Figure 5.3. The distributions are displayed with kernel density estimation [Chen, 2017]. The core cohort and the Heidelberg validation cohort both usually had cardiac catheterization before the cMRI, whereas in Ulm, the majority of MRI procedures were conducted before hemodynamics. Ulm also had the highest average time difference between the procedures at 7.8 days. In the center of Stuttgart, the distribution is balanced and resembles a normal distribution.

The mean LVEDP observed in the core cohort was 16.8 mmHg, in Ulm 18.9 mmHg and 16.0 mmHg in the validation cohort from Heidelberg (Figure 5.4). The mean LVEDP seen in patients from Stuttgart was as high as 20.6 mmHg. Since an elevated LVEDP is an indication of impaired cardiac health, the Stuttgart cohort represents a more severely ill patient population

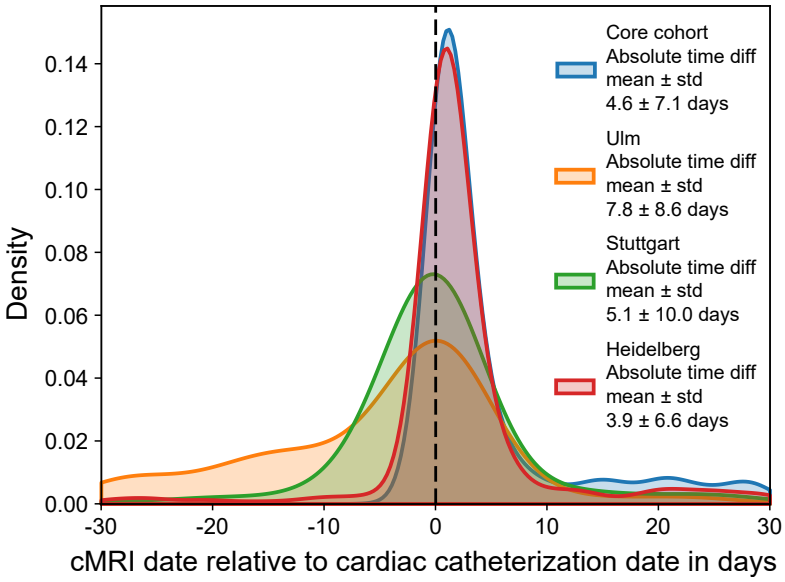


Figure 5.3: Distribution of time differences between cardiac catheterization and cMRI examinations for all cohorts.

compared to other centers. This is additionally reflected in the significantly lower ejection fraction.

5.4.2 Model evaluation on validation cohorts

The fixed AI-LVEDP model was evaluated on the validation cohorts without any adaptation. The results, including various metrics, are presented in Table 5.6. The artificial intelligence agent was capable of predicting LVEDP (normal vs. elevated) with almost equal performance compared to the core cohort (AUC = 0.78): 0.71 (University Hospital Ulm)and 0.75 (Robert-Bosch Hospital Stuttgart), AUC of 0.78 (University Hospital Heidelberg).

Metric	Core cohort	Heidelberg	Ulm	Stuttgart
ROC AUC	0.78 (0.75-0.81)	0.75 (0.73-0.77)	0.71 (0.68-0.74)	0.75 (0.73-0.78)
Accuracy	0.75 (0.70-0.80)	0.69 (0.65-0.74)	0.66 (0.61-0.70)	0.76 (0.75-0.77)
MAE (mmHg)	5.27 (5.14-5.41)	5.95 (5.84-6.06)	6.71 (6.65-6.77)	6.62 (6.49-6.76)
F1-Score	0.63 (0.57-0.71)	0.61 (0.56-0.69)	0.50 (0.41-0.59)	0.79 (0.76-0.82)
Specificity	0.76 (0.70-0.81)	0.59 (0.53-0.65)	0.40 (0.30-0.51)	0.81 (0.73-0.89)
Sensitivity	0.56 (0.45-0.66)	0.63 (0.57-0.70)	0.77 (0.67-0.87)	0.77 (0.75-0.80)

Table 5.6: AI-LVEDP model performance core cohort and validation cohorts. In brackets, the 95 % confidence intervals are stated.

The accuracy benchmarks ranged from 0.66 to 0.76. The mean absolute difference of individual predictions to the measured pressure values was 5.95 mmHg in Heidelberg. The mean absolute error in Ulm and Stuttgart was above 6.5 mmHg. This is partly explained by the higher ratio of patients with elevated LVEDP in those centers, allowing a larger margin of error.

A scatterplot of all data points is shown in Figure 5.5. The Pearson correlation coefficient between LVEDP ground truth and AI prediction was 0.46 on the core cohort, 0.36 in Ulm, 0.46 in Stuttgart and 0.48 in Heidelberg. The data from Ulm included the least number of patients. Hence, the confidence intervals for the metrics calculated on this data subset are the widest,

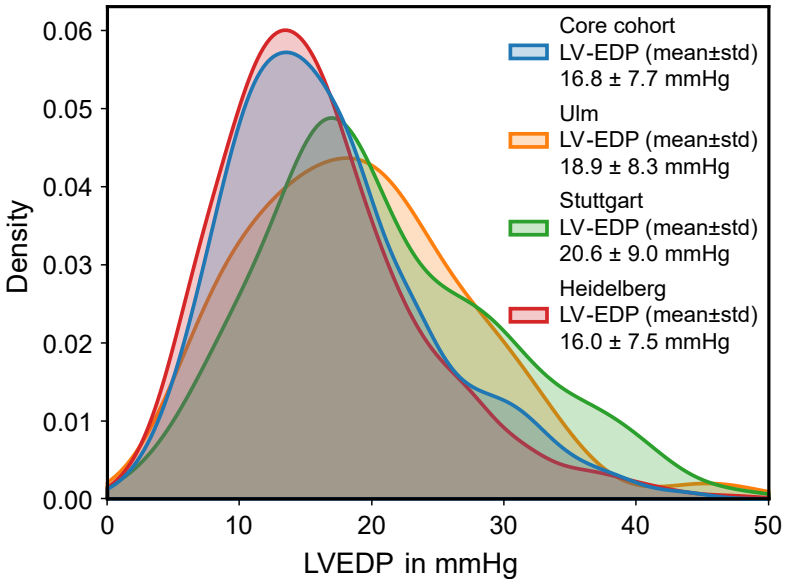


Figure 5.4: LVEDP distribution for the core cohort and all validation cohorts.

especially for the F1 score, specificity and sensitivity.

Despite the strong heterogeneity in the validation datasets, e.g. different MRI vendors/models/protocols and diverse patient spectra, AI-LVEDP maintains comparable performance. This is likely due to the methods applied during model development forcing the AI-LVEDP to generalize. The major techniques used to promote generalization were MRI preprocessing to establish data consistency and data augmentation to prevent overfitting.

5.5 Comparison to state-of-the-art "biomarkers" from echocardiography

Benchmarking performance is essential to establish a comprehension of the model's capabilities. However, it is similarly important to put these numbers into context. Therefore, comparing benchmarks to related and competing methods is crucial.

The gold standard for non-invasive assessment of LVEDP are currently biomarkers measured during echocardiography [Jones et al., 2021]. Both the E/A and E/E' quotients were routinely calculated within Doppler echocardiography examinations in the University Hospital Heidelberg. Pulse Doppler echocardiography is a procedure in which Doppler ultrasound is used to measure the speed and direction of blood flow by utilizing the Doppler effect [Oglat et al., 2018]. The E/A quotient refers to the ratio of early (E) to late (A) diastolic filling velocities of blood in the left ventricle of the heart [Cohen et al., 1996]. However, to calculate the E/E' quotient, the early diastolic mitral annular velocity (E') is used instead of the late diastolic filling velocities. The early diastolic mitral annular velocity reflects the velocity of movement of the mitral annulus (the base of the mitral valve) toward the apex of the heart, which is measured by tissue Doppler imaging [Ommen et al., 2000].

All available echocardiography examinations from patients in the core cohort were extracted from the research data warehouse, including calculated E/A and E/E' ratios. Due to the loss of the A-wave in patients with atrial fibrillation, the number of available E/A measurements (n=430)

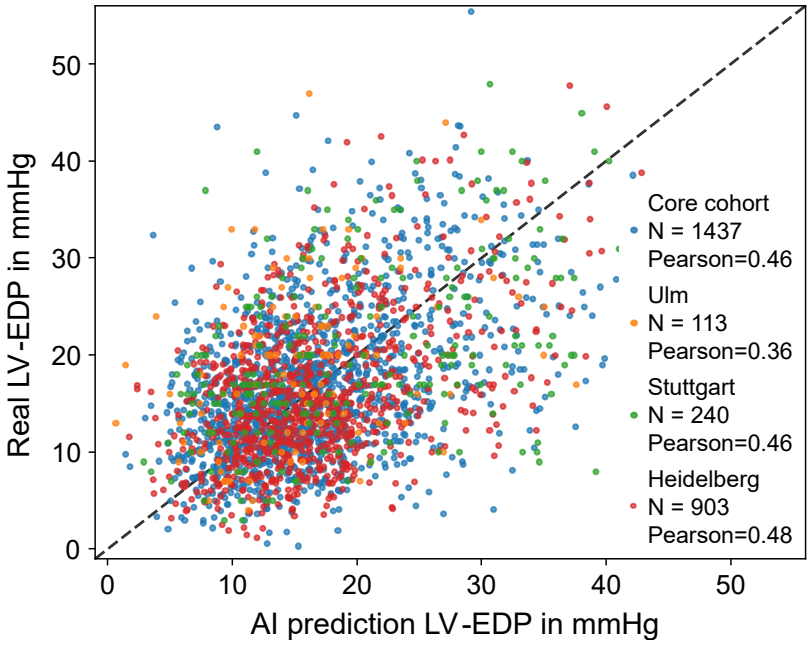


Figure 5.5: Individual AI-LVEDP predictions versus actual left-ventricular pressure for each center.

is less than the number of E/E' quotients (n=487). The E/E' ratio reached an AUC of 0.69, distinguishing between elevated LVEDP (>20 mmHg) and normal LVEDP (<14 mmHg), and the E/A ratio presented an AUC of 0.72 (Figure 5.6). Both echocardiographic parameters slightly underperformed when compared to AI-LVEDP (AUC=0.78) on the exact same patient subset. The ROC analyses were statistically compared using the DeLong test [Robin et al., 2018]. The resulting p-value for E/E' and AI-LVEDP was p=0.01. Comparing the ROC curves between E/A and AI-LVEDP yielded p=0.08, indicating a tendency towards superiority of the AI.

ROC AUC	AI-LVEDP	E/E'	E/A
<14 vs. >20	0.78	0.69	0.72
>12 mmHg	0.67	0.64	0.66
>14 mmHg	0.69	0.63	0.66
>15 mmHg	0.70	0.63	0.65
>20 mmHg	0.74	0.67	0.69

Table 5.7: Echo biomarkers versus AI-LVEDP for different classification thresholds.

Scatterplots of the measured LVEDP values and the echocardiography biomarkers result in a Pearson correlation of 0.39 for E/A and 0.41 for E/E'. Furthermore, confidence intervals including 95% of prediction were calculated (Figure 5.6 top panels). The slope and offset of the linear functions were determined by a linear fit utilizing the least squares method. The confidence interval borders, stated in absolute LVEDP deviation (mmHg), were then generated in such a way that 95 % of data points were comprised in the enclosed area. The calculated 95 % prediction intervals were ±16.8 mmHg for E/A and ±17.9 mmHg for E/E'. The equivalent confidence intervals obtained from AI-LVEDP were only ±15.45 mmHg, thus favoring the novel approach.

In Table 5.7, the performance of AI-LVEDP, E/E' and E/A for different classification thresholds is stated. In addition to the dichotomization into elevated and normal filling pressure, also singular

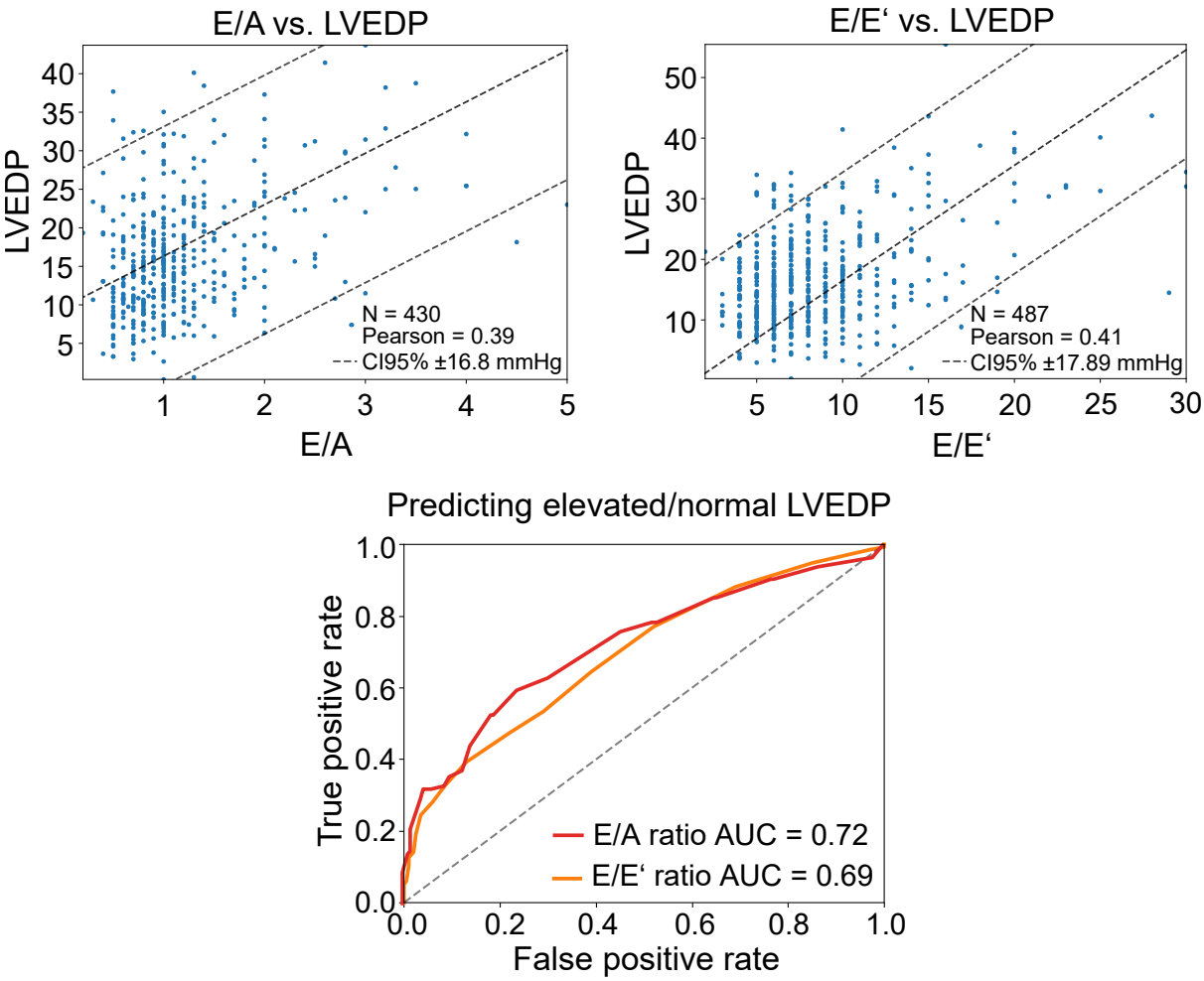


Figure 5.6: Echocardiography biomarker correlation with LVEDP.

thresholds are evaluated. Regardless of the exact scenario, AI-LVEDP scored higher AUCs than the echocardiography biomarkers. However, the magnitude of the difference varied. For classification thresholds close to 12 mmHg, where the majority of data points lay, the biovariability of LVEDP (Section 3.5) introduces the strongest effective noise levels. Thus, these tasks are more difficult for each biomarker and dampen their respective differences.

5.6 Comparison to cardiac MRI experts

Next, the performance of the deep learning model was compared to human experts. For this, the capabilities of four expert cardiologists to predict LV filling pressures from cMRI sequences were accessed. However, there is no evidence-based methodology of estimating LV-EDP from cardiac MRI imaging sequences, both for patients with reduced ejection fraction and patients with preserved ejection fraction.

300 cMRI/LVEDP couplets from the core cohort dataset were randomly selected and an app was developed with sequential visualization of the cMRIs and with the possibility for the observer to enter a numerical value for the LVEDP. In order to access the intra-observer reliability, 25 of the 300 cMRIs were duplicated and, therefore, presented to the experts twice. In addition to the cMRI, the experts were provided with the patient's age, weight, sex and heart rate during the

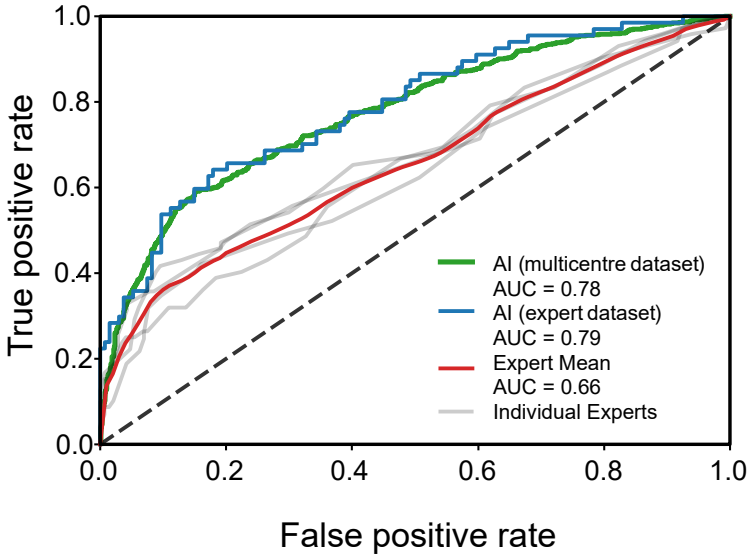


Figure 5.7: ROC analysis of expert cardiologists and comparison to AI-LVEDP.

MRI. Hence, the physicians were provided with exactly the same information as AI-LVEDP had. The experts evaluated the LVEDP independently using their extensive experience in cardiac imaging. There were no time constraints enforced during the assessment, so the physicians could individually decide how much time they needed for the prediction.

The experts reached an overall AUC of 0.66 (CI95% 0.62-0.69), distinguishing between elevated LVEDP (>20 mmHg) and normal LVEDP (<14 mmHg). The AI model showed a better performance with an AUC of 0.79 on the same data subset (Figure 5.7, blue curve) and an overall AUC of 0.78 on the complete multicenter dataset including all available patients (green curve).

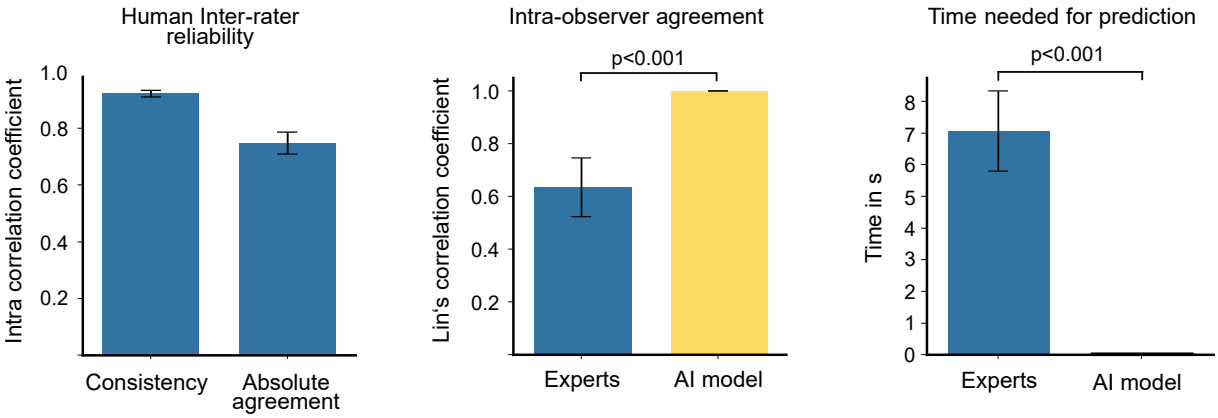


Figure 5.8: Inter- and intra-correlation coefficients for LVEDP prediction using cMRI by expert cardiologists and time required for prediction.

The experts spent an average time period of 7 seconds on each cMR sequence to make their prediction, whereas the AI needed only 0.02 seconds (Figure 5.8, right panel), showcasing the fast inference time of AI-LVEDP, enabling real-time real-time application possibilities. The hardware used for this benchmark was an Intel Core i9-9940X CPU paired with two NVIDIA TITAN RTX GPUs.

The intra-observer reliability was quantized by Lin's concordance correlation coeffi-

Predict LVEDP<14 and LVEDP>20 mmHg	**ROC AUC**
Expert 1	0.67
Expert 2	0.63
Expert 3	0.65
Expert 4	0.68
Experts mean ± std	0.66 ± 0.02
AI model	0.79
Time spent per cMRI	**mean ± std [s]**
Expert 1	8.8 ± 4.4
Expert 2	5.5 ± 3.5
Expert 3	7.7 ± 5.1
Expert 4	6.2 ± 4.0
Experts mean ± std	7.1 ± 1.3
AI model	0.02 ± 0.01
Intra-observer reliability	**Lin's concordance correlation coefficient**
Experts	0.63 (CI95%0.43-0.84)
AI model	1.00 (CI95%1.00-1.00)
Inter-observer reliability	
Absolute agreement (ICC3)	0.75 (CI95%0.71-0.79)

Table 5.8: Expert predictions of LVEDP based on cMRI.

cient [Lawrence and Lin, 1989]. The intra-observer reliability was 0.63 (CI95% 0.43-0.84). Since AI-LVEDP is a completely deterministic model, the intra-observer reliability of the deep learning model was 1. The inter-observer reliability was evaluated by intraclass correlation [Shrout and Fleiss, 1979]. The inter-observer agreement was 0.75 (CI95% 0.71-0.79) in terms of absolute agreement and 0.92 (CI95% 0.91-0.94) in terms of consistency (Figure 4B, middle panel and left panel, respectively). The individual expert performance measures are shown in Table 5.8.

Furthermore, the experts were interviewed to provide insights on their approach for estimating LVEDP from the cine cMR sequences. They stated that signs of atrial enlargement, ventricular dilation, paradox septal movements of the left ventricle, septal bounce and post-systolic septal displacement were important features influencing their estimation of LVEDP. Additionally, they took into consideration the presence of mitral prolaps, visual contrast perturbance and significant jets potentially representing mitral insufficiency.

5.7 Explaining AI-LVEDP predictions - Model introspection by attention mapping

After investigating human approaches for the determination of LVEDP using cardiac MRI, this section attempts to unveil the approach used by artificial intelligence. Especially in medicine, the ability to inspect and interpret functionality can improve trust and acceptance of machine learning models [Teng et al., 2022]. Attention mapping (Section 2.10) is an excellent instrument to provide a visual indication of a model's functionalism by highlighting the most important information for creating the prediction in the input image.

Attention maps were generated by systematically obscuring different regions of the cMRI

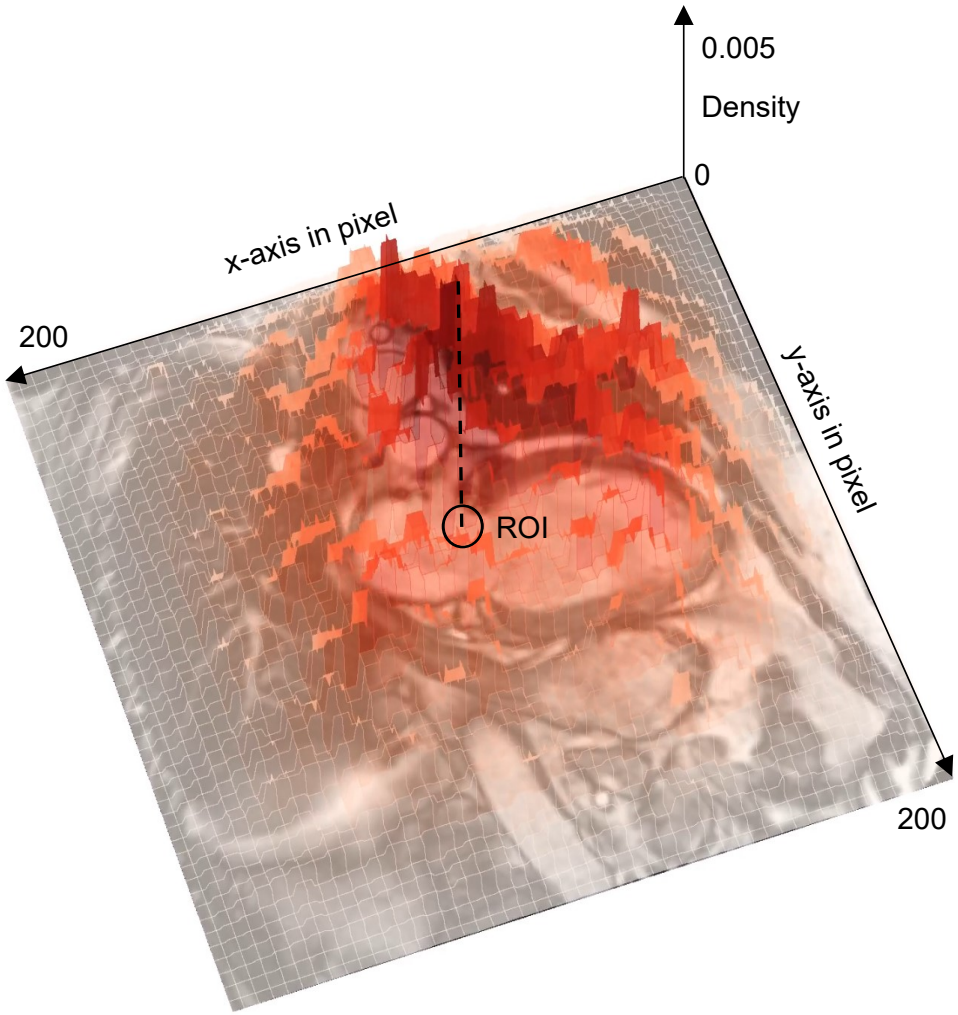


Figure 5.9: Cumulative attention map. Generated by overlaying all individual attention maps from the core cohort dataset.

sequences. In detail, a 5x5 pixel window was slid across the image stack (x-,y- and time-dimension) to occlude a small part of the image. The obscured pixels were substituted with the mean grayscale pixel value throughout the particular cMRI sequence. To assess the significance of the occluded region for the LVEDP prediction, the baseline prediction from the normal cMRI video was compared with the prediction from the occluded cMRI video (Figure 2.11). Utilizing this approach, attention maps for every frame in the cMRI video were calculated. In the end, overlays of the attention maps with the cMRI were generated to visualize the most relevant regions.

The directionality of the LVEDP prediction was also obtained and denoted by color coding. Regions marked as green contributed to a lower LVEDP prediction, whereas areas colored in a red shade were signaling elevated filling pressure. The transparency of the color quantified the strength of the impact. The lower the transparency, the higher the impact.

An overlay of all generated attention maps from the core cohort is displayed in Figure 5.9. A reference cMRI is shown below to aid localization. The analysis reveals that the neural network's attention is predominantly centered on the mitral valve area. However, the observed dispersion may appear larger than it actually is, since the MRIs producing the individual attention maps were not perfectly aligned.

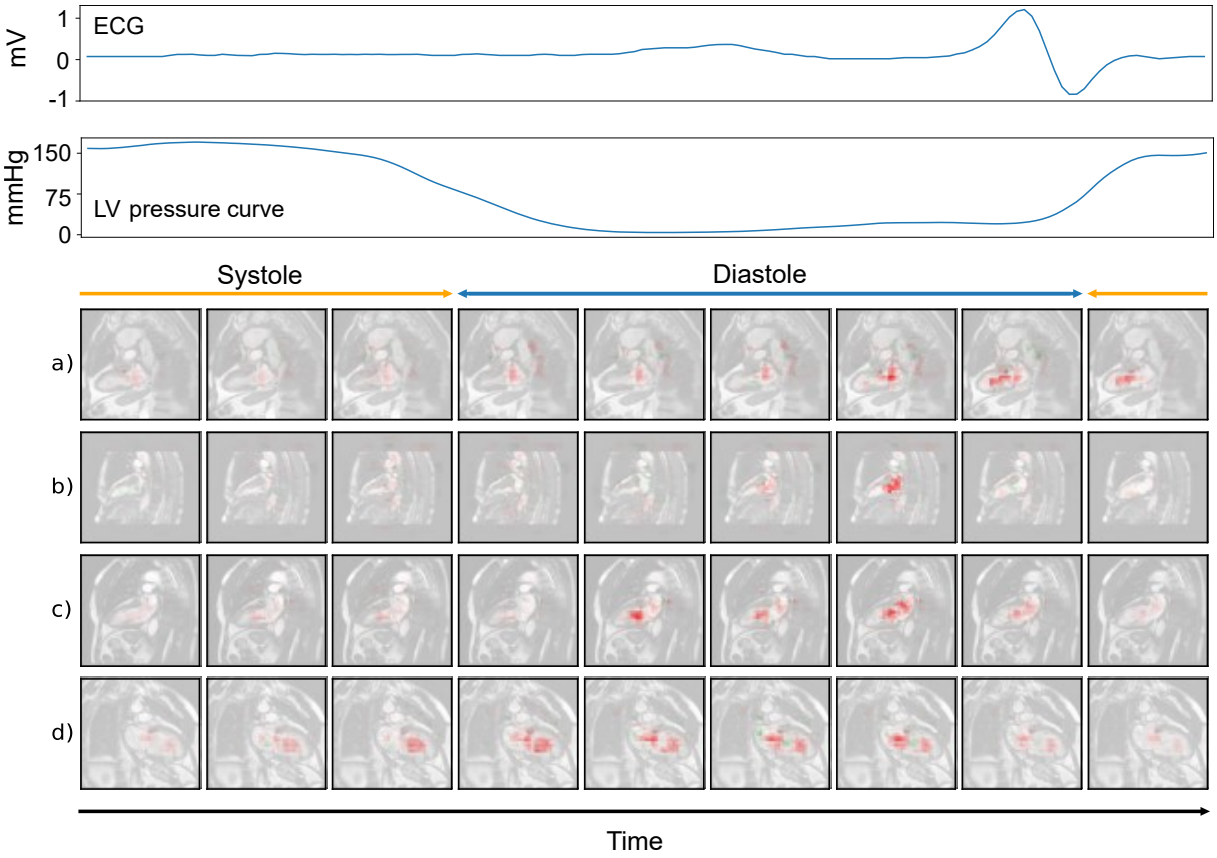


Figure 5.10: AI-LVEDP attention maps for different patients over time.

Attention maps for individual patients are shown in Figure 5.10. Each row (a-d) refers to a different patient. The image sequence displayed in each row shows the AI-LVEDP attention over time during exactly one heart cycle. For reference, an ECG trace and a left-ventricular pressure are aligned with the time axis. Especially during the diastolic phase of the cardiac cycle, the majority of highlighted areas can be observed. Prior to the opening of the mitral valve, the attention focus is on the left atrium in the vicinity of the mitral valve. Upon valve opening, the attention moves to follow the blood flow through the mitral valve into the left ventricle (Figure 5.10 (a)). After the blood flow stops, the visualized attention vanishes.

Furthermore, small levels of attention outside the cardiac cavities could be observed. This is not necessarily wrong, since peripheral water retention or other potential biomarkers found in the cMRI could aid the prediction, e.g. as an indication of heart insufficiency. To investigate the effect of this periphery attention, every pixel outside the heart was masked (replaced) by black pixels. The AI model was retrained based on the masked dataset. However, this had no significant impact on model performance, suggesting that extracardiac information was not largely relevant for the AI prediction.

Overall, findings emphasize the correct functionality of AI-LVEDP and found concepts are in agreement with established echocardiography-based estimations of diastolic function [Lancellotti et al., 2017].

6 AI-based Diagnosis of Cardiomyopathies using Cardiac MRI (AI-CMP)

Cardiomyopathies (CMPs) show an increasing prevalence worldwide [Roth et al., 2020]. The diverse nature of cardiomyopathies makes diagnosis a challenging task [Belloni et al., 2008]. In this chapter, it is shown that AI agents (named AI-CMP) can accurately diagnose cardiomyopathies from a single 4-chamber cardiac MRI. Furthermore, a thorough model introspection shows interesting insights into which heart structures were important for the AI for diagnosis discovery.

6.1 Cardiomyopathies

Cardiomyopathies are a group of pathologies that affect the heart muscle, leading to structural and functional abnormalities [Ciarambino et al., 2021]. The cardiomyopathies included in AI-CMP are hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM), ischemic cardiomyopathy (ICM) and cardiac amyloidosis, which is one of the leading causes of restrictive cardiomyopathy [Kittleson et al., 2020].

Hypertrophic cardiomyopathy (HCM) is characterized by the thickening of the heart muscle, particularly the left ventricle. Genetics are the most important factor for the manifestation of the disease [Maron and Maron, 2013]. Hence, it is often inherited. The thickening can obstruct blood flow out of the heart, making the heart less effective at pumping blood.

Dilated cardiomyopathy (DCM) is a condition expressed by the dilation or enlargement of the heart's chambers. DCM can be caused by a variety of factors, including infections, genetic traits or exposure to drugs and toxins [Weintraub et al., 2017]. The dilation weakens the heart muscle, leading to impaired cardiac function.

Ischemic cardiomyopathy (ICM) is a type of cardiomyopathy that develops as a result of coronary artery disease, which leads to inadequate blood supply to the heart muscle due to the buildup of plaque [Moroni et al., 2021]. The reduced blood flow deprives the heart muscle of oxygen and nutrients, which leads to damage to the heart muscle and prevents it from pumping blood effectively.

Cardiac amyloidosis has a significantly lower incidence than the other mentioned cardiomyopathies [Rubin and Maurer, 2020]. The cause of cardiac amyloidosis is the deposition of abnormal protein fibers called amyloids in the heart tissue. These amyloid deposits lead to stiffening and thickening of the heart muscle, impairing its function over time.

Common consequences of cardiomyopathies are shortness of breath, fatigue, arrhythmias, chest pain, heart failure or even sudden cardiac death [McKenna et al., 2017].

Cardiac MRI is gold-standard for diagnosis of cardiomyopathies [Salerno et al., 2017]. However, cardiologists usually do not rely on only a single cMRI view and often include other modalities. Especially for late gadolinium enhancement (LGE), gadolinium is a cMRI contrast agent, enables information about potential cardiac fibrosis and can therefore aid the physician heavily in concluding a diagnosis. LGE can be used to diagnose myocarditis (ICM), showing patterns of patchy midmyocardial or focal subepicardial LGE as a sign of potentially irreversible damage (necrosis, fibrosis, edema). Also, DCM (midmyocardial to subepicardial LGE), HCM (patchy to streaky midmyocardial LGE) and amyloidosis (subendocardial LGE) can be diagnosed by different LGE patterns [Becker et al., 2018, Weng et al., 2016]. The additional information from the LGE would likely improve the classification accuracy of the model developed here. However, by only relying on a single frame 4-chamber cardiac MRI, the requirements for usage are kept minimal. The injection of the contrast agent gadolinium is regarded safe, but could cause nausea, headache, abnormal skin sensation, or, in rare cases, an allergic reaction. In-vitro studies have found gadolinium-based contrast agents to be neurotoxic, but the long-term effects of gadolinium remain unclear, even after several retrospective studies [Bower et al., 2019, Olchowy et al., 2017]. As a consequence, LGE is only used conservatively in clinical practice [Blumfield et al., 2017, Organization et al., 2010].

6.2 Model development of AI-CMP

More than 11,000 MRI examinations of the heart were carried out at the University Hospital of Heidelberg, between 2006 and 2023. From those patients, structural cardiomyopathy phenotypes or confirmed no structural heart disease could be verified in 4,390 patients via semi-automated doctoral letter screening (Fig. 1.1). In recent cMRI reports the diagnosis has been specified in a fully structured way (dropdown selection). Before, the list of diagnoses was stated in the semi-structured diagnosis section from the doctoral letters. If this section was detected, the relevant diagnoses were extracted automatically. If the diagnosis section was not found by the automated script, a keyword search (HCM, Amyloidosis, …) was conducted and the letter was evaluated manually with highlighted keywords in the free text. The keyword highlighting proved to be very beneficial in terms of manual screening speed. Patients without any structural heart disease indication during the cMRI were marked as controls.

The 4-chamber cine MRIs were preprocessed via brightness adaptation and bi-cubic interpolation [Keys, 1981]. Deep convolutional neural networks were trained in a 5-fold cross-validation scenario to benchmark performance. The model was trained to distinguish between four types of cardiomyopathies and controls. An architecture search was conducted by varying the number of layers and layer configurations. The best results were achieved with scaled exponential linear unit (SELU) activation [Klambauer et al., 2017]. The optimizer was Adam. The data was divided into train/validation/test set with a split of 60/20/20 %. The validation set played a crucial role in monitoring overfitting and implementing early stopping. In each epoch, the current model is checked to determine whether it is the best model with regard to the validation data. If this is the case, the model is saved. After training, the performance is reported by evaluating the saved

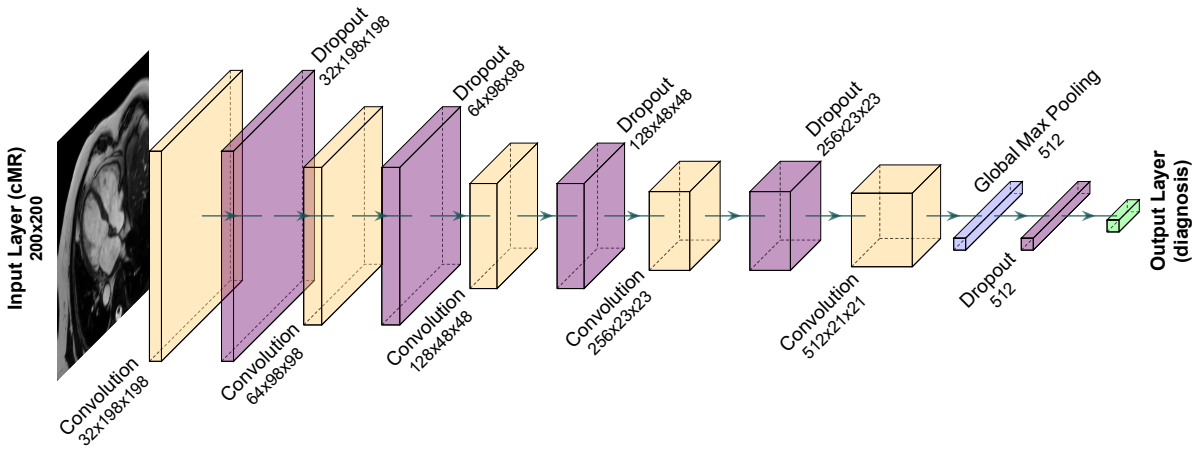


Figure 6.1: AI-CMP architecture.

model on the test set.

During training, data augmentation was utilized to improve generalization. The augmentations included rotation, chopping, brightness variation, contrast variation, sharpness variation and distortion. The data augmentation proved to be essential for regularizing the models. The preprocessing and data augmentation techniques used for AI-CMP were similar to the methods described in Section 5.2. The batch size during training was 250. The output activation was softmax activation and the loss function was categorical cross-entropy. The final AI-CMP architecture is shown in Figure 6.1, with the type and output dimension of each layer specified.

6.3 AI-CMP results and benchmarking

Model outputs are the class probabilities for cardiomyopathies and control. Only a singular frame of the 4-chamber MRI was required to solve the classification problem. Therefore, the final model input was specified accordingly. By requiring only a single frame, the examination time could be significantly reduced compared to a full cine sequence acquisition. This could benefit patient throughput and potentially impact future cMRI protocols.

ROC AUCs for specific classes were calculated in a one-vs-all scenario. Plots of the receiver-operator curves can be found in Figure 6.2. The class AUCs ranged from 0.82-0.92. In gray, the individual curves from each cross-validation are shown. The low variation in the individual ROCs

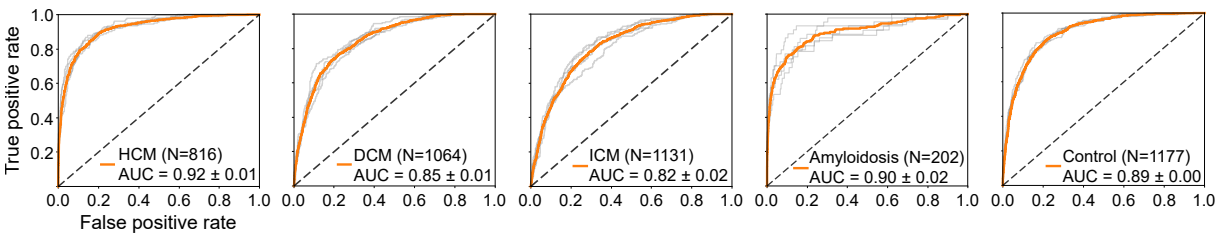


Figure 6.2: ROC curves for AI diagnosis from cardiac MRI.

highlights the large amounts of data. The class-specific accuracies ranged from 0.78 to 0.96

(Table 6.1). However, the high accuracy for amyloidosis is biased by an already high majority class baseline of 95 %. The overall accuracy for the five-class classification problem was 64 %. In comparison, the control group was with 27 % the most occurring class in the dataset.

Diagnosis	Number of patients	AUC	Accuracy
HCM	816	0.92 (CI95% 0.91-0.94)	0.89 (CI95% 0.88-0.90)
DCM	1,064	0.85 (CI95% 0.84-0.87)	0.81 (CI95% 0.78-0.84)
ICM	1,131	0.82 (CI95% 0.79-0.84)	0.78 (CI95% 0.76-0.80)
Amyloidosis	202	0.90 (CI95% 0.86-0.93)	0.96 (CI95% 0.96-0.97)
Control	1,077	0.89 (CI95% 0.89-0.90)	0.83 (CI95% 0.82-0.84)

Table 6.1: Class-specific AI-CMP performance predicting diagnosis using cMRI.

A detailed confusion matrix is shown in Figure 6.3. The AI was able to distinguish between DCM and HCM with an accuracy of 94 %. The distinction between DCM/Control (Accuracy=82 %) or ICM/DCM (Accuracy=76 %) was much more difficult for the AI model, which also reflects the actual task difficulty for experts [Mahrholdt et al., 2005].

A subgroup analysis was conducted to investigate the robustness and potential bias of AI-CMP. The average patient was 58 years old. In the group of patients older than 60 years, the overall accuracy decreased to 0.6 and class-specific ROC AUCs ranged from 0.76-0.90, with the ICM classification being the most difficult. In the group of patients younger than 60 years, the overall accuracy increased to 0.68 (AUCs 0.81-0.94). The number of male patients was 72 %. In the subgroup of male patients, the accuracy was 0.63, whereas in the female patients group, the overall accuracy was 0.66. The ejection fraction (EF) was on average 52 %. No major differences in model performance could be found between the low EF (≤50) and the high EF (>50) group (Accuracy 0.63 vs. 0.65).

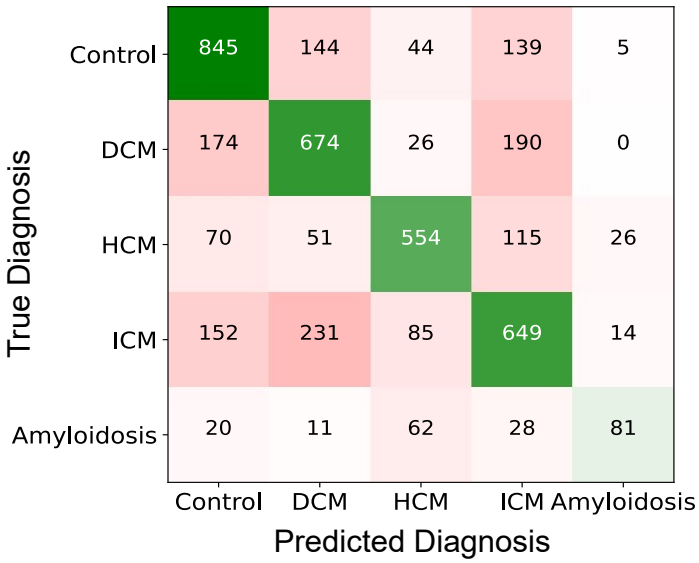


Figure 6.3: Confusion matrix AI-CMP predictions utilizing cardiac MRI.

Across the spectrum of analyzed patient characteristics no substantial model bias was revealed.The developed application could aid physicians making a differential diagnosis and function as a gatekeeper for LGE. Also, in smaller healthcare providers less common diseases

such as amyloidosis can be easily missed. Furthermore, there are potential areas of application in teaching.

6.4 Introspection of AI-CMP by attention mapping

Attention mapping by occlusion was employed to highlight the MRI areas that had a strong impact on the AI model's predictions. Further technical details on the attention mapping setup are provided in Section 5.7.

Overall, the highlighted areas are in concordance with areas described as relevant by MRI expert readers (Figure 6.4). For the diagnosis of HCM, the neural network recognizes the thickened interventricular septum as the most pertinent information. Furthermore, it is intriguing that the AI recognizes the LV apex aneurysm, which is a common symptom of ischemic cardiomyopathy. In

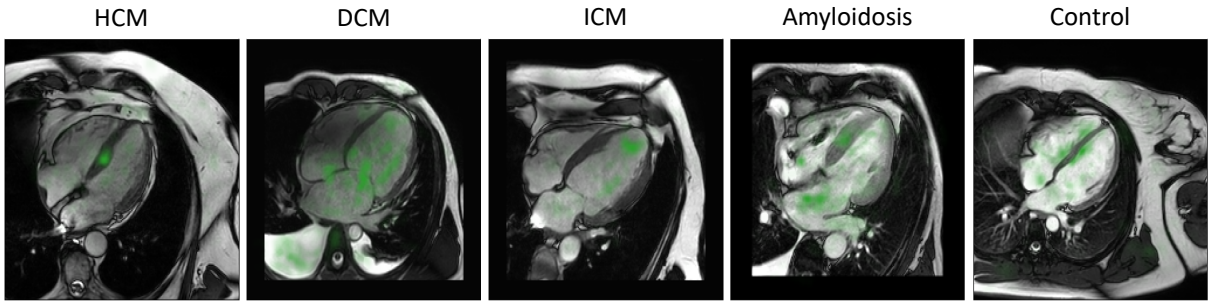


Figure 6.4: Attention maps for diagnosis predictions from cardiac MRI.

the DCM cMRI, the AI's attention is concentrated on the areas outside the heart's periphery. The marked pleural effusion (water retention) is characteristic of heart insufficiency, which is often caused by cardiomyopathies [Maisch et al., 2012]. Hence, the presence of pleural effusions is a valid biomarker, which could provide important information regarding this classification task. Secondly, the focus was also on the dilated left ventricle and the left atrium. A variety of disease patterns are frequently observed in cardiac amyloidosis, including myocardial thickening, atrial dilatation, dark blood pools and pericardial effusion [Maceira et al., 2005]. In concordance, the network attention is extensively distributed over the whole heart region. In controls, the mapping usually highlighted the cardiac septum area.

7 Discussion and Conclusion

Diagnosis of diastolic heart failure remains one of the most challenging problems in cardiology. The left ventricular end-diastolic pressure (LVEDP) is recognized as the key indicator for the assessment of diastolic heart function. However, this biomarker is difficult to obtain in clinical routine, as it involves invasive intracardiac pressure measurements, which are naturally associated with interventional risks for the patient. Therefore, LVEDP is often unavailable in suspected heart failure patients. The AI-LVEDP model, as developed in the context of this dissertation, constitutes a non-invasive option to infer the LVEDP from commonly available cardiac MRI recordings [Lehmann et al., 2024].

For this, a collective of 66,936 patients who underwent cardiac catheterization including hemodynamic measurements was evaluated and filtered for patients additionally receiving a cardiac MRI examinations within a 30 day window. The automated annotation of left ventricular pressure curves resulting from the hemodynamics, with over 183,000 individual samples, by a self developed algorithm was essential in providing unbiased training objectives for AI-LVEDP. Furthermore, the extended analysis yielded quantified measures of LVEDP-biovariability and the effects of contrast agents. Artificial neural networks were trained to reveal the hidden information content inherent in LV pressure curves, such as the patient's age and sex. A subsequent investigation concluded that the AI model was also inferring the patient's cardiac health status during the age prediction. Furthermore, the model's capability to diagnose coronary artery disease has clinical applicability in aiding as a possible gatekeeper for coronary angiography [Patel et al., 2010]. Following this analysis, the convolutional AI-LVEDP network was trained using cardiac MRI as the input data, with the automated LVEDP labels serving as the objective function.

AI-LVEDP could deliver more precise LVEDP estimates than those achieved by human cMRI experts and routinely used echocardiographic parameters for diastolic dysfunction. Although echocardiography, including AI-enhanced echocardiography, is widely used for predicting filling pressures, incorporating AI-supported cMRI can enhance the accuracy and efficiency of diastolic function assessment [Lancellotti et al., 2017, Tromp et al., 2022]. This integration could also provide valuable new annotations for large-scale biobanks [Gomes et al., 2024].

AI estimates for filling pressure showed a significant correlation with natriuretic peptides, indicating heart insufficiency, further highlighting its capability to dichotomize patients into clinically relevant subgroups [Members et al., 2008, Oremus et al., 2014]. AI-LVEDP only requires a common 2-chamber MRI sequence as input and does not rely on contrast agent requiring or time-intensive protocols, thus making possible integration of AI-LVEDP into the clinical workflow feasible. A large number of patients could benefit from earlier diagnosis of diastolic heart failure through this commonly used modality.

An extensive introspection strategy using attention mapping was employed to address the model's explainability, demonstrating that blood flow and, specifically, the mitral valve area had a pivotal role in driving the LVEDP prediction. The performance of AI-LVEDP was consistent

regardless of age, sex, atrial fibrillation status and left-ventricular ejection fraction. The analysis on ejection fraction subgroups indicated that signs of heart insufficiency are identified by the AI as an important contributing factor, but demonstrates that the recognition of elevated LVEDP is not primarily driven by the diagnosis of systolic heart failure itself [Lehmann et al., 2024].

While AI models have performed well in initial research cohorts, they frequently encounter issues when deployed in other centers or on varied datasets [Vardas et al., 2022]. Therefore, adopting a comprehensive validation strategy is crucial, as the absence of generalization severely limits the clinical application [Cabitza et al., 2021].

The complete dataset gathered for the development of AI-LVEDP included cMRI sequences from various different MRI scanners, vendors, field strengths and imaging protocols and was collected over a time span of 18 years. More than 100 investigators captured the left ventricular pressure curves throughout the same timeframe. The anticipated inter-observer variability was addressed through the automated analysis of pressure curves. Even though the dataset exhibited considerable heterogeneity, attributable to the external validation across two separate clinical centers and an extra validation cohort from Heidelberg, the AI demonstrated the ability to address this diversity effectively [Lehmann et al., 2024].

Additionally to the automated inference of diastolic function, the AI-CMP model was developed to diagnose various types of cardiomyopathies including hypertrophic, dilated, and ischemic cardiomyopathy (ICM) and cardiac amyloidosis, which constitute another major burden for health care systems [Brieler et al., 2017]. AI-CMP was extensively validated via attention mapping and sensitivity analysis. The methodology of AI-CMP, as revealed by introspection, used for diagnosis was comparable to guideline approaches by human experts. Further, a subgroup analysis yielded no significant bias across different patient groups. AI-CMP only requires a single cMRI frame for diagnosis, which could enable shorter cMRI protocols in the future. Also, experts rely on contrast agent-enhanced sequences, whereas the AI only relied on standard procedures, keeping the prerequisites minimal for application. Especially in centers with lower patient throughput, AI-assisted diagnosis of less frequent conditions, such as cardiac amyloidosis, could benefit diagnosis precision.

Finally, to summarize the key implication of this dissertation, AI-enhanced cardiac magnetic resonance imaging is capable of inferring left ventricular filling pressures and diagnosing heart diseases. If cardiac MRI is conducted, the additional availability of AI-CMP and AI-LVEDP can add to clinical decision-making without further effort. These automated and highly scalable systems could facilitate diagnostics in an ever-increasing number of patients.

However, this study is not without its limitations. An essential prerequisite in the development of AI-LVEDP was that the LVEDP was comparable at the time of the cardiac catheterization and the cMRI. This was ensured by excluding emergency cases and patients with procedures likely affecting the filling pressure during that time window.

Even direct invasive measurements of LVEDP in stable patients still reveal considerable variability within the same individual, with LVEDP showing notable changes with each cardiac cycle. The average variability in LVEDP, measured as the difference between minimum and maximum LVEDP extracted per beat over the course of the cardiac catheterization, was 7.91 mmHg. This limitation was partly addressed by aggregating the LVEDP from individual heartbeats and deriving robust LVEDP annotations utilizing outlier detection and statistical tests. This possibly

led to a reduction of noise in the annotations, but can not eliminate the inherent uncertainty in its entirety.

The use of diuretics, which was not considered an exclusion criterion, is known to influence the filling pressure by overall blood volume reduction. Data on patient medications showed that 42% were on diuretics or aldosterone antagonists, which could impact LVEDP and possibly impair the AI's prediction accuracy for filling pressure. While this effect was not observed for the entire cohort, it may have led to misclassification in certain individual cases [Lehmann et al., 2024].

The data used for the AI models came from retrospective sources. Retrospective data rarely matches the quality of clinical trial data [DeMaria, 2008]. It is crucial to emphasize the necessity for additional real-world applications of the AI models. A registered clinical trial, potentially in a multicenter setting, could be the final test to validate the robustness of the AI models. Incorporating a diverse range of minorities and rare-disease patients is also essential for a well-rounded assessment, as extending the monitoring of model performances across additional subgroups reveals the full scope of application and biases [Zohuri and Moghaddam, 2020].

Incorporating federated learning could be an effective approach, with the algorithm traveling securely to different clinical sites to train or validate the AI model while minimizing the risk of data leaks [Rieke et al., 2020]. However, there is also the possibility of reconstructing training data from neural network weights. This form of data dissemination is still subject to ongoing research [Haim et al., 2022]. Preventive strategies likely involve using large training datasets and applying strong model regularization, as these issues are related to signs of overfitting.

Heart segmentation was considered as a possible preprocessing step of AI-LVEDP and AI-CMP to remove less relevant and potentially distracting information from the heart's periphery. Investigative analysis revealed that the additional segmentation did not lead to an immediate improvement in performance. Although modern segmentation models yield robust performance, it is still sometimes the case that the segmentation result is unstable for no obvious reason [Sander et al., 2020, Alabed et al., 2022]. This consideration, combined with the idea that peripheral information, such as water retention associated with heart insufficiency, could yield useful additional information for the AI (as observed in attention mapping), led to the decision to exclude segmentation in the final models. The main concern was the segmentation, which introduced another layer of possible failure, especially with regard to the heterogeneous validation data. However, the potential of upstream segmentation for AI-LVEDP and AI-CMP is still evident and could be further investigated in subsequent studies. In particular, the use of segmentation for precise centering of the heart in the sequences presented to the AI models, adding another important level of data consistency, could yield a positive impact on future performance.

Multimodal approaches, incorporating additional information such as blood biomarkers, contrast agent enhanced cMRI sequences and echocardiography in the AI models, would also likely lead to improved precision. However, this comes with multiple implications. A strong limitation would be the lack of complete data, since it is unlikely that a large number of patients has the complete information across the multiple modalities available for training and for prospective prediction. Applying data imputation is not a promising approach, as these methods rely on linear interpolation, which is insufficient in the context of the complex non-linear patterns recognized by deep neural networks [Azur et al., 2011]. A better option is to use sophisticated AI models that are inherently capable of handling incomplete data. These models should be designed

to train with incomplete information, in order to not limit applicability, while still being able to utilize additional information whenever available. Such large-scale models would certainly be superior to existing solutions and could yield groundbreaking advances similar to Chat-GPT in the domain of large language models [Zhao et al., 2023]. However, the vast amount of data needed for training such ambitious models would require more data accumulation initiatives and digital infrastructure, following the example of the UK Biobank [Bycroft et al., 2018].

AI-LVEDP and AI-CMP could aid in the early diagnosis of cardiomyopathies and diastolic heart failure. To receive the necessary cMRI, patients must already be in contact with the hospital, which is often the case only if the patient is already exhibiting symptoms. There is a great need for screening tools that can assess the patient's cardiac status before severe symptoms manifest. Yet, there are currently no reliable screening options available. Population-wide screening for early-stage heart failure is not feasible using cMRI. Smartwatch-based ECGs are among the most promising approaches [Isakadze and Martin, 2020]. Smartwatch ECGs have been previously used for the diagnosis of arrhythmias, which are normally identified using normal ECG [Bumgarner et al., 2018, Perez et al., 2019]. Those results are promising. However, it remains uncertain whether valuable information about cardiomyopathies and diastolic heart failure can be extracted from ECG. The relevance of even a moderately effective screening tool would be substantial. Consequently, this area should be explored extensively in future research.

To further address the point of early diagnosis, it is not clear if the diagnosis and early treatment of cardiomyopathies and diastolic heart failure is the optimal point of intervention when it comes to maximizing the impact on public health. Although the genetic aspect of heart failure is well established, lifestyle also has a large impact in preventing disease or mitigating symptoms [Morita et al., 2005, Hershberger et al., 2018]. Drug abuse, excessive sodium intake, obesity and smoking are key factors promoting heart failure [Ponikowski et al., 2014]. In analogy, a healthy diet, physical exercise and leisure activity are known to significantly reduce the incidence of negative cardiovascular events and mortality [Woolf, 2008, Nayor and Vasan, 2015, Aggarwal et al., 2018]. Hence, promoting prevention via adopting lifestyle changes could even have a higher impact on public health than novel diagnostics or treatment options.

Estimating the impact of prevention on healthcare systems is complex. The anticipated effect of preventive measures is that people live longer. Nevertheless, they will still require medical care at some point, albeit with a delay of a few years. Despite the benefits of prevention in extending healthy years and improving quality of life, the onset of age-related symptoms is ultimately unavoidable. Significant relief for the healthcare system can probably only be achieved through the automation of processes and the associated increase in patient throughput. This inevitable evolution of future healthcare will be heavily directed by AI.

List of Abbreviations

ACS	Acute coronary syndrome
Adam	Adaptive moment estimation
Afib	Atrial fibrillation
AI	Artificial intelligence
atm	Atmosphere
AUC	Area under curve
bpm	Beats per minute
CC	Cardiac catheterization
CI	Confidence interval
CMP	Cardiomyopathies
CNN	Convolutional neural network
DCM	Dilated cardiomyopathy
DFT	Discrete Fourier transform
EDD	End-diastolic dimension
EF	Ejection fraction
ESD	End-systolic dimension
FFT	Fast Fourier transform
g	Gram
GAN	Generative adversarial networks
Grad-CAM	Gradient-weighted class activation mapping
HCM	Hypertrophic cardiomyopathy
HF	Heart failure
HF-pEF	Heart failure with preserved ejection fraction
HF-rEF	Heart failure with reduced ejection fraction
hPa	Hectopascal
hsTnT	High-sensitivity troponin T
Hz	Hertz
i.e.	Id est
ICM	Ischemic cardiomyopathy
kg	Kilogram
l	Litre
LA	Left atrium
LGE	Late Gadolinium enhancement
LSTM	Long short-term memory
LV	Left ventricle
LV-DP	Left-ventricular minimum diastolic pressure

LV-SP	Left-ventricular peak systolic pressure
LVEDP	Left-ventricular end-diastolic pressure
MAD	Median absolute deviation
MAE	Mean absolute error
mmHg	Millimetre of mercury
MRI	Magnetic resonance imaging
ms	Milliseconds
MSE	Mean squared error
ng	Nanogram
NN	Neural network
NT-proBNP	N-terminal pro–B-type natriuretic peptide
NYHA	New York heart association classification
PACS	Picture archiving and communication system
ReLU	Rectified linear unit
ResNet	Residual neural network
RIFE	Real-time intermediate flow estimation
RMSprop	Root mean square propagation
ROC	Receiver operating characteristic
R^2	Coefficient of determination
SELU	Scaled exponential linear unit
SGLT2	Sodium glucose transporter 2
std	Standard deviation
VAE	Variational autoencoder

Bibliography

[Aalaei-Andabili and Bavry, 2019] Aalaei-Andabili, S. H. and Bavry, A. A. (2019). Left ventricular diastolic dysfunction and transcatheter aortic valve replacement outcomes: a review. *Cardiology and therapy*, 8(1):21–28.

[Abadi et al., 2016] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

[Abdolrasol et al., 2021] Abdolrasol, M. G., Hussain, S. S., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., Ali, J. A., Mekhilef, S., and Milad, A. (2021). Artificial neural networks based optimization techniques: A review. *Electronics*, 10(21):2689.

[Agarwal et al., 2017] Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., and Ma, T. (2017). Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199.

[Aggarwal et al., 2018] Aggarwal, M., Bozkurt, B., Panjrath, G., Aggarwal, B., Ostfeld, R. J., Barnard, N. D., Gaggin, H., Freeman, A. M., Allen, K., Madan, S., et al. (2018). Lifestyle modifications for preventing and treating heart failure. *Journal of the American College of Cardiology*, 72(19):2391–2405.

[Akbiyik, 2023] Akbiyik, M. E. (2023). Data augmentation in training cnns: injecting noise to images. *arXiv preprint arXiv:2307.06855*.

[Akoglu, 2018] Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3):91–93.

[Al-Hijji et al., 2019] Al-Hijji, M. A., Lennon, R. J., Gulati, R., El Sabbagh, A., Park, J. Y., Crusan, D., Kanwar, A., Behfar, A., Lerman, A., Holmes, D. R., et al. (2019). Safety and risk of major complications with diagnostic cardiac catheterization. *Circulation: Cardiovascular Interventions*, 12(7):e007791.

[Alabed et al., 2022] Alabed, S., Maiter, A., Salehi, M., Mahmood, A., Daniel, S., Jenkins, S., Goodlad, M., Sharkey, M., Mamalakis, M., Rakocevic, V., et al. (2022). Quality of reporting in ai cardiac mri segmentation studies–a systematic review and recommendations for future studies. *Frontiers in Cardiovascular Medicine*, 9:956811.

[Anker et al., 2021] Anker, S. D., Butler, J., Filippatos, G., Ferreira, J. P., Bocchi, E., Böhm, M., Brunner-La Rocca, H.-P., Choi, D.-J., Chopra, V., Chuquiure-Valenzuela, E., et al. (2021). Empagliflozin in heart failure with a preserved ejection fraction. *New England Journal of Medicine*, 385(16):1451–1461.

[Antoniou et al., 2017] Antoniou, A., Storkey, A., and Edwards, H. (2017). Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*.

[Arbelo et al., 2023] Arbelo, E., Protonotarios, A., Gimeno, J. R., Arbustini, E., Barriales-Villa, R., Basso, C., Bezzina, C. R., Biagini, E., Blom, N. A., de Boer, R. A., et al. (2023). 2023 esc guidelines for the management of cardiomyopathies: Developed by the task force on the management of cardiomyopathies of the european society of cardiology (esc). *European heart journal*, 44(37):3503–3626.

[Avendi et al., 2016] Avendi, M. R., Kheradvar, A., and Jafarkhani, H. (2016). A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri. *Medical image analysis*, 30:108–119.

[Azur et al., 2011] Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49.

[Baim and Grossman, 2006] Baim, D. and Grossman, W. (2006). *Grossman's Cardiac Catheterization, Angiography, and Intervention*. LWW medical book collection. Lippincott Williams & Wilkins.

[Becker et al., 2018] Becker, M. A., Cornel, J. H., Van de Ven, P. M., van Rossum, A. C., Allaart, C. P., and Germans, T. (2018). The prognostic value of late gadolinium-enhanced cardiac magnetic resonance imaging in nonischemic dilated cardiomyopathy: a review and meta-analysis. *JACC: Cardiovascular Imaging*, 11(9):1274–1284.

[Bellman, 1954] Bellman, R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515.

[Belloni et al., 2008] Belloni, E., De Cobelli, F., Esposito, A., Mellone, R., Perseghin, G., Canu, T., Del Maschio, A., et al. (2008). Mri of cardiomyopathy. *AJR Am J Roentgenol*, 191(6):1702–1710.

[Bianco et al., 2018] Bianco, S., Cadene, R., Celona, L., and Napoletano, P. (2018). Benchmark analysis of representative deep neural network architectures. *IEEE access*, 6:64270–64277.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[Blumfield et al., 2017] Blumfield, E., Moore, M. M., Drake, M. K., Goodman, T. R., Lewis, K. N., Meyer, L. T., Ngo, T. D., Sammet, C., Stanescu, A. L., Swenson, D. W., et al. (2017). Survey of gadolinium-based contrast agent utilization among the members of the society for pediatric radiology: a quality and safety committee report. *Pediatric radiology*, 47:665–673.

[Bochner and Chandrasekharan, 1949] Bochner, S. and Chandrasekharan, K. (1949). *Fourier transforms*. Number 19. Princeton University Press.

[Boureau et al., 2010] Boureau, Y.-L., Ponce, J., and LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118.

[Bower et al., 2019] Bower, D. V., Richter, J. K., von Tengg-Kobligk, H., Heverhagen, J. T., and Runge, V. M. (2019). Gadolinium-based mri contrast agents induce mitochondrial toxicity and cell death in human neurons, and toxicity increases with reduced kinetic stability of the agent. *Investigative radiology*, 54(8):453–463.

[Brieler et al., 2017] Brieler, J., Breeden, M. A., and Tucker, J. (2017). Cardiomyopathy: an overview. *American family physician*, 96(10):640–646.

[Brienesse et al., 2018] Brienesse, S. C., Davies, A. J., Khan, A., and Boyle, A. J. (2018). Prognostic value of lvedp in acute myocardial infarction: a systematic review and meta-analysis. *Journal of cardiovascular translational research*, 11:33–35.

[Brigham, 1988] Brigham, E. O. (1988). *The fast Fourier transform and its applications*. Prentice-Hall, Inc.

[Bumgarner et al., 2018] Bumgarner, J. M., Lambert, C. T., Hussein, A. A., Cantillon, D. J., Baranowski, B., Wolski, K., Lindsay, B. D., Wazni, O. M., and Tarakji, K. G. (2018). Smartwatch algorithm for automated detection of atrial fibrillation. *Journal of the American College of Cardiology*, 71(21):2381–2388.

[Bycroft et al., 2018] Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209.

[Cabitza et al., 2021] Cabitza, F., Campagner, A., Soares, F., de Guadiana-Romualdo, L. G., Challa, F., Sulejmani, A., Seghezzi, M., and Carobene, A. (2021). The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Computer Methods and Programs in Biomedicine*, 208:106288.

[Campesato, 2020] Campesato, O. (2020). *Artificial intelligence, machine learning, and deep learning*. Mercury Learning and Information.

[Castelvecchi, 2016] Castelvecchi, D. (2016). Can we open the black box of ai? *Nature News*, 538(7623):20.

[Chemla et al., 2009] Chemla, D., Nitenberg, A., Teboul, J.-L., Richard, C., Monnet, X., Le Clesiau, H., Valensi, P., and Brahimi, M. (2009). Subendocardial viability index is related to the diastolic/systolic time ratio and left ventricular filling pressure, not to aortic pressure: an invasive study in resting humans. *Clinical and Experimental Pharmacology and Physiology*, 36(4):413–418.

[Chen, 2017] Chen, Y.-C. (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1):161–187.

[Chicco et al., 2021] Chicco, D., Warrens, M. J., and Jurman, G. (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science*, 7:e623.

[Choi et al., 2020] Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., and Campbell, J. P. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational vision science & technology*, 9(2):14–14.

[Chollet, 2018] Chollet, F. (2018). *Deep learning mit python und keras: das praxis-handbuch vom entwickler der keras-bibliothek*. MITP-Verlags GmbH & Co. KG.

[Chung et al., 2015] Chung, C. S., Shmuylovich, L., and Kovács, S. J. (2015). What global diastolic function is, what it is not, and how to measure it. *American Journal of Physiology-Heart and Circulatory Physiology*, 309(9):H1392–H1406.

[Ciarambino et al., 2021] Ciarambino, T., Menna, G., Sansone, G., and Giordano, M. (2021). Cardiomyopathies: an overview. *International journal of molecular sciences*, 22(14):7722.

[Cohen et al., 1996] Cohen, G. I., Pietrolungo, J. F., Thomas, J. D., and Klein, A. L. (1996). A practical guide to assessment of ventricular diastolic function using doppler echocardiography. *Journal of the American College of Cardiology*, 27(7):1753–1760.

[Collins and Moons, 2019] Collins, G. S. and Moons, K. G. (2019). Reporting of artificial intelligence prediction models. *The Lancet*, 393(10181):1577–1579.

[Cooley and Tukey, 1965] Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301.

[Cormen et al., 2022] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2022). *Introduction to algorithms*. MIT press.

[Corsini et al., 2022] Corsini, A., Cercenelli, L., Zecchi, M., Marcelli, E., and Corazza, I. (2022). Chapter 30 - basic hemodynamic parameters. In Karimov, J. H., Fukamachi, K., and Gillinov, M., editors, *Advances in Cardiovascular Technology*, pages 463–474. Academic Press.

[Cybenko, 1989] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

[Dagdelen et al., 2001] Dagdelen, S., Eren, N., Karabulut, H., Akdemir, I., Ergelen, M., Saglam, M., Yüce, M., Alhan, C., and Caglar, N. (2001). Estimation of left ventricular end-diastolic pressure by color m-mode doppler echocardiography and tissue doppler imaging. *Journal of the American Society of Echocardiography*, 14(10):951–958.

[Dai et al., 2021] Dai, Z., Liu, H., Le, Q. V., and Tan, M. (2021). Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34:3965–3977.

[Dal Canto et al., 2022] Dal Canto, E., Remmelzwaal, S., van Ballegooijen, A. J., Handoko, M. L., Heymans, S., van Empel, V., Paulus, W. J., Nijpels, G., Elders, P., and Beulens, J. W. (2022). Diagnostic value of echocardiographic markers for diastolic dysfunction and heart failure with preserved ejection fraction. *Heart failure reviews*, 27(1):207–218.

[Data et al., 2016] Data, M. C., Salciccioli, J. D., Crutain, Y., Komorowski, M., and Marshall, D. C. (2016). Sensitivity analysis and model validation. *Secondary analysis of electronic health records*, pages 263–271.

[Deisenroth et al., 2020] Deisenroth, M. P., Faisal, A. A., and Ong, C. S. (2020). *Mathematics for Machine Learning*, page 149. Cambridge University Press.

[DeMaria, 2008] DeMaria, A. N. (2008). Clinical trials and clinical judgment.

[Denslow and Buckles, 1993] Denslow, S. and Buckles, D. S. (1993). Pulse oximetry-gated acquisition of cardiac mr images in patients with congenital cardiac abnormalities. *AJR. American journal of roentgenology*, 160(4):831–833.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Dokainish et al., 2008] Dokainish, H., Sengupta, R., Pillai, M., Bobek, J., and Lakkis, N. (2008). Assessment of left ventricular systolic function using echocardiography in patients with preserved ejection fraction and elevated diastolic pressures. *The American journal of cardiology*, 101(12):1766–1771.

[Dolgin et al., 1994] Dolgin, M., Association, N. Y. H., Committee, C., et al. (1994). Nomenclature and criteria for diagnosis of diseases of the heart and great vessels. *(9th ed. Boston, Lippincott Williams and Wilkins)*.

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

[Dreiseitl and Ohno-Machado, 2002] Dreiseitl, S. and Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 35(5-6):352–359.

[Du et al., 2015] Du, L.-J., Dong, P.-S., Jia, J.-J., Fan, X.-M., Yang, X.-M., Wang, S.-X., Yang, X.-S., Li, Z.-J., and Wang, H.-L. (2015). Association between left ventricular end-diastolic pressure and coronary artery disease as well as its extent and severity. *International journal of clinical and experimental medicine*, 8(10):18673.

[Du et al., 2019] Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR.

[Duchon, 1979] Duchon, C. E. (1979). Lanczos filtering in one and two dimensions. *Journal of Applied Meteorology and Climatology*, 18(8):1016–1022.

[El Naqa and Murphy, 2015] El Naqa, I. and Murphy, M. J. (2015). *What is machine learning?* Springer.

[Esteva et al., 2017] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118.

[Falk et al., 2012] Falk, M., Marohn, F., and Tewes, B. (2012). *Foundations of statistical analyses and applications with SAS*. Birkhäuser.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874. ROC Analysis in Pattern Recognition.

[Felker et al., 2020] Felker, G. M., Ellison, D. H., Mullens, W., Cox, Z. L., and Testani, J. M. (2020). Diuretic therapy for patients with heart failure: Jacc state-of-the-art review. *Journal of the American College of Cardiology*, 75(10):1178–1195.

[Forsyth and Ponce, 2002] Forsyth, D. A. and Ponce, J. (2002). *Computer vision: a modern approach*. prentice hall professional technical reference.

[Fukushima, 1980] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

[Fürnkranz et al., 2012] Fürnkranz, J., Gamberger, D., and Lavrač, N. (2012). *Foundations of rule learning*. Springer Science & Business Media.

[Gaggin and Januzzi Jr, 2013] Gaggin, H. K. and Januzzi Jr, J. L. (2013). Biomarkers and diagnostics in heart failure. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1832(12):2442–2450.

[García-Martín et al., 2019] García-Martín, E., Rodrigues, C. F., Riley, G., and Grahn, H. (2019). Estimation of energy consumption in machine learning. *Journal of Parallel and Distributed Computing*, 134:75–88.

[Garthwaite and McMahon, 2004] Garthwaite, S. M. and McMahon, E. G. (2004). The evolution of aldosterone antagonists. *Molecular and cellular endocrinology*, 217(1-2):27–31.

[Gasquet and Witomski, 2013] Gasquet, C. and Witomski, P. (2013). *Fourier analysis and applications: filtering, numerical computation, wavelets*, volume 30. Springer Science & Business Media.

[Ginat et al., 2011] Ginat, D. T., Fong, M. W., Tuttle, D. J., Hobbs, S. K., and Vyas, R. C. (2011). Cardiac imaging: Part 1, mr pulse sequences, imaging planes, and basic anatomy. *American Journal of Roentgenology*, 197(4):808–815.

[Glorot and Bengio, 2010] Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

[Gomes et al., 2024] Gomes, B., Singh, A., O'Sullivan, J. W., Schnurr, T. M., Goddard, P. C., Loong, S., Amar, D., Hughes, J. W., Kostur, M., Haddad, F., et al. (2024). Genetic architecture of cardiac dynamic flow volumes. *Nature Genetics*, 56(2):245–257.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[Goodfellow et al., 2020] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

[Goodfellow and Vinyals, 2014] Goodfellow, I. J. and Vinyals, O. (2014). Qualitatively characterizing neural network optimization problems. *CoRR*, abs/1412.6544.

[Grandini et al., 2020] Grandini, M., Bagli, E., and Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

[Graves and Schmidhuber, 2005] Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

[Greten and Andrassy, 2010] Greten, H. and Andrassy, K. (2010). Innere medizin, 13., vollst. überarb. u. erw. aufl.

[Gripenberg, 2003] Gripenberg, G. (2003). Approximation by neural networks with a bounded number of nodes at each level. *Journal of Approximation Theory*, 122(2):260–266.

[Haim et al., 2022] Haim, N., Vardi, G., Yehudai, G., Shamir, O., and Irani, M. (2022). Reconstructing training data from trained neural networks. *Advances in Neural Information Processing Systems*, 35:22911–22924.

[Hall, 2005] Hall, C. (2005). Nt-probnp: the mechanism behind the marker. *Journal of cardiac failure*, 11(5):S81–S83.

[Hanin, 2018] Hanin, B. (2018). Which neural net architectures give rise to exploding and vanishing gradients? *Advances in neural information processing systems*, 31.

[Hasebroock and Serkova, 2009] Hasebroock, K. M. and Serkova, N. J. (2009). Toxicity of mri and ct contrast agents. *Expert opinion on drug metabolism & toxicology*, 5(4):403–416.

[Haykin, 1999] Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation 2Nd Ed.* Prentice-Hall Of India Pvt. Limited.

[He et al., 2019] He, F., Liu, T., and Tao, D. (2019). Control batch size and learning rate to generalize well: Theoretical and empirical evidence. *Advances in neural information processing systems*, 32.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

[Hershberger et al., 2018] Hershberger, R. E., Givertz, M. M., Ho, C. Y., Judge, D. P., Kantor, P. F., McBride, K. L., Morales, A., Taylor, M. R., Vatta, M., and Ware, S. M. (2018). Genetic evaluation of cardiomyopathy—a heart failure society of america practice guideline. *Journal of cardiac failure*, 24(5):281–302.

[Hinton et al., 2012a] Hinton, G., Srivastava, N., and Swersky, K. (2012a). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2.

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

[Hinton et al., 2012b] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2012b). Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.

[Hochreiter, 1991] Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen netzen.

[Hochreiter et al., 2001] Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., et al. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Huang et al., 2020] Huang, Z., Zhang, T., Heng, W., Shi, B., and Zhou, S. (2020). Real-time intermediate flow estimation for video frame interpolation.

[Hubel and Wiesel, 1962] Hubel, D. and Wiesel, T. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167.

[Isakadze and Martin, 2020] Isakadze, N. and Martin, S. S. (2020). How useful is the smartwatch ecg? *Trends in cardiovascular medicine*, 30(7):442–448.

[Isensee et al., 2021] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211.

[Jaganathan et al., 2019] Jaganathan, K., Panagiotopoulou, S. K., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., Schwartz, G. B., et al. (2019). Predicting splicing from primary sequence with deep learning. *Cell*, 176(3):535–548.

[Jain et al., 2020] Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumdar, S., Afzal, S., Sharma Mittal, R., and Munigala, V. (2020). Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3561–3562.

[Januzzi Jr and Myhre, 2020] Januzzi Jr, J. L. and Myhre, P. L. (2020). The challenges of nt-probnp testing in hfpef: shooting arrows in the wind.

[Jayalakshmi and Santhakumaran, 2011] Jayalakshmi, T. and Santhakumaran, A. (2011). Statistical normalization and back propagation for classification. *International Journal of Computer Theory and Engineering*, 3(1):1793–8201.

[Johnson et al., 2021] Johnson, K. B., Wei, W.-Q., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., Zhao, J., and Snowdon, J. L. (2021). Precision medicine, ai, and the future of personalized health care. *Clinical and translational science*, 14(1):86–93.

[Jones et al., 2021] Jones, R., Varian, F., Alabed, S., Morris, P., Rothman, A., Swift, A. J., Lewis, N., Kyriacou, A., Wild, J. M., Al-Mohammad, A., et al. (2021). Meta-analysis of echocardiographic quantification of left ventricular filling pressure. *ESC heart failure*, 8(1):566–576.

[Josephson, 2008] Josephson, M. E. (2008). *Clinical cardiac electrophysiology: techniques and interpretations*. Lippincott Williams & Wilkins.

[Jylhävä et al., 2017] Jylhävä, J., Pedersen, N. L., and Hägg, S. (2017). Biological age predictors. *EBioMedicine*, 21:29–36.

[Kaiser and Hudgins, 1994] Kaiser, G. and Hudgins, L. H. (1994). *A friendly guide to wavelets*, volume 300. Springer.

[Kariyanna et al., 2020] Kariyanna, P. T., Aurora, L., Jayarangaiah, A., Das, S., Gonzalez, J. C., Hegde, S., and McFarlane, I. M. (2020). Neurotoxicity associated with radiological contrast agents used during coronary angiography: a systematic review. *American journal of medical case reports*, 8(2):60.

[Keevers, 2019] Keevers, T. L. (2019). Cross-validation is insufficient for model validation. *Joint and Operations Analysis Division, Defence Science and Technology Group: Victoria, Australia*.

[Kelley, 1999] Kelley, C. T. (1999). *Iterative methods for optimization*. SIAM.

[Keskar et al., 2016] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

[Keys, 1981] Keys, R. (1981). Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

[Kirchhof et al., 2014] Kirchhof, P., R&D, E. C., on Personalized Medicine, E. A. W. S., Sipido, K. R., R&D, E. C., on Personalized Medicine, E. A. W. S., Cowie, M. R., R&D, E. C., on Personalized Medicine, E. A. W. S., Eschenhagen, T., R&D, E. C., on Personalized Medicine, E. A. W. S., Fox, K. A., R&D, E. C., on Personalized Medicine, E. A. W. S., et al. (2014). The continuum of personalized cardiovascular medicine: a position paper of the european society of cardiology. *European heart journal*, 35(46):3250–3257.

[Kittleson et al., 2020] Kittleson, M. M., Maurer, M. S., Ambardekar, A. V., Bullock-Palmer, R. P., Chang, P. P., Eisen, H. J., Nair, A. P., Nativi-Nicolau, J., Ruberg, F. L., Failure, A. H. A. H., and of the Council on Clinical Cardiology, T. C. (2020). Cardiac amyloidosis: evolving diagnosis and management: a scientific statement from the american heart association. *Circulation*, 142(1):e7–e22.

[Klambauer et al., 2017] Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). Self-normalizing neural networks. *Advances in neural information processing systems*, 30.

[Koç et al., 2019] Koç, M. M., Aslan, N., Kao, A. P., and Barber, A. H. (2019). Evaluation of x-ray tomography contrast agents: A review of production, protocols, and biological applications. *Microscopy research and technique*, 82(6):812–848.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

[Kukačka et al., 2017] Kukačka, J., Golkov, V., and Cremers, D. (2017). Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*.

[Labach et al., 2019] Labach, A., Salehinejad, H., and Valaee, S. (2019). Survey of dropout methods for deep neural networks. *arXiv preprint arXiv:1904.13310*.

[Lancellotti et al., 2017] Lancellotti, P., Galderisi, M., Edvardsen, T., Donal, E., Goliasch, G., Cardim, N., Magne, J., Laginha, S., Hagendorff, A., Haland, T. F., et al. (2017). Echo-doppler estimation of left ventricular filling pressure: results of the multicentre eacvi euro-filling study. *European Heart Journal-Cardiovascular Imaging*, 18(9):961–968.

[Larose, 2005] Larose, D. T. (2005). An introduction to data mining. *Traduction et adaptation de Thierry Vallaud*.

[Lawrence and Lin, 1989] Lawrence, I. and Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. 521:436–44.

[LeCun et al., 1989] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551.

[LeCun et al., 1998] LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R. (1998). Efficient backprop. In *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*, pages 9–50, London, UK, UK. Springer-Verlag.

[Lee Rodgers and Nicewander, 1988] Lee Rodgers, J. and Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.

[Lehmann, 2018] Lehmann, D. H. (2018). Neural network patching. Master's thesis, Technische Universität Darmstadt, Germany.

[Lehmann et al., 2024] Lehmann, D. H., Gomes, B., Vetter, N., Braun, O., Amr, A., Hilbel, T., Müller, J., Köthe, U., Reich, C., Kayvanpour, E., et al. (2024). Prediction of diagnosis and diastolic filling pressure by ai-enhanced cardiac mri: a modelling study of hospital data. *The Lancet Digital Health*, 6(6):e407–e417.

[Leistner et al., 2020] Leistner, D. M., Dietrich, S., Erbay, A., Steiner, J., Abdelwahed, Y., Siegrist, P. T., Schindler, M., Skurk, C., Haghikia, A., Sinning, D., et al. (2020). Association of left

ventricular end-diastolic pressure with mortality in patients undergoing percutaneous coronary intervention for acute coronary syndromes. *Catheterization and Cardiovascular Interventions*, 96(4):E439–E446.

[Leshno et al., 1993] Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867.

[Leys et al., 2013] Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, 49(4):764–766.

[Libby and Theroux, 2005] Libby, P. and Theroux, P. (2005). Pathophysiology of coronary artery disease. *Circulation*, 111(25):3481–3488.

[Lifshits, 2013] Lifshits, M. A. (2013). *Gaussian random functions*, volume 322. Springer Science & Business Media.

[Lin et al., 2013] Lin, M., Chen, Q., and Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

[Liu et al., 2019] Liu, T., Tian, Y., Zhao, S., Huang, X., and Wang, Q. (2019). Automatic whole heart segmentation using a two-stage u-net framework and an adaptive threshold window. *IEEE Access*, 7:83628–83636.

[Lu et al., 2019] Lu, L., Shin, Y., Su, Y., and Karniadakis, G. E. (2019). Dying relu and initialization: Theory and numerical examples. *arXiv preprint arXiv:1903.06733*.

[Maceira et al., 2005] Maceira, A. M., Joshi, J., Prasad, S. K., Moon, J. C., Perugini, E., Harding, I., Sheppard, M. N., Poole-Wilson, P. A., Hawkins, P. N., and Pennell, D. J. (2005). Cardiovascular magnetic resonance in cardiac amyloidosis. *Circulation*, 111(2):186–193.

[Maharana et al., 2022] Maharana, K., Mondal, S., and Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91–99.

[Mahrholdt et al., 2005] Mahrholdt, H., Wagner, A., Judd, R. M., Sechtem, U., and Kim, R. J. (2005). Delayed enhancement cardiovascular magnetic resonance assessment of non-ischaemic cardiomyopathies. *European Heart Journal*, 26(15):1461–1474.

[Maisch et al., 2012] Maisch, B., Noutsias, M., Ruppert, V., Richter, A., and Pankuweit, S. (2012). Cardiomyopathies: classification, diagnosis, and treatment. *Heart failure clinics*, 8(1):53–78.

[Mandinov et al., 2000] Mandinov, L., Eberli, F. R., Seiler, C., and Hess, O. M. (2000). Diastolic heart failure. *Cardiovascular research*, 45(4):813–825.

[Marcus et al., 2022] Marcus, G., Davis, E., and Aaronson, S. (2022). A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*.

[Maron and Maron, 2013] Maron, B. J. and Maron, M. S. (2013). Hypertrophic cardiomyopathy. *The Lancet*, 381(9862):242–255.

[Maroules et al., 2018] Maroules, C. D., Hamilton-Craig, C., Branch, K., Lee, J., Cury, R. C., Maurovich-Horvat, P., Rubinshtein, R., Thomas, D., Williams, M., Guo, Y., and Cury, R. C. (2018). Coronary artery disease reporting and data system (cad-radstm): Inter-observer agreement for assessment categories and modifiers. *Journal of Cardiovascular Computed Tomography*, 12(2):125–130.

[McCullough, 2007] McCullough, P. A. (2007). Coronary artery disease. *Clinical Journal of the American Society of Nephrology*, 2(3):611–616.

[McDonagh et al., 2021] McDonagh, T. A., Metra, M., Adamo, M., Gardner, R. S., Baumbach, A., Böhm, M., Burri, H., Butler, J., Čelutkienė, J., Chioncel, O., et al. (2021). 2021 esc guidelines for the diagnosis and treatment of acute and chronic heart failure: Developed by the task force for the diagnosis and treatment of acute and chronic heart failure of the european society of cardiology (esc) with the special contribution of the heart failure association (hfa) of the esc. *European heart journal*, 42(36):3599–3726.

[McKenna et al., 2017] McKenna, W. J., Maron, B. J., and Thiene, G. (2017). Classification, epidemiology, and global burden of cardiomyopathies. *Circulation research*, 121(7):722–730.

[McKie and Burnett, 2016] McKie, P. M. and Burnett, J. C. (2016). Nt-probnp: the gold standard biomarker in heart failure.

[Members et al., 2008] Members, A. F., Dickstein, K., Cohen-Solal, A., Filippatos, G., McMurray, J. J., Ponikowski, P., Poole-Wilson, P. A., Strömberg, A., van Veldhuisen, D. J., Atar, D., et al. (2008). Esc guidelines for the diagnosis and treatment of acute and chronic heart failure 2008: the task force for the diagnosis and treatment of acute and chronic heart failure 2008 of the european society of cardiology. developed in collaboration with the heart failure association of the esc (hfa) and endorsed by the european society of intensive care medicine (esicm). *European heart journal*, 29(19):2388–2442.

[Mensah et al., 2019] Mensah, G. A., Roth, G. A., and Fuster, V. (2019). The global burden of cardiovascular diseases and risk factors: 2020 and beyond.

[Mhaskar, 1996] Mhaskar, H. N. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural computation*, 8(1):164–177.

[Mielniczuk et al., 2007] Mielniczuk, L. M., Lamas, G. A., Flaker, G. C., Mitchell, G., Smith, S. C., Gersh, B. J., Solomon, S. D., Moyé, L. A., Rouleau, J. L., Rutherford, J. D., et al. (2007). Left ventricular end-diastolic pressure and risk of subsequent heart failure in patients following an acute myocardial infarction. *Congestive Heart Failure*, 13(4):209–214.

[Mikołajczyk and Grochowski, 2018] Mikołajczyk, A. and Grochowski, M. (2018). Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122. IEEE.

[Montavon et al., 2018] Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

[Morita et al., 2005] Morita, H., Seidman, J., Seidman, C. E., et al. (2005). Genetic causes of human heart failure. *The Journal of clinical investigation*, 115(3):518–526.

[Moroni et al., 2021] Moroni, F., Gertz, Z., and Azzalini, L. (2021). Relief of ischemia in ischemic cardiomyopathy. *Current Cardiology Reports*, 23(7):80.

[Mottram and Marwick, 2005] Mottram, P. M. and Marwick, T. H. (2005). Assessment of diastolic function: what the general cardiologist needs to know. *Heart*, 91(5):681–695.

[Mueller et al., 2007] Mueller, C., Breidthardt, T., Laule-Kilian, K., Christ, M., and Perruchoud, A. P. (2007). The integration of bnp and nt-probnp into clinical medicine. *Swiss medical weekly*, 137(0102):4–12.

[Mullens et al., 2019] Mullens, W., Damman, K., Harjola, V.-P., Mebazaa, A., Brunner-La Rocca, H.-P., Martens, P., Testani, J. M., Tang, W. W., Orso, F., Rossignol, P., et al. (2019). The use of diuretics in heart failure with congestion—a position statement from the heart failure association of the european society of cardiology. *European journal of heart failure*, 21(2):137–155.

[Nadar and Shaikh, 2019] Nadar, S. K. and Shaikh, M. M. (2019). Biomarkers in routine heart failure clinical care. *Cardiac failure review*, 5(1):50.

[Nagueh et al., 1996] Nagueh, S. F., Kopelen, H. A., and Quin ones, M. A. (1996). Assessment of left ventricular filling pressures by doppler in the presence of atrial fibrillation. *Circulation*, 94(9):2138–2145.

[Nayor and Vasan, 2015] Nayor, M. and Vasan, R. S. (2015). Preventing heart failure: the role of physical activity. *Current opinion in cardiology*, 30(5):543–550.

[Neelakantan et al., 2015] Neelakantan, A., Vilnis, L., Le, Q. V., Sutskever, I., Kaiser, L., Kurach, K., and Martens, J. (2015). Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*.

[Ng, 2004] Ng, A. Y. (2004). Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 78–, New York, NY, USA. ACM.

[Niculescu-Mizil and Caruana, 2005] Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

[Nielsen, 2015] Nielsen, M. A. (2015). *Neural networks and deep learning*, volume 25. Determination press San Francisco, CA, USA.

[Oglat et al., 2018] Oglat, A. A., Matjafri, M., Suardi, N., Oqlat, M. A., Abdelrahman, M. A., and Oqlat, A. A. (2018). A review of medical doppler ultrasonography of blood flow in general and especially in common carotid artery. *Journal of medical ultrasound*, 26(1):3–13.

[Oh et al., 2023] Oh, J. K., Miranda, W. R., and Kane, G. C. (2023). Diagnosis of heart failure with preserved ejection fraction relies on detection of increased diastolic filling pressure, but how?

[Olchowy et al., 2017] Olchowy, C., Cebulski, K., Łasecki, M., Chaber, R., Olchowy, A., Kałwak, K., and Zaleska-Dorobisz, U. (2017). The presence of the gadolinium-based contrast agent depositions in the brain and symptoms of gadolinium neurotoxicity-a systematic review. *PloS one*, 12(2):e0171704.

[Ommen et al., 2000] Ommen, S. R., Nishimura, R. A., Appleton, C. P., Miller, F., Oh, J. K., Redfield, M. M., and Tajik, A. (2000). Clinical utility of doppler echocardiography and tissue doppler imaging in the estimation of left ventricular filling pressures: a comparative simultaneous doppler-catheterization study. *Circulation*, 102(15):1788–1794.

[Oremus et al., 2014] Oremus, M., McKelvie, R., Don-Wauchope, A., Santaguida, P. L., Ali, U., Balion, C., Hill, S., Booth, R., Brown, J. A., Bustamam, A., et al. (2014). A systematic review of bnp and nt-probnp in the management of heart failure: overview and methods. *Heart failure reviews*, 19:413–419.

[Organization et al., 2010] Organization, W. H. et al. (2010). Pharmaceuticals: Restrictions in use and availability. Technical report, World Health Organization.

[Owens et al., 2008] Owens, J. D., Houston, M., Luebke, D., Green, S., Stone, J. E., and Phillips, J. C. (2008). Gpu computing. *Proceedings of the IEEE*, 96(5):879–899.

[Oymak and Soltanolkotabi, 2020] Oymak, S. and Soltanolkotabi, M. (2020). Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105.

[Pan et al., 2008] Pan, S. J., Kwok, J. T., Yang, Q., et al. (2008). Transfer learning via dimensionality reduction. In *AAAI*, volume 8, pages 677–682.

[Park and Kwak, 2017] Park, S. and Kwak, N. (2017). Analysis on the dropout effect in convolutional neural networks. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 189–204. Springer.

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

[Patel et al., 2010] Patel, M. R., Peterson, E. D., Dai, D., Brennan, J. M., Redberg, R. F., Anderson, H. V., Brindis, R. G., and Douglas, P. S. (2010). Low diagnostic yield of elective coronary angiography. *New England Journal of Medicine*, 362(10):886–895.

[Perez and Wang, 2017] Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

[Perez et al., 2019] Perez, M. V., Mahaffey, K. W., Hedlin, H., Rumsfeld, J. S., Garcia, A., Ferris, T., Balasubramanian, V., Russo, A. M., Rajmane, A., Cheung, L., et al. (2019). Large-scale assessment of a smartwatch to identify atrial fibrillation. *New England Journal of Medicine*, 381(20):1909–1917.

[Pinedo et al., 2010] Pinedo, M., Villacorta, E., Tapia, C., Arnold, R., López, J., Revilla, A., Gómez, I., Fulquet, E., and San Román, J. A. (2010). Inter- and intra-observer variability in the echocardiographic evaluation of right ventricular function. *Revista Española de Cardiología (English Edition)*, 63(7):802–809.

[Pinkus, 1999] Pinkus, A. (1999). Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195.

[Pitt and Myung, 2002] Pitt, M. A. and Myung, I. J. (2002). When a good fit can be bad. *Trends in cognitive sciences*, 6(10):421–425.

[Ponikowski et al., 2014] Ponikowski, P., Anker, S. D., AlHabib, K. F., Cowie, M. R., Force, T. L., Hu, S., Jaarsma, T., Krum, H., Rastogi, V., Rohde, L. E., et al. (2014). Heart failure: preventing disease and death worldwide. *ESC heart failure*, 1(1):4–25.

[Poplin et al., 2018] Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., and Webster, D. R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158–164.

[Posina et al., 2013] Posina, K., McLaughlin, J., Rhee, P., Li, L., Cheng, J., Schapiro, W., Gulotta, R. J., Berke, A. D., Petrossian, G. A., Reichek, N., et al. (2013). Relationship of phasic left atrial volume and emptying function to left ventricular filling pressure: a cardiovascular magnetic resonance study. *Journal of Cardiovascular Magnetic Resonance*, 15(1):99.

[Press et al., 1986] Press, W. H., Flannery, B., Teukolsky, S., and Vetterling, W. (1986). Numerical recipices, the art of scientific computing. *Cambridge U. Press, Cambridge, MA*.

[Rajpurkar et al., 2022] Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. (2022). Ai in health and medicine. *Nature medicine*, 28(1):31–38.

[Rieke et al., 2020] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., et al. (2020). The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7.

[Robin et al., 2018] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., Siegert, S., and Doering, M. (2018). proc: display and analyze roc curves. *R package version*, 1(5).

[Rokach and Maimon, 2005] Rokach, L. and Maimon, O. (2005). Decision trees. *Data mining and knowledge discovery handbook*, pages 165–192.

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

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

[Roth et al., 2020] Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A. Z., Benjamin, E. J., Benziger, C. P., et al. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the gbd 2019 study. *Journal of the American College of Cardiology*, 76(25):2982–3021.

[Rubin and Maurer, 2020] Rubin, J. and Maurer, M. S. (2020). Cardiac amyloidosis: overlooked, underappreciated, and treatable. *Annual Review of Medicine*, 71:203–219.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–.

[Russell and Norvig, 2016] Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.

[Salerno et al., 2017] Salerno, M., Sharif, B., Arheden, H., Kumar, A., Axel, L., Li, D., and Neubauer, S. (2017). Recent advances in cardiovascular magnetic resonance: techniques and applications. *Circulation: Cardiovascular Imaging*, 10(6):e003951.

[Sander et al., 2020] Sander, J., de Vos, B. D., and Išgum, I. (2020). Automatic segmentation with detection of local segmentation failures in cardiac mri. *Scientific Reports*, 10(1):21769.

[Santurkar et al., 2018] Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. (2018). How does batch normalization help optimization? *Advances in neural information processing systems*, 31.

[Schmidhuber, 2014] Schmidhuber, J. (2014). Deep learning in neural networks: An overview. *CoRR*, abs/1404.7828.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

[Selvaraju et al., 2016] Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., and Batra, D. (2016). Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*.

[Shah et al., 2024] Shah, S. J., Fine, N., Garcia-Pavia, P., Klein, A. L., Fernandes, F., Weissman, N. J., Maurer, M. S., Boman, K., Gundapaneni, B., Sultan, M. B., et al. (2024). Effect of tafamidis on cardiac function in patients with transthyretin amyloid cardiomyopathy: A post hoc analysis of the attr-act randomized clinical trial. *JAMA cardiology*, 9(1):25–34.

[Shang and Wah, 1996] Shang, Y. and Wah, B. W. (1996). Global optimization for neural network training. *Computer*, 29(3):45–54.

[Shannon, 1949] Shannon, C. (1949). Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21.

[Shorten and Khoshgoftaar, 2019] Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

[Shrikumar et al., 2016] Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.

[Shrout and Fleiss, 1979] Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

[Singh et al., 2023] Singh, S. K., Kumar, S., and Mehra, P. S. (2023). Chat gpt & google bard ai: A review. In *2023 International Conference on IoT, Communication and Automation Technology (ICICAT)*, pages 1–6. IEEE.

[Sinning et al., 2010] Sinning, C., Keller, T., Abegunewardene, N., Kreitner, K.-F., Münzel, T., and Blankenberg, S. (2010). Tako-tsubo syndrome: dying of a broken heart? *Clinical Research in Cardiology*, 99:771–780.

[Smilkov et al., 2017] Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

[Smiseth and Tendera, 2008] Smiseth, O. A. and Tendera, M. (2008). *Diastolic heart failure*. Springer.

[Solomon et al., 2021] Solomon, S. D., de Boer, R. A., DeMets, D., Hernandez, A. F., Inzucchi, S. E., Kosiborod, M. N., Lam, C. S., Martinez, F., Shah, S. J., Lindholm, D., et al. (2021). Dapagliflozin in heart failure with preserved and mildly reduced ejection fraction: rationale and design of the deliver trial. *European journal of heart failure*, 23(7):1217–1225.

[Sorajja et al., 2020] Sorajja, P., Lim, M. J., and Kern, M. J. (2020). *Kern's cardiac catheterization handbook*. Elsevier, Philadelphia, Pa, seventh edition. edition.

[Spertus et al., 2021] Spertus, J. A., Fine, J. T., Elliott, P., Ho, C. Y., Olivotto, I., Saberi, S., Li, W., Dolan, C., Reaney, M., Sehnert, A. J., et al. (2021). Mavacamten for treatment of symptomatic obstructive hypertrophic cardiomyopathy (explorer-hcm): health status analysis of a randomised, double-blind, placebo-controlled, phase 3 trial. *The Lancet*, 397(10293):2467–2475.

[Srivastava et al., 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

[Strömberg and Mårtensson, 2003] Strömberg, A. and Mårtensson, J. (2003). Gender differences in patients with heart failure. *European Journal of Cardiovascular Nursing*, 2(1):7–18.

[Sutskever et al., 2013] Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA. PMLR.

[Szegedy et al., 2017] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

[Szegedy et al., 2016] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

[Szeliski, 2022] Szeliski, R. (2022). *Computer vision: algorithms and applications*. Springer Nature.

[Tabuada and Gharesifard, 2020] Tabuada, P. and Gharesifard, B. (2020). Universal approximation power of deep residual neural networks via nonlinear control theory. *arXiv preprint arXiv:2007.06007*.

[Tan et al., 2018] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 270–279. Springer.

[Tan and Le, 2019] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

[Tavakol et al., 2012] Tavakol, M., Ashraf, S., and Brener, S. J. (2012). Risks and complications of coronary angiography: a comprehensive review. *Global journal of health science*, 4(1):65.

[Taylor, 2018] Taylor, A. M. (2018). Chapter 10 - left-ventricular hemodynamics, heart failure, and shock. In Ragosta, M., editor, *Textbook of Clinical Hemodynamics (Second Edition)*, pages 216–248. Elsevier, second edition edition.

[Taylor, 2001] Taylor, B. (2001). The international system of units (si), 2001 edition.

[Teng et al., 2022] Teng, Q., Liu, Z., Song, Y., Han, K., and Lu, Y. (2022). A survey on the interpretability of deep learning in medical diagnosis. *Multimedia Systems*, 28(6):2335–2355.

[Tikk et al., 2003] Tikk, D., Kóczy, L. T., and Gedeon, T. D. (2003). A survey on universal approximation and its limits in soft computing techniques. *International Journal of Approximate Reasoning*, 33(2):185–202.

[Touvron et al., 2023] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[Tromp et al., 2022] Tromp, J., Seekings, P. J., Hung, C.-L., Iversen, M. B., Frost, M. J., Ouwerkerk, W., Jiang, Z., Eisenhaber, F., Goh, R. S., Zhao, H., et al. (2022). Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study. *The Lancet Digital Health*, 4(1):e46–e54.

[Tsao et al., 2018] Tsao, C. W., Lyass, A., Enserro, D., Larson, M. G., Ho, J. E., Kizer, J. R., Gottdiener, J. S., Psaty, B. M., and Vasan, R. S. (2018). Temporal trends in the incidence of and mortality associated with heart failure with preserved and reduced ejection fraction. *JACC: Heart Failure*, 6(8):678–685.

[Van Calster et al., 2023] Van Calster, B., Steyerberg, E. W., Wynants, L., and van Smeden, M. (2023). There is no such thing as a validated prediction model. *BMC medicine*, 21(1):70.

[Van der Meer, 2002] Van der Meer, F. (2002). *Technical Principles of X-Ray Angiography*, pages 593–601. Springer Berlin Heidelberg, Berlin, Heidelberg.

[van Smeden et al., 2022] van Smeden, M., Heinze, G., Van Calster, B., Asselbergs, F. W., Vardas, P. E., Bruining, N., de Jaegere, P., Moore, J. H., Denaxas, S., Boulesteix, A. L., and Moons, K. G. M. (2022). Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease . *European Heart Journal*, 43(31):2921–2930.

[Vardas et al., 2022] Vardas, P. E., Asselbergs, F. W., van Smeden, M., and Friedman, P. (2022). The year in cardiovascular medicine 2021: digital health and innovation. *European heart journal*, 43(4):271–279.

[Vieillard-Baron et al., 2019] Vieillard-Baron, A., Millington, S., Sanfilippo, F., Chew, M., Diaz-Gomez, J., McLean, A., Pinsky, M., Pulido, J., Mayo, P., and Fletcher, N. (2019). A decade of progress in critical care echocardiography: a narrative review. *Intensive care medicine*, 45:770–788.

[Voulodimos et al., 2018] Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., et al. (2018). Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018.

[Walpole et al., 1993] Walpole, R. E., Myers, R. H., Myers, S. L., and Ye, K. (1993). *Probability and statistics for engineers and scientists*, volume 5. Macmillan New York.

[Wang et al., 2007] Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007). Statistics in medicine—reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21):2189–2194.

[Weintraub et al., 2017] Weintraub, R. G., Semsarian, C., and Macdonald, P. (2017). Dilated cardiomyopathy. *The Lancet*, 390(10092):400–414.

[Weng et al., 1992] Weng, J., Ahuja, N., and Huang, T. S. (1992). Cresceptron: a self-organizing neural network which grows adaptively. In *International Joint Conference on Neural Networks (IJCNN)*, volume 1, pages 576–581. IEEE.

[Weng et al., 2016] Weng, Z., Yao, J., Chan, R. H., He, J., Yang, X., Zhou, Y., and He, Y. (2016). Prognostic value of lge-cmr in hcm: a meta-analysis. *JACC: Cardiovascular Imaging*, 9(12):1392–1402.

[Wong et al., 2016] Wong, S. C., Gatt, A., Stamatescu, V., and McDonnell, M. D. (2016). Understanding data augmentation for classification: when to warp? In *2016 international conference on digital image computing: techniques and applications (DICTA)*, pages 1–6. IEEE.

[Woolf, 2008] Woolf, S. H. (2008). The power of prevention and what it requires. *Jama*, 299(20):2437–2439.

[Wright, 1921] Wright, S. (1921). Correlation and causation. *Journal of agricultural research*, 20(7):557.

[Yadav et al., 2015] Yadav, N., Yadav, A., Kumar, M., et al. (2015). *An introduction to neural network methods for differential equations*, volume 1. Springer.

[Yang and Wang, 2020] Yang, G. R. and Wang, X.-J. (2020). Artificial neural networks for neuroscientists: a primer. *Neuron*, 107(6):1048–1070.

[Yang and Berdine, 2017] Yang, S. and Berdine, G. (2017). The receiver operating characteristic (roc) curve. *The Southwest Respiratory and Critical Care Chronicles*, 5(19):34–36.

[Yarotsky, 2017] Yarotsky, D. (2017). Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114.

[Ying, 2019] Ying, X. (2019). An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing.

[Yosinski et al., 2015] Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

[Youssef et al., 2023] Youssef, A., Pencina, M., Thakur, A., Zhu, T., Clifton, D., and Shah, N. H. (2023). External validation of ai models in health should be replaced with recurring local validation. *Nature Medicine*, 29(11):2686–2687.

[Zdravkovic et al., 2013] Zdravkovic, V., Mladenovic, V., Colic, M., Bankovic, D., Lazic, Z., Petrovic, M., Simic, I., Knezevic, S., Pantovic, S., Djukic, A., et al. (2013). Nt-probnp for prognostic and diagnostic evaluation in patients with acute coronary syndromes. *Polish Heart Journal (Kardiologia Polska)*, 71(5):472–479.

[Zeiler and Fergus, 2013] Zeiler, M. D. and Fergus, R. (2013). Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901.

[Zeki, 1993] Zeki, S. (1993). *A vision of the brain.* Blackwell scientific publications.

[Zhang et al., 2019] Zhang, N., Yang, G., Gao, Z., Xu, C., Zhang, Y., Shi, R., Keegan, J., Xu, L., Zhang, H., Fan, Z., et al. (2019). Deep learning for diagnosis of chronic myocardial infarction on nonenhanced cardiac cine mri. *Radiology*, 291(3):606–617.

[Zhao et al., 2017] Zhao, B., Lu, H., Chen, S., Liu, J., and Wu, D. (2017). Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics*, 28(1):162–169.

[Zhao et al., 2023] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

[Zhou, 2020] Zhou, D.-X. (2020). Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794.

[Zohuri and Moghaddam, 2020] Zohuri, B. and Moghaddam, M. (2020). Deep learning limitations and flaws. *Mod. Approaches Mater. Sci*, 2:241–250.

[Zoph et al., 2020] Zoph, B., Cubuk, E. D., Ghiasi, G., Lin, T.-Y., Shlens, J., and Le, Q. V. (2020). Learning data augmentation strategies for object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 566–583. Springer.

[Zou and Hastie, 2005] Zou, H. and Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.