Validation and evaluation metrics for medical and biomedical image synthesis

Tereza Nečasová, Ninon Burgos, David Svoboda

Chapter 25

Validation and evaluation metrics for medical and biomedical image synthesis

Tereza Nečasová[a], **Ninon Burgos**[b], **and David Svoboda**[a,*]

[a]Centre for Biomedical Image Analysis, Faculty of Informatics, Masaryk University, Czech Republic
[b]Sorbonne Université, Institut du Cerveau – Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, F-75013, Paris, France
[*]Corresponding author: svoboda@fi.muni.cz

Abstract

Synthetic image data play an important role in the verification of medical and biomedical image analysis algorithms. However the usage of such data strongly relies on their quality and plausibility. Despite the emergence of many frameworks for image synthesis in recent years, the quality of the generated images has not been sufficiently assessed in many cases, or the methodology varied across the publications. If we want to use synthetic image data for the verification of biomedical analysis tools, then the images should resemble the real ones as much as possible with evidence about their similarity.

Initially, the hardware available for simulations was limited. Therefore, the validation was not under the scope of interest. With the technological improvements, the expectations put on synthetic data have arisen. Proper validation of synthetic image data is nowadays becoming essential.

Keywords: Image synthesis, Image simulation, Similarity, Distance Metrics, Image Datasets

N. Burgos and D. Svoboda (Eds.), *Biomedical Image Synthesis and Simulation*, Elsevier
https://doi.org/10.1016/B978-0-12-824349-7.00032-3

1. Introduction

Regardless of the application domain, each newly introduced method requires a proper validation procedure. When designing a segmentation algorithm, for example, one is excepted to provide some tests showing how the algorithm performs and whether the results are better compared with state of the art methods. In the field of image synthesis, one can legitimately ask about the explained variability in the synthetic data, whether the generated data look realistic, and whether they are sufficiently similar to their real counterpart. However, no standard pipeline currently exists for synthetic image quality assessment (see Fig. 1). Some authors evaluate their computer-generated data with the statement of having visually similar results, some attempt to measure the similarity using expert's assessment or various techniques. Without any broader comparison of synthesized data, there is no warranty regarding their quality. The quality assessment of the synthesized images should be an integral part of the image synthesis process. The user of the proposed synthetic image data should be assured about their plausibility before using them in further steps.

This chapter includes an overview of the validation methods applied in the field of biomedical and medical image synthesis. Three strategies can be considered, and possibly combined, when evaluating the quality of generators producing synthetic images. In the first case, the data are validated using the strength of **expert knowledge** to support the plausibility of the generated images. The second class is focused on synthetic images **paired** with their real counterparts. The third class of evaluation is done by comparing the characteristics of the **whole dataset** of real images against the characteristics in the whole dataset of synthetic images. Although the assessment in the first class can be further evaluated, it is rather more qualitative than quantitative compared with the two other classes. The main approaches of each strategy will be described in the following subsections.

2. Expert Knowledge

When developing a new framework responsible for image synthesis where the outputs are expected to resemble the real images, the most expected approach is to report the visual inspection of the synthetic data.

2.1 Rating based on the visual plausibility of synthetic data

A first task given to the experts can be to rate the plausibility of the synthetic data. For example, the sensitivity and specificity of a classification task is utilized when validating a framework for simulation of bright-field microscopy

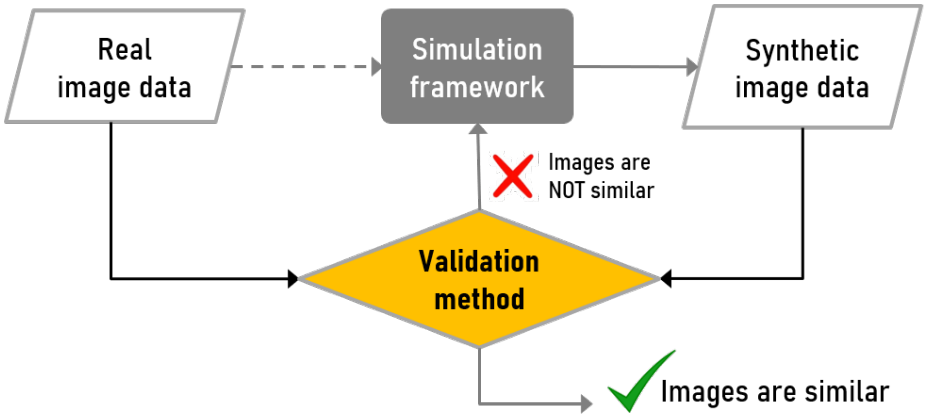

Figure 1: Processes connected with the validation of synthetic data. Synthetic images obtained from the simulation framework should be compared with real images with a clear conclusion given by a validation method (see in the yellow box) whether they are sufficiently similar or not.

images depicting Pap Smear specimens [1]. The sensitivity relates to the ability of accurate detection of real images as being real and the specificity to the ability of accurate detection of synthetic images as synthetic. These measures are computed from the number of true positives/real images (TP), true negatives/synthetic images (TN), false positives (FP) and false negative cases (FN) as:

$$\text{sensitivity} = \frac{TP}{TP + FN} \tag{1}$$

$$\text{specificity} = \frac{TN}{FP + TN}. \tag{2}$$

Both values range from 0 to 1 (=*ideal*). In this work, six experts from different research fields were asked to recognize the origin of a set of real and synthetic images. The experts were showed a tightly cropped, randomly selected, view of the real and synthetic scenes projected through lightly frosted glass, to account for limitations of existing display devices. They were showed only for two seconds to imitate the real conditions for assessment. The results of the experts' classification is reviewed and compared to a random classification in Table 1.

A similar validation approach can be found in [2], where the experts were also asked to recognize real or synthetic images resulting from a generator of magnetic resonance imaging (MRI) brain slices. They also compute TP, TN, FP, FN to derive accuracy, precision, sensitivity and F1-score.

In [3], the synthetic whole slide histological images accompanied by the reference images are evaluated by observing the image data at different levels of magnification. The goal is to discover in which level of detail (magnification)

Classifier	TP	FP	TN	FN	Sensitivity	Specificity
Expert 1	35	29	37	17	0.67	0.56
Expert 2	39	29	29	21	0.65	0.50
Expert 3	50	16	38	14	0.78	0.70
Expert 4	49	20	36	13	0.79	0.64
Expert 5	51	17	39	11	0.82	0.70
Expert 6	58	32	21	7	0.89	0.40
Random	37	31	24	26	0.54	0.48

Table 1: Results of the experts' evaluation in [1]. The classifiers considered included six experts and a random classifier. TP is the number of correctly classified real images - true positives, TN is true negatives - correctly classified synthetic images, FP is false positives and FN is false negative cases.

the synthetic image is not looking realistic anymore. The measurement of realism of synthetic images is evaluated as follows:

1. open a synthetic image in a software slide viewer on a standard computer screen;
2. decrease the magnification until it "feels realistic";
3. slowly increase the magnification until it "feels wrong";
4. the current magnification is then written down as the highest realistic magnification for the current image.

The highest realistic magnification is regarded as the score that evaluates the quality of the image simulation. In order the keep some biological correctness, only experts were asked to assess the data plausibility.

The five-point Likert scale (*very poor / poor / satisfactory / good / very good*) has been used to assess the quality of synthetic tissue microscopy images generated from pre-segmented Haematoxylin and Eosin images of brain tumours [4]. Here, the expert histopathologists were asked to classify the images according to the following three criteria:

1. similarity to real-world tissue microscopy images,
2. reproducibility of nuclei morphometry and
3. reproducibility in nuclei texture.

However, the particular results of the expert have not been reported.

The Likert scale has also been applied in [5], where 21 experts in ultrasound were asked to rank (*fake*, *rather fake*, *cannot decide*, *rather real*, *real*) the

	H 40×	**H** 20×	**WD** 40×	**WD** 20×	**MD** 40×	**MD** 20×	**PD** 40×	**PD** 20×
Architecture	5	5	5	4	4	4	5	5
Crypt shape	5	5	5	5	5	5	4.5	4.5
Lumen	5	5	5	5	5	5	-	-
Goblet cells	4	4	-	-	-	-	-	-
Epithelial cells	4	4	4	4	4	4	4	4
Stromal cells	3	3	3	3	3	3	3	4

Table 2: Results of the experts' evaluation in [7]. The values report the average evaluation of the appearance of synthetic images by three pathologists. Healthy (H), well differentiated (WD), moderately differentiated (MD), and poorly differentiated (PD) images were evaluated at magnifications 20× and 40×. (*1 = Not realistic at all, 5 = Very realistic, '-' means feature is not relevant*)

realism of the simulated ultrasound scans. The scans were randomly presented to the experts without giving them any hint whether observing the real or simulated one. The time spent during the analysis of each particular image was also taken into account when assessing the experts' reliability.

The objective of the work of Gong et al. [6] was to reduce gadolinium dose in contrast-enhanced brain MRI. Two neuro-radiologists were asked to assess the quality of the post-contrast MR images (low-dose, synthesized full-dose and true full-dose). They rated the general image quality, suppression of aliasing/motion artifacts and degree of enhancement compared against pre-contrast MR images using a five-point Likert scale ranging from 1 (poor) to 5 (excellent).

In [7], a model of healthy and cancerous colonic crypt micro-environment was proposed and successfully implemented to show the ability to control cancer grade, cellularity, cell overlap ratio, and image resolution. Here, the histopathologists were asked to grade the quality of synthesis on a scale from 1 to 5 (*5 = very realistic*). The averages of the grades for a combination of four types of cell objects, architecture and number of crypts, two different magnifications, and four differentiation levels are reported in Table 2.

When comparing the real and synthetic (also called as *fake*) images, one can also use a Visual Turing Test (VTT) [8]. This test is a procedure during which a stochastic sequence of binary questions is generated and given to some respondents. When applied in the field of medical image analysis, the experts are asked to distinguish between the real and synthetic images in a sequence.

VTT was applied for example in [9] for the validation of generated brain MR images. In another work [10], the authors tested the quality of synthesized lung nodules for X-ray computed tomography (CT) image augmentation

potentially used for object detection. Chuquicusma et al. [11] performed VTT on images of malignant and benign lung nodules for a computer-aided diagnosis system generated by a deep convolutional generative adversarial network (DC-GAN). In the context of anomaly detection, Schlegl et al. [12] used the VTT to quantify the quality of the generated normal images.

2.2 Rating based on the usability of synthetic data

Experts have been asked to not only rate the plausibility of synthetic data but also their usability.

A simulation framework called SIMCEP [13] forms a cornerstone in image synthesis dedicated to fluorescence microscopy. To validate the generated images, mediated experts were asked to use four different image processing tools developed for automated image cytometry (specifically for cell enumeration). Five sets of ten images, each containing 1000 cells and different levels of overlap were analyzed with each out of four tools developed by independent research groups and the results were compared in a plot. The tools gave similar results supporting the expectation that for worse conditions such as overlapping cells, the number of enumerated cells will be lower.

In order to perform a clinical assessment of a method using a generative adversarial network (GAN) to synthesize standard-dose PET images from low-dose ones [14], the experts were asked to perform two tasks. First, they assigned a score (1-5) for each synthesized image, where range 1-3 was considered to be low quality and 4-5 high quality. Second, the experts gave the amyloid status (positive vs. negative) for each image, the amyloid status being a biomarker used in the differential diagnosis of dementia. The status defined on the standard-dose ground truths and the synthesized images were compared. The consistency between the amyloid status showed whether the method was able to maintain the pathological features.

As stated in [15], several works try to convince the readers about correlation of proposed metrics with human evaluation [16, 17]. However, they also state that expert evaluation can be biased towards the visual quality of synthesized images and neglect the overall distributional characteristics, which are important for unsupervised learning.

3. Pairwise Comparison

In medical image synthesis, a majority of approaches require paired images in their training process, for example when learning to synthesise CT from MR images [18, 19, 20, 21, 22], generate MR images of a certain sequence from MR images of another sequence [23, 24, 25], denoise low-dose CT images [26, 27] or perform super-resolution [28]. Cross-modality synthesis is also present in

microscopy imaging, where the attempt to reduce time-consuming and laborious tissue preparation results in synthesizing fluorescence images from the bright-field pairs [29]. A consequence of the need for paired images during training is that both reference and synthetic images are also often available for evaluation. The quality of the pairwise estimates is typically controlled by measuring the difference to the so called *reference image* (also known as *ground truth* or *annotation*) in the pair. The term reference image is also a reason why the pairwise comparison is sometimes called a *full reference* image quality assessment.

3.1 Generic pairwise performance measures

The measures the most frequently used to evaluate the synthesis accuracy by comparing real and synthetic images in a pairwise manner are listed below.

3.1.1. Mean absolute error

One of the most common measures is the mean absolute error (MAE) [18, 19, 20, 21, 22, 30]. It is defined as the absolute difference between intensities in pixels of the simulated and ground truth image:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}, \tag{3}$$

where x_i and y_i are the intensity values of the i-th pixel/voxel of the real and synthesized image, respectively, and n is the number of pixel/voxel pairs.

In the context of PET attenuation correction [18] or MR-only radiotherapy treatment planning [19], the MAE is often chosen as error metric as it is well suited when comparing CT images due to their quantitative nature.

3.1.2. Peak signal-to-noise ratio

The peak signal-to-noise ratio (PSNR) is a measure derived from the mean squared error (MSE):

$$MSE = \frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}, \tag{4}$$

with the same notations as for the MAE: x_i and y_i are the intensity values of the i-th pixel/voxel of the real and synthesized image, respectively, and n is the number of pixel/voxel pairs.

PSNR compares the maximum of the intensity in the image with the error between the estimated and the ground truth image given by the MSE:

$$PSNR = 10 \log_{10} \left(\frac{MAX^2}{MSE} \right), \tag{5}$$

where MAX is the maximum possible intensity of the image. Higher values of PSNR relate to better simulation. The application of PSNR to validate the plausibility of synthetic images is apparent in many papers [21, 22, 24, 25, 26, 27, 28, 29, 30].

3.1.3. Structural similarity

Even though they are simple to compute, the ability of the MAE, MSE and PSNR measures to perceive visual quality is limited [31]. Exploiting known characteristics of the human visual system, Wang et al. [31] proposed the structural similarity (SSIM), which compares local patterns of pixel intensities that have been normalised for luminance and contrast. SSIM has been widely adopted by the image synthesis community [24, 25, 27, 28, 29].

SSIM is computed as a function of three components, luminance, contrast and structure, as follows:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{6}$$

where x is the simulated image, y is the ground truth image, μ_i is the mean value of image i, σ_i is the variance of image i, σ_{xy} is the covariance of images x and y. The constants C_j are included to avoid instability when $\mu_x^2 + \mu_y^2$, respectively $\sigma_x^2 + \sigma_y^2$, is very close to zero. Therefore, they are set to $C_j = (K_j \cdot L)^2$, where L is the dynamic range of the pixel values (e.g. 255 for 8-bit grayscale images), and $K \ll 1$ is a small constant. The original values of K_j were set to $K_1 = 0.01$ and $K_2 = 0.03$ [31]. The higher the value of SSIM, the better the quality of the synthesized image.

In [32], the relationship between SSIM and PSNR was investigated with the conclusion that the values of the PSNR can be predicted from the SSIM and vice-versa. Additionally, the PSNR and SSIM mainly differ in the sensitivity to image degradations. Similar conclusions can be found in [33], where the association between SSIM and MSE was shown.

In their work, Mason et al. [34] measured the correlations between ten pairwise metrics and the subjective score of five radiologists when assessing the quality of MR images. They showed that metrics such as SSIM and PSNR are potentially not ideal surrogate measures of MR image quality as determined by radiologist evaluation.

At this point, the reader can seriously wonder whether it is manageable to perform a proper data validation with all the previously-described metrics, together with all those that will be introduced in the following text. Fortunately, it is not needed. The above mentioned metrics (MAE, PSNR, SSIM) are the most common distance metrics and are accepted as a de-facto standard among the pair-wise metrics. Therefore, if you plan to generate some dataset that

form pairs (e.g. MRI–CT), you will likely be expected to evaluate at least one of these. All the metrics that come in the following paragraphs are not less important but rather mostly proposed/derived for specific purposes, so their usage is somewhat limited.

3.2 Application-specific pairwise performance measures

3.2.1. Pseudo-healthy image synthesis

To detect anomalies and better understand changes induced by diseases, Xia et al. [35] proposed to create subject-specific pseudo-healthy images from pathological ones using a CycleGAN. They assessed the quality of the image synthesis process using generic metrics, but they also designed new metrics tailored to their application: the so called *healthiness*, *identity*, and *deformation correction* metrics.

Healthiness The healthiness (h) expresses a fraction of (unwanted) pathology areas present in pseudo-healthy images and can be computed as:

$$h = 1 - \frac{\mathbb{E}_{x_p \sim \mathcal{P}}[N(f_p(G(x_p)))]}{\mathbb{E}_{m_p \sim \mathcal{P}_m}[N(f_p(x_p))]} \tag{7}$$

where x_p is the pathological image, f_p the function providing a segmentation of pathological regions, $G(\cdot)$ the CycleGAN deriving the pseudo-healthy image from the pathological image, m_p the ground truth mask of pathological regions and $N(\cdot)$ the number of pixels labelled as pathological by f_p. One should pay attention that the *healthiness* can be strongly influenced by the quality of segmentation function f_p.

Identity To measure the CycleGAN ability to preserve the subject identity (iD), one can measure the structural similarity of the original real pathology image and the derived pseudo-healthy image outside the pathological regions:

$$iD = \text{MS-SSIM}[(1 - m_p) \odot G(x_p), (1 - m_p) \odot x_p] \tag{8}$$

where MS-SSIM stands for multiscale structural similarity [36], x_p is a pathological image, m_p is its corresponding pathology mask, $G(\cdot)$ is the generator of pseudo-healthy images, and $\odot$ is the pixel-wise multiplication.

Deformation correction In the case where brain tissue has recovered after some surgery or noninvasive therapy, there may be apparent structural changes. The deformation correction measure aims to assess whether such deformations have taken place. To avoid brightness influence, the images are first converted into edge maps. The classifier, that was trained over the set of

edge maps of healthy images, is used to judge whether the image of treated brain contains some deformations. The output of such classifier is a continuous number between 0 and 1 [35].

In this section, the validation methods used for pairs of images were reported. Note that the frameworks for synthesis of paired data are often based on algorithms using a loss function. The loss function should differ from the validation method since we want to avoid overfitting.

4. Dataset comparison

Some synthetic data are the result of frameworks generating a whole set of images according to the given input parameters. A set of generated images resulting from these tools should be comparable with the set of real images corresponding to the particular application of the synthesized data.

If we want to compare the images not only in the corresponding pixels but also in some quantitative characteristics, the use of image descriptors is suggested. Those image descriptors can represent characteristics such as color, shape, texture or some features from the frequency domain. A survey of image feature descriptors was published in [37]. The values of a particular hidden layer of a neural network can also be considered as a descriptor (e.g. for the computation of the Fréchet Inception Distance - discussed further in 4.1.7). The most commonly derived descriptors used in validation methods of synthetic images are:

- **Haralick texture descriptors**
  Haralick descriptors [38] represent a set of texture descriptors derived from so called gray level co-occurrence matrix. In [39], contrast, correlation, homogeneity, and maximal correlation coefficient are reported to show the similarity between the real and synthetic image data. Haralick descriptors are an input parameter to affinity propagation in [7].

- **Central moments**
  A number of central moments, e.g. variance, skewness, kurtosis, can be calculated over pixel intensities. They were used for instance in [40, 41, 1].

- **Subcellular location features** [42]
  This respectably extensive pack of descriptors include various types of patterns, such as texture features, Zernike moment features, object skeleton features and many others. In [43], the contribution of features on the classification of multiple cell objects was compared.

- **Local binary pattern** [44]
  Local binary pattern (LBP) is a texture operator which assigns to each pixel of an image a binary 8-bit number by thresholding its eight neighbors. LBP has been applied to ultrasound images to validate the method described in [45].

- **Mean square displacement** [46]
  The mean square displacement (MSD) descriptor is used for the validation of tracking objects. Svoboda et al. [39] compared MSD using a histogram, displacement profiles and ensemble-average MSD curves.

- **Spectral and spatial sharpness** [47]
  In [48], the spectral and spatial sharpness accompanied the standard SSIM and PSNR metrics to validate the quality of super-resolved MR images.

The above-mentioned list of descriptors is not limited. One can easily derive a data-specific descriptor that fits particular needs. For example, in the field of live cell analysis, the derived features such as the number of tracks, the average number of time points in tracks, or the average speed of particles may help to effectively compare the time-lapse image sequences [49].

One should keep in mind that image descriptors need not be necessarily used for the comparison of synthetic and real datasets. In case one is missing the real data, it is common to analyze solely the synthetic data and study the behaviour of the measured characteristics. This approach is also known as *no reference* image quality assessment.

The image descriptors are commonly used as input parameters to the simulation frameworks (number of elements/objects, shape, speed of motion, etc.) or are an integral part of loss functions in GANs or variational autoencoders. One should pay attention that the use of image descriptors as the inputs for the simulation together with the subsequent validation of the same features over the generated data is not only meaningful but may be even misleading. The results would be biased and therefore worthless.

Below, the reader can find the detailed explanation of the most prominent methods that utilize image descriptors and that are relevant to the comparison of sets of biomedical images.

4.1 Measures based on a comparison of probability distributions

4.1.1. Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test (K-S test) [50] is used when comparing the empirical distribution of a quantitative variable with a particular theoretical

distribution (e.g. the normal distribution). The K-S test can be used as well when comparing two samples and their empirical distributions (not necessarily lying under certain model distributions) [51, 39].

The tested null hypothesis is that the cumulative distribution functions are the same for both samples $A = \{a_i | i = 1, \ldots, n_A; a_i \in \mathbb{R}\}$ and $B = \{b_i | i = 1, \ldots, n_B; b_i \in \mathbb{R}\}$. The empirical cumulative distribution for sample A is defined as $F_A(x) = \frac{\#\{a \in A | a \leq x\}}{n_A}$ and in the same manner for sample B. In our case, A can stand for a sample of real data and B for a sample of synthetic data. The test statistic for the two samples is based on the largest distance (see Fig. 2) between the two empirical distribution functions, which is

$$KS(A, B) = \sup_{x \in \mathbb{R}} |F_A(x) - F_B(x)| \,. \tag{9}$$

The value of the Kolmogorov-Smirnov statistic is then compared to the critical value and the null hypothesis is rejected if $KS(A, B) > \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \kappa_\alpha$, where κ_α is chosen according to the level of significance α. Note that the Kolmogorov-Smirnov test is a non-parametric technique, which means it has no assumption put on the given data.

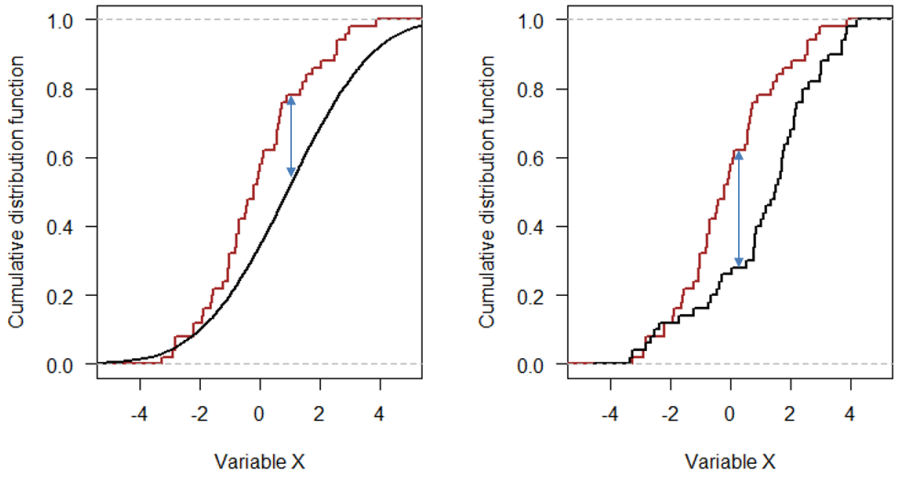

Figure 2: The distance between two distribution functions for an equation of Kolmogorov-Smirnov test. In the one-sample case (left) the distance is measured between the specific theoretical cumulative distribution (black) and the empirical distribution (red). For the two-sample case (right) the distance is computed between the two empirical distribution functions.

4.1.2. Kullback-Leibler divergence

The Kullback-Leibler (K-L) divergence, sometimes called *relative entropy* [52], belongs to the family of probability measures. It is a measure of the dissim-

ilarity between two random quantities, in particular between two probability measures, as the Kolmogorov-Smirnov distance [53].

The K-L divergence measures the inefficiency of assuming that a given distribution is $P_B(x)$ when the true distribution is $P_A(x)$. It is a kind of Bregman divergence and it is defined for discrete samples as:

$$D_{KL}(A \parallel B) = \sum_{x \in X} P_A(x) \log \frac{P_A(x)}{P_B(x)}, \tag{10}$$

where $P_A(x)$ and $P_B(x)$ are two probability distributions [54].

Compared with statistical distances, the statistical divergence is not symmetric, i.e., it is not a metric. The K-L divergence returns the value of divergence of two probability distributions. It reveals 0 if and only if $P_A(x) = P_B(x)$ otherwise it is non-negative.

This measure was applied and visualized using boxplots for comparison of synthetic and real data of microscopy nuclei images in [4].

4.1.3. Kernel maximum mean discrepancy

Maximum mean discrepancy (MMD) [55] is another statistic measuring the difference between two probability distributions on the basis of samples drawn from each of them. More specifically, the aim is to find smooth functions resembling the sample values.

The test statistic is the difference between the mean function values in terms of kernel functions:

$$MMD^2(\mathbb{P}_r, \mathbb{P}_q) = \mathbb{E}_{\substack{x_r, x_r' \sim \mathbb{P}_r \\ x_q, x_q' \sim \mathbb{P}_q}} \left[k(x_r, x_r') - 2k(x_r, x_q) + k(x_q, x_q') \right] \tag{11}$$

where $\mathbb{P}_r$ and $\mathbb{P}_q$ are the probability distributions for some fixed kernel function k, where r and r' are independent variables with distribution $\mathbb{P}_r$; q and q' are independent variables with distribution $\mathbb{P}_q$. The biased empirical estimate of MMD is given by substitution of empirical estimates for expected values to:

$$MMD[R, S] = \left[\frac{1}{m^2} \sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(y_i, y_j) \right]^{\frac{1}{2}} \tag{12}$$

where R and S ($|R| = m, |S| = n$) are two samples to be compared.

Large differences indicate that samples are from different distributions, whereas small MMD suggests that the distributions are identical. In their empirical study, Xu et al. [15] showed that MMD satisfies most of the desirable properties and provided that the distances between samples are computed in

a suitable feature space. To reach this conclusion, they designed a score that includes dropping and collapsing modes and is able to detect overfitting.

MMD was applied for quality assessment of generated data for example in [56] and [57].

4.1.4. Mutual information

Mutual information (MI) is related to joint entropy as introduced simultaneously by Viola [58] and Maes [59].

Given the images I and J, the joint entropy is defined as

$$H(I,J) = -\sum_{a,b} p_{IJ}(a,b) \log p_{IJ}(a,b) \tag{13}$$

where p_{IJ} is the joint probability distribution of pixel intensities associated with images I and J. The joint entropy is minimized when the images I and J are similar. The individual (marginal) entropy for I is

$$H(I) = -\sum_{a} p_I(a) \log p_I(a) \tag{14}$$

and for J in the same manner. MI is then given by

$$MI = H(I) + H(J) - H(I,J) \ . \tag{15}$$

In [30], mutual information was utilized to measure the similarity between original MR image and synthesized CT image resulting from a CycleGAN [60], when the paired reference image was not available.

4.1.5. Regional mutual information

Regional mutual information (RMI) was introduced by [61]. Compared with mutual information, the RMI takes into account information from the neighborhood of each pixel.

RMI is computed for two images I and J (each with corresponding pixels $[I_{ij}, J_{ij}]$) in the following manner:

1. For each of the pixels $[i,j]$, a vector $\boldsymbol{v_{ij}}$ is created with the values coming from the two co-occurence matrices [38] from a neighborhood of radius r. The vector $\boldsymbol{v_{ij}}$ is then considered as a point p_k in a d-dimensional space where $d = (2r+1)^2$. Given the radius r and $m \times n$ images, there is a distribution of $N = (m-2r)(n-2r)$ points represented by a $d \times N$ matrix $P = [p_1, ..., p_N]$.
2. The points are consequently centered by the mean: $P_0 = P - \frac{1}{N}\sum_k^N p_k$.
3. The covariance of the points is then computed as $C = \frac{1}{N} P_0 P_0^T$.

4. The joint entropy $H_g(C)$ is estimated.
5. The marginal entropies $H_g(C_I)$ and $H_g(C_J)$ are estimated where C_I is a matrix in the top left of C and C_J in the bottom right, both of size $\frac{d}{2} \times \frac{d}{2}$.
6. Finally, the RMI is calculated as

$$RMI = H_g(C_I) + H_g(C_J) + H_g(C) \ . \tag{16}$$

As it is also mentioned in [61], the corresponding pixels in the edges of the image can be handled in a number of ways.

In [4], the authors used RMI pair-wise, not only for the real and synthetic counterparts, but also for each pair of real-synthetic images, and also for pairs of real images to express the measure of "alikeness". For the vizualization of the results, the boxplots were used for RMI values between real images next to boxplot of RMI between real and synthetic images.

4.1.6. Inception score

The inception score (IS) [16] is a measure for objective evaluation of trained GANs which was designed to correlate very well with subjective human judgement.

A deep convolutional network model (Google Inception v3 network [62]), pre-trained on the large scale ImageNet dataset [63], is used to classify the generated images from the proposed trained model G. The probabilities of images belonging to particular classes are computed and used for the evaluation of IS. The IS for a trained model G is computed as:

$$IS(G) = \exp\left(\mathbb{E}_{X \sim p_g} KL(p(y|X) \parallel p(y))\right) \tag{17}$$

where $X \sim p_g$ means that an image X is sampled from p_g, $p(y|X)$ is the conditional label distribution (class label y conditional on image X), and $p(y) = \int_X p(y|X)p_g(X)$ is the marginal class distribution. KL is the K-L divergence (as defined in 4.1.2), $\mathbb{E}$ is an expected value, which is further exponentiated for an easier comparison.

Two main properties are covered in IS - image quality (similarity of an image to a specific object) and image diversity (whether the dataset contains a wide range of generated objects). The distribution should have a low entropy for meaningful objects. Also, the marginal integral $\int p(y|X = G(z))dz$ should have high entropy to provide varied images by the generator. The lowest value of IS is 1, the highest value is the number of classes defined in the pre-trained model. The calculation of IS assumes a large enough number of samples (i.e. 50000 images) to be able measure the diversity.

As a validation method, IS was applied in [64] for two proposed models designed to generate multi-parameter MRI data. IS was also evaluated in [2] to show the quality of synthesized MRI brain slices using GAN.

4.1.7. Fréchet distance and Fréchet inception distance

The Fréchet distance [65] is a measure of distance between curves that takes into account the location and ordering of the points along the curves. It is well known for its simile to walking a dog on a leash, who is crossing the curved path. A discrete variant of this measure, which can be evaluated over any feature vectors, exists [66]. However, the most popular application in the field of image synthesis is the use of Inception feature vectors from Inception v3 model (called as the Fréchet inception distance (FID)), performed for example in [56, 67]. It was applied also on SonoNet in [68] as Fréchet SonoNet distance.

FID summarizes the distance between the Inception feature vectors for real and generated images in the same domain. More specifically, the pre-trained classification model is applied to both the real and generated images and the Inception feature vectors from the hidden layer are extracted, thus the final classification probabilities are not used. It was proposed in [69] as an improvement over the existing IS (described in section 4.1.6), which also uses the Inception feature vectors v3 for evaluating the quality of generated images by GAN models. They showed in the paper that FID is more consistent with the noise level than IS. FID is a distance while IS is a score with a maximum value.

The definition of FID is given by:

$$d^2 = ||\mu_1 - \mu_2||^2 + Tr(C_1 + C_2 - 2 * \sqrt{C_1 * C_2}), \tag{18}$$

where μ_1 and μ_2 refer to the feature-wise mean of the real and generated images, where each element is the mean feature observed across the images. C_1 and C_2 are the covariance matrices for the real and generated feature vectors. Tr is a trace – a linear algebra operation – the sum of the elements on a diagonal of the square matrix.

The best value of FID is 0.0 indicating that the two groups of images are identical, thus low FID values mean that the two groups of images are similar (or they have similar statistics).

4.2 Measures based on clustering and classification

4.2.1. Affinity propagation

Affinity propagation (AP) [70] is a method of clustering data. The data points create a network of nodes, where each data point is a potential exemplar. The data points given to the input of this procedure are iteratively examined according to the measure of similarity (e.g. squared error in Euclidean distance) until a good set of exemplars and corresponding clusters emerges.

Initially, the input matrix s is computed over all combinations and each "similarity" is set to some optimization criterion, for example a negative squared

error:

$$s(i,k) = -\|x_i - x_k\|^2 \tag{19}$$

for points x_i and x_k. The values of elements belonging to the diagonal ($s(i,i)$) are able to prefer some of the points to be an exemplar.

Two types of "messages" are evaluated in each step for two points: the "responsibility" message $r(i,k)$ sent from point i to point k and the "availability" $a(i,k)$ sent from point k to point i. The first message, "responsibility":

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\} \tag{20}$$

reflects the actual evidence for how well-suited point k is to serve as the exemplar for point i (regarding the potential exemplars of point i). The second message, "availability"

$$a(i,k) \leftarrow \min\left\{0, r(k,k) + \sum_{i' \notin \{i,k\}} \max\{0, r(i',k)\}\right\} \tag{21}$$

reflects the accumulated evidence for how appropriate it would be for point i to choose point k as its exemplar (regarding the support from other points that point k should be an exemplar). These are initially set to zeros.

The "self-availability" $a(k,k)$ is updated in each step as

$$a(k,k) \leftarrow \sum_{i' \neq k} \max\{0, r(i',k)\} \ . \tag{22}$$

The messages are sent until either the cluster boundaries remain unchanged over a number of iterations, or after some predetermined number of iterations. The advantages of AP are that no specific number of clusters has to be predefined.

In [7], affinity propagation was applied to 13 Haralick texture features of 20 images of real nuclei to get different phenotypes of images. These were compared to the phenotypes clustered on synthetic images. The frequencies of phenotypes in both groups were compared using histograms.

4.2.2. Classification

We should keep in mind that the aim of validation methods is not to find the differences between the synthetic and real data, but to explore the similarity of the synthetic data compared with the real data. One could suggest that a classification task distinguishing real and synthetic data is a suitable solution. However, the bad result of any classifier does not prove that the data are not similar. For example in Fig. 3, it is not possible to distinguish the two classes of points (blue and yellow) with any discrimination function, but at first sight

one can see that the groups are not homogeneous. Furthermore, the result of a classification could be confused with a bad choice of classifier.

Some works employed classification not for differentiation of real versus synthetic dataset, but for verification that the synthetic images will be classified to the appropriate subcategories as well as real data. For instance, in [43] the synthetic images for ten classes of organelles were generated. The support vector machine classification was applied to both real and synthetic data, and results (such as confusion matrix) were reported for comparison. Similarly, in [71] the k-nearest neighbor to classify bacteria classes based on their shape was performed.

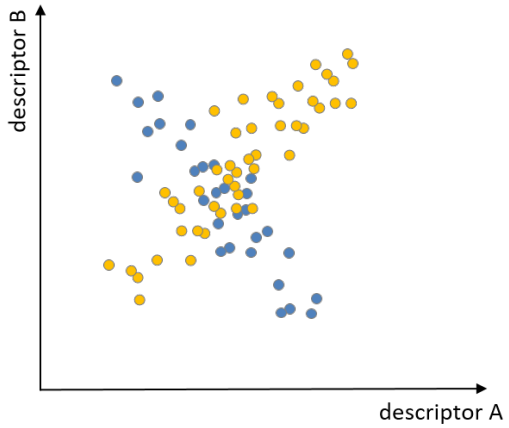

Figure 3: A classification problem. It is not possible to distinguish the two groups of points (blue and yellow) with any discriminator without transformation. However, at first sight, one can see that the groups are not homogeneous.

4.3 Measures based on a visualization of transformed data

4.3.1. Histogram

A histogram is a widely used plot for visualization of absolute or relative frequencies of a categorized quantitative variable. The quantitative variable (e.g., an image descriptor) determines the axis x with a specific binning (categorization), and the axis y counts the frequency – relative (f) or absolute ($f*$) – in the bins:

$$f(j) = \frac{n_j/n}{d_j}, f^*(j) = \frac{n_j}{d_j} \ , \tag{23}$$

where n is the total number of observations, n_j is the number of observations for a particular category j, and d_j is the width of this category. The shape of

the histogram should approximate the probability density function. However, the approximation is very sensitive to the choice of the number of bins.

A quality assessment of the synthesized images at the image descriptor level was used in [39], where the distributions of five Haralick texture descriptors [38] were compared using histograms and the Kolmogorov-Smirnov test. In [7], histograms were used to compare the length of minor axis and ratio between minor and major axes of healthy crypts in colon tissue in real and synthetic images. The Gamma distribution of these two measures was fitted to the real data and subsequently plotted into both histograms.

4.3.2. Quantile-quantile plot

The quantile-quantile (Q-Q) plot [72] is designed for the visualization of distribution comparisons. The quantiles of two quantitative variables are plotted in the x-y figure against each other (see Fig. 4). The quantile matches (and therefore identical distribution) reveals if the points lie along the straight line with the slope of 1. Distant points from this line indicate deviation from the same distribution. The method is non-parametric with no assumption put on the input data. In case of image descriptors, the Q-Q plot can help to compare the distribution of both groups of data in each descriptor separately, i.e. univariately [51, 1, 73, 43].

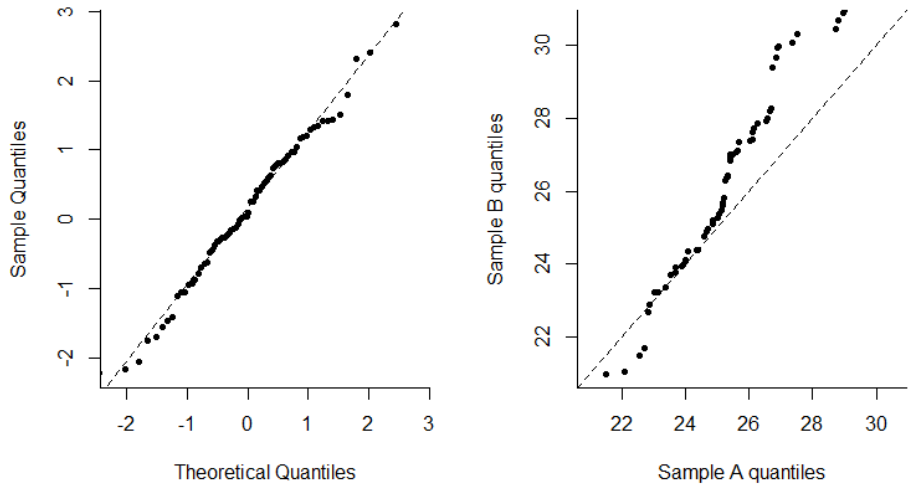

Figure 4: Example of Q-Q plot in case of one-sample comparison (left) and two-sample comparison (right). The optimal match of distributions is achieved when the points lie on the dashed line.

4.3.3. Fourier ring correlation

The Fourier ring correlation (FRC) [74] is a spatial frequency correlation function that measures the degree of correlation of two images at different spatial frequencies. The images are initially transformed to the frequency domain. Afterwards, the normalized average correlation is computed for N_r concentric rings of increasing radius, which corresponds to increasing spatial frequencies centered around the (0,0) spatial frequency.

FRC for images I and J (which are transformed to the frequency domain) is computed as:

$$FRC(R) = \frac{\sum_{i \in R} I(\mathbf{r}_i) \cdot J(\mathbf{r}_i)^*}{\sqrt{(\sum_{i \in R} |I(\mathbf{r}_i)|^2)(\sum_{i \in R} |J(\mathbf{r}_i)|^2)}} \tag{24}$$

where R is the ring number and $\mathbf{r}_i$ are the spatial coordinates. The values of the FRC draw a curve in a plot, which can be further investigated (see Fig. 5).

This approach is found to be useful in electron and fluorescence microscopy [75, 76] for determining the resolution at which both images are consistent. Here, the 2σ criterion is specified as

$$F_{2\sigma}(R) = \frac{2}{\sqrt{N_p(R)/2}} \tag{25}$$

where $N_p(R)$ is the number of pixels in the ring R. The optimal resolution is then found as the crossing of FRC and $F_{2\sigma}$ curves.

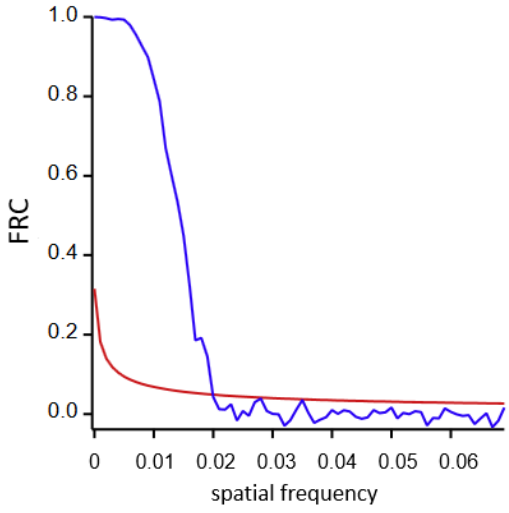

Figure 5: Fourier ring correlation. The values of FRC for each ring are plotted as a curve (blue). The red curve depicts the $F_{2\sigma}$. Adopted from [76].

4.3.4. Overlapping subspaces

In [77], the validation method is based on the comparison of explained variability of both real and synthetic data in the same feature space. The descriptors (Haralick descriptors in this case) are initially preprocessed by principal component analysis to reduce the original number of dimensions into only three, easy-to-visualize, dimensions. The real and synthetic images are represented as data points in this feature subspace. Finally, the overlap of the clusters created around real and synthetic data (Fig. 6) is evaluated via Jaccard index as a quantitative measure of this technique. However this validation method assumes that the three principal components are able to explain the majority of the original feature space given by the descriptors.

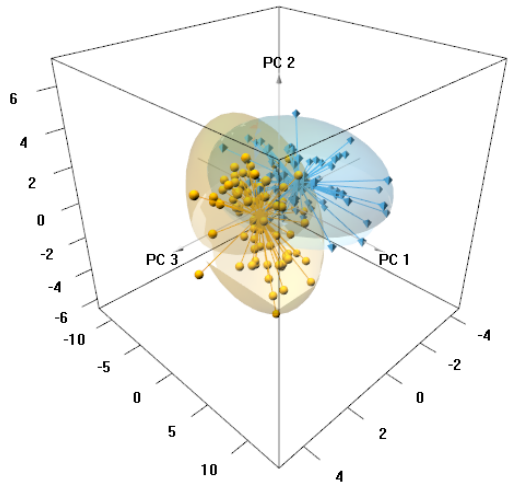

Figure 6: Interactive visualizer showing real images (yellow) and synthetic images (blue) in the reduced feature space. The axes are the three main components given by principal component analysis. Each point corresponds to a particular image [77].

4.3.5. t-distributed stochastic neighbor embedding

t-distributed stochastic neighbor embedding (t-SNE) [78] is also a reduction (nonlinear in this case) from a multidimensional feature space into two or three dimensions designed for visualisation. The method is an improved variation of stochastic neighbor embedding (SNE) [79] with easier optimization and better results in terms of spread of the points in the 2D or 3D map.

The images are represented as high-dimensional data points in a feature space. Like SNE, the Euclidean distances between data points are converted into conditional probabilities representing similarities between the images. The

similarity of data point x_j to data point x_i in SNE is expressed as:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2)} \tag{26}$$

where σ_i is the variance of the Gaussian that is centered on data point x_i. The probability corresponds to the probability that x_j would be in a neighborhood of x_i in proportion to their probability density under the Gaussian centered at x_i. The values of $p_{i|i}$ are set to zero. The low-dimensional counterparts for data points x_i, x_j in a high-dimensional space are labeled as y_i, y_j. The conditional probability $q_{j|i}$ for the low-dimensional counterparts is evaluated in a similar way as in equation 26, but the variance of the Gaussian is set to $\frac{1}{\sqrt{2}}$. With the use of symmetric SNE, the joint probabilities are employed instead of conditional probabilities for a faster computation. They are set to $p_{ij} = \frac{p_{j|i}+p_{i|j}}{2n}$.

Compared with SNE, t-SNE employs a heavy-tailed distribution – a Student t-distribution with one degree of freedom – instead of Gaussian in the low-dimensional map. The final joint probabilities q_{ij} are defined as

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}} \ . \tag{27}$$

The particular locations of the data points in the map are determined by minimizing the K-L divergence over all data points using gradient descent.

A validation of synthetic data using t-SNE was applied as a qualitative method in amyloid brain PET image synthesis by [56]. Qualitative visual assessment of the real and synthetic distribution in the 2D plot via t-SNE was also applied in [10]. They compared two settings of training GAN and real images in the map. t-SNE was used in [80] to show the homogeneity of generated and real images with (or without) lesions. In this work, the aim was to generate normal-looking counterpart for the abnormal images with lesions. Similarly, in [81] t-SNE of the original and augmented cervical histopathology was a part of the qualitative validation of the training set. In [82], the authors used t-SNE in a different way – for a comparison among generated methods, not real images.

4.4 Measures applied to time-lapse sequences

4.4.1. Linear mixed models

Linear mixed models (LMM) also known as *hierarchical linear models* or *linear models with mixed effects* are an extension of linear models for dependent measurements. The extension is needed to overcome two important assumptions put on basic linear models, which are violated in data with dependent

measurements. One of them is that the residuals have to be of homogeneous variance and the other that the residuals should not be correlated. The dependency can result from longitudinal data, repeated measurements or clustered data, because one observation is related to another. This method is one of the possibilities for statistical comparison of time-lapse sequenced image data as well as other dependent datasets of images.

The linear mixed model [83] is considered to be of formula:

$$\boldsymbol{Y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\varepsilon_i} \ , \tag{28}$$

where $\boldsymbol{Y_i} = (Y_{i1}, ..., Y_{in_i})'$ is the vector of repeated quantitative outcome measurements for object i, $\boldsymbol{\beta}$ is treated as a vector of fixed effects, i.e. population-average regression coefficients, and $\boldsymbol{X_i}$ is a known design matrix linking $\boldsymbol{\beta}$ to $\boldsymbol{Y_i}$. The effect of $\boldsymbol{\beta}$ is the same for all the objects. In the case of validation of synthetic data, we are interested in the statistical significance of β_j, which is one element of vector $\boldsymbol{\beta}$. A coefficient β_j is the estimated effect of $\boldsymbol{X_j}$, a binary variable holding the information about being a synthetic or a real image. Naturally, β_j can be the only element of $\boldsymbol{\beta}$ if $\boldsymbol{X_j}$ is the only explanatory variable in the design matrix $\boldsymbol{X_i}$. Random effects are held by $\boldsymbol{b_i}$, a vector of q object-specific regression coefficients. The columns in the $\boldsymbol{Z_i}$ matrix represent observed values for the q predictor variables for the i-th subject, which have effects on the continuous response variable that vary randomly across subjects. In many cases, predictors with effects that vary randomly across subjects are represented in both the $\boldsymbol{X_i}$ matrix and the $\boldsymbol{Z_i}$ matrix.

The residuals $\boldsymbol{\varepsilon_i}$ are distributed as $\boldsymbol{\varepsilon_i} \sim N(\boldsymbol{0}, \boldsymbol{R_i})$, where $\boldsymbol{R_i}$ is a covariance matrix and depends on i only through its dimension n_i. The $\boldsymbol{b_i}$ are distributed as $\boldsymbol{b_i} \sim N(\boldsymbol{0}, \boldsymbol{D})$, independently of each other and of the $\boldsymbol{\varepsilon_i}$, where $\boldsymbol{D}$ is a covariance matrix of the random effects [83, 84, 85].

LMM were applied in [86] to images of tubular network of endothelial cells. For assessment of the effect of being an image from a group of real or synthetic dataset, two standard measures were chosen: box counting fractal dimension and lacunarity [87, 88], both describing the complexity of the structures depicted in the analyzed images.

4.4.2. Dynamic time-warping

Dynamic time warping (DTW) is an algorithm for measuring similarity between two time series which may vary in timing. In another words, one can differ from another only in being slower or faster in any part of the series ("warped"). The aim of the algorithm is to find an optimal alignment. The algorithm was originally applied for comparing speech patterns in automatic speech recognition [89].

In [90], the DTW for two time series $X := (x_1, x_2, ...x_N)$ of length $N \in \mathbb{N}$ and $Y := (y_1, y_2, ...y_M)$ of length $M \in \mathbb{N}$ is defined using a *local cost measure*.

The local cost measure is a function

$$c : F \times F \to \mathbb{R}_{\geq 0} \ , \tag{29}$$

comparing two features x, y from a fixed feature space F. Local cost measures for each pair of elements X and Y are composed into *cost matrix* $C \in \mathbb{R}^{N \times M}$ defined by

$$C(n, m) := c(x_n, y_m) \ . \tag{30}$$

The goal is to find an alignment between X and Y with the minimal overall cost. A warping path is a sequence $p = (p_1, ...p_L)$ with $p_l = (n_l, m_l) \in [1 : N] \times [1 : M]$ for $l \in [1 : L]$ that corresponds to the possibility to align X and Y with a particular total cost $c_p(X, Y)$ with respect to the local cost measure c:

$$c_p(X, Y) := \sum_{l=1}^{L} c(x_{n_l}, y_{m_l}) \ . \tag{31}$$

Finally, the optimal warping path is the one having minimal total cost among all possible warping paths. This task is a standard optimization problem which can be solved, e.g., by using dynamic programming.

DTW as a validation method was applied in [91]. On a dataset of tubular networks of epithelial cells, the descriptor computing the number of lagoons in the network was evaluated for each frame of the sequence. The curves resulted from real and synthetic image sequences showing the development of the networks were compared against each other.

5. Conclusion

A wide range of methods for the validation of generated data were described in this chapter. Even though the mathematical models offer a powerful tool in the inspection of synthetic data, they should be accompanied by a human-driven qualitative assessment. However, the human eye can be deceived by many factors, e.g. the level of details and number of dimensions, thus it is not sufficient as such. The qualitative and quantitative assessment should go hand-in-hand together and the method should be carefully chosen with regard to suitability for a particular application.

Acknowledgments

We thank Yang Song from University of New South Wales for her invaluable feedback to this chapter.

N. Burgos received funding from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" programme reference ANR-10-IAIHU-0006 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6).

References

[1] Malm P, Brun A, Bengtsson E (2015) Simulation of bright-field microscopy images depicting Pap-smear specimen. Cytometry, Part A 87:212–226

[2] Calimeri F, Marzullo A, Stamile C, Terracina G (2017) Biomedical data augmentation using generative adversarial neural networks. In: Lintas A, Rovetta S, Verschure PF, Villa AE (eds) Artificial Neural Networks and Machine Learning – ICANN 2017, Springer International Publishing, Cham, pp 626–634

[3] Apou G, Feuerhake F, Forestier G, Naegel B, Wemmert C (2015) Synthesizing whole slide images. In: 2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA), pp 154–159

[4] Glotsos D, Kostopoulos S, Ravazoula P, Cavouras D (2018) Image quilting and wavelet fusion for creation of synthetic microscopy nuclei images. Computer Methods and Programs in Biomedicine 162:177–186, DOI "10.1016/j.cmpb.2018.05.023"

[5] Vitale S, Orlando JI, Iarussi E, Larrabide I (2020) Improving realism in patient-specific abdominal ultrasound simulation using cyclegans. Int J Comput Assist Radiol Surg 15(2):183–192

[6] Gong E, Pauly JM, Wintermark M, Zaharchuk G (2018) Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. Journal of Magnetic Resonance Imaging 48(2):330–340, DOI 10.1002/jmri.25970

[7] Kovacheva VN, Snead D, Rajpoot NM (2016) A model of the spatial tumour heterogeneity in colorectal adenocarcinoma tissue. BMC bioinformatics 17(1):255

[8] Geman D, Geman S, Hallonquist N, Younes L (2015) Visual turing test for computer vision systems. Proceedings of the National Academy of Sciences 112(12):3618–3623, DOI 10.1073/pnas.1422953112

[9] Han C, Hayashi H, Rundo L, Araki R, Shimoda W, Muramatsu S, Furukawa Y, Mauri G, Nakayama H (2018) Gan-based synthetic brain mr image generation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp 734–738, DOI 10.1109/ISBI.2018.8363678

[10] Han C, Kitamura Y, Kudo A, Ichinose A, Rundo L, Furukawa Y, Umemoto K, Li Y, Nakayama H (2019) Synthesizing diverse lung nodules wherever massively: 3d multi-conditional gan-based ct image augmentation for object detection. 1906.04962

[11] Chuquicusma MJM, Hussein S, Burt J, Bagci U (2018) How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. 1710.09762

[12] Schlegl T, Seeböck P, Waldstein SM, Langs G, Schmidt-Erfurth U (2019) F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Medical Image Analysis 54:30–44, DOI 10.1016/j.media.2019.01.010

[13] Lehmussola A, Ruusuvuori P, Selinummi J, Huttunen H, Yli-Harja O (2007) Computational framework for simulating fluorescence microscope images with cell populations. IEEE Transactions on Medical Imaging 26(7):1010–1016, DOI 10.1109/tmi.2007.896925

[14] Ouyang J, Chen TK, Gong E, Pauly J, Zaharchuk G (2019) Ultra-low-dose pet reconstruction using generative adversarial network with feature matching and task-specific perceptual loss. Medical Physics pp 3555–3564

[15] Xu Q, Huang G, Yuan Y, Guo C, Sun Y, Wu F, Weinberger K (2018) An empirical study on evaluation metrics of generative adversarial networks. 1806.07755

[16] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X, Chen X (2016) Improved techniques for training gans. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 29, pp 2234–2242

[17] Lopez-Paz D, Oquab M (2018) Revisiting classifier two-sample tests. 1610.06545

[18] Burgos N, Cardoso MJ, Thielemans K, Modat M, Pedemonte S, Dickson J, Barnes A, Ahmed R, Mahoney CJ, Schott JM, Duncan JS, Atkinson D, Arridge SR, Hutton BF, Ourselin S (2014) Attenuation correction synthesis for hybrid PET-MR scanners: Application to brain studies. IEEE Transactions on Medical Imaging 33(12):2332–2341, DOI 10.1109/TMI.2014.2340135

[19] Dowling JA, Sun J, Pichler P, Rivest-Hénault D, Ghose S, Richardson H, Wratten C, Martin J, Arm J, Best L, Chandra SS, Fripp J, Menk FW, Greer PB (2015) Automatic Substitute Computed Tomography Generation and Contouring for Magnetic Resonance Imaging (MRI)-Alone External Beam Radiation Therapy From Standard MRI Sequences. International Journal of Radiation Oncology · Biology · Physics 93(5):1144–1153, DOI 10.1016/j.ijrobp.2015.08.045

[20] Han X (2017) MR-based synthetic CT generation using a deep convolutional neural network method. Medical Physics 44(4):1408–1419, DOI 10.1002/mp.12155

[21] Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Isgum I (2017) Deep MR to CT synthesis using unpaired data. CoRR abs/1708.01155

[22] Nie D, Trullo R, Lian J, Petitjean C, Ruan S, Wang Q, Shen D (2017) Medical Image Synthesis with Context-Aware Generative Adversarial Networks. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017, Springer, Cham, Lecture Notes in Computer Science, pp 417–425, DOI 10.1007/978-3-319-66179-7_48

[23] Roy S, Carass A, Prince JL (2013) Magnetic Resonance Image Example-Based Contrast Synthesis. IEEE Transactions on Medical Imaging 32(12):2348–2363, DOI 10.1109/TMI.2013.2282126

[24] Jog A, Carass A, Roy S, Pham DL, Prince JL (2017) Random forest regression for magnetic resonance image synthesis. Medical Image Analysis 35:475–488, DOI 10.1016/j.media.2016.08.009

[25] Dar SU, Yurt M, Karacan L, Erdem A, Erdem E, Çukur T (2019) Image Synthesis in Multi-Contrast MRI With Conditional Generative Adversarial Networks. IEEE Transactions on Medical Imaging 38(10):2375–2388, DOI 10.1109/TMI.2019.2901750

[26] Wolterink JM, Leiner T, Viergever MA, Išgum I (2017) Generative Adversarial Networks for Noise Reduction in Low-Dose CT. IEEE Transactions on Medical Imaging 36(12):2536–2545, DOI 10.1109/TMI.2017.2708987

[27] Yang Q, Yan P, Zhang Y, Yu H, Shi Y, Mou X, Kalra MK, Zhang Y, Sun L, Wang G (2018) Low-Dose CT Image Denoising Using a Generative Adversarial Network With Wasserstein Distance and Perceptual Loss. IEEE Transactions on Medical Imaging 37(6):1348–1357, DOI 10.1109/TMI.2018.2827462

[28] Chen Y, Shi F, Christodoulou AG, Xie Y, Zhou Z, Li D (2018) Efficient and Accurate MRI Super-Resolution Using a Generative Adversarial Network and 3D Multi-level Densely Connected Network. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, Cham, Lecture Notes in Computer Science, pp 91–99, DOI 10.1007/978-3-030-00928-1_11

[29] Lee G, Oh J, Kang M, Her N, Kim M, Jeong W (2018) Deephcs: Bright-field to fluorescence microscopy image conversion using deep learning for label-free high-content screening. In: Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference,

Granada, Spain, September 16-20, 2018, Proceedings, Part II, pp 335–343, DOI 10.1007/978-3-030-00934-2_38

[30] Hiasa Y, Otake Y, Takao M, Matsuoka T, Takashima K, Carass A, Prince JL, Sugano N, Sato Y (2018) Cross-modality image synthesis from unpaired data using CycleGAN - effects of gradient consistency loss and training data size. In: Simulation and Synthesis in Medical Imaging - Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings, Springer, LNCS 11037, pp 31–41, DOI 10.1007/978-3-030-00536-8_4

[31] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13(4):600–612, DOI 10.1109/TIP.2003.819861

[32] Hore A, Ziou D (2010) Image quality metrics: Psnr vs. ssim. In: 2010 20th International Conference on Pattern Recognition, pp 2366–2369, DOI 10.1109/ICPR.2010.579

[33] Dosselmann R, Yang XD (2011) A comprehensive assessment of the structural similarity index. Signal Image Video Process 5(1):81–91

[34] Mason A, Rioux J, Clarke SE, Costa A, Schmidt M, Keough V, Huynh T, Beyea S (2020) Comparison of Objective Image Quality Metrics to Expert Radiologists' Scoring of Diagnostic Quality of MR Images. IEEE Transactions on Medical Imaging 39(4):1064–1072, DOI 10.1109/TMI.2019.2930338

[35] Xia T, Chartsias A, Tsaftaris SA (2020) Pseudo-healthy synthesis with pathology disentanglement and adversarial learning. Medical Image Anal 64:101719

[36] Wang Z, Simoncelli EP, Bovik AC (2003) Multi-scale structural similarity for image quality assessment. In: in Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers, pp 1398–1402

[37] Majtner T (2015) Texture-based image description in fluorescence microscopy. Doctoral theses, dissertations, Masaryk University, Faculty of Informatics, Brno

[38] Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics SMC-3(6):610–621

[39] Svoboda D, Ulman V (2017) MitoGen: A Framework for Generating 3D Synthetic Time-Lapse Sequences of Cell Populations in Fluorescence Microscopy. IEEE Transactions on Medical Imaging 36(1):310–321

[40] Svoboda D, Kašík M, Maška M, Hubený J, Stejskal S, Zimmermann M (2007) On simulating 3D fluorescent microscope images. In: CAIP, Springer, Lecture Notes in Computer Science, vol 4673, pp 309–316

[41] Svoboda D, Homola O, Stejskal S (2011) Generation of 3D digital phantoms of colon tissue. In: Proceedings of 8th International Conference on Image Analysis and Recognition, Springer-Verlag, Berlin, Heidelberg, pp 31–39, DOI http://dx.doi.org/10.1007/978-3-642-21596-4_4

[42] Boland MV, Murphy RF (2001) A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. Bioinformatics 17(12):1213–1223, DOI 10.1093/bioinformatics/17.12.1213

[43] Zhao T, Murphy RF (2007) Automated learning of generative models for subcellular location: Building blocks for systems biology. Cytometry Part A 71A(12):978–990, DOI 10.1002/cyto.a.20487

[44] Ojala T, Pietikainen M, Harwood D (1994) Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In: Proceedings of 12th International Conference on Pattern Recognition, vol 1, pp 582–585, DOI 10.1109/ICPR.1994.576366

[45] Singh P, Mukundan R, Ryke RD (2017) Texture based quality analysis of simulated synthetic ultrasound images using Local Binary Patterns. Journal of Imaging 4(1):3, DOI 10.3390/jimaging4010003

[46] Michalet X (2010) Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium. Physical Review E 83(4), DOI 10.1103/physreve.82.041914

[47] Vu CT, Chandler DM (2009) S3: A spectral and spatial sharpness measure. In: 2009 First International Conference on Advances in Multimedia, pp 37–43, DOI 10.1109/MMEDIA.2009.15

[48] Zhao C, Dewey BE, Pham DL, Calabresi PA, Reich DS, Prince JL (2021) Smore: A self-supervised anti-aliasing and super-resolution algorithm for mri using deep learning. IEEE Transactions on Medical Imaging 40(3):805–817, DOI 10.1109/TMI.2020.3037187

[49] Paavolainen L, Kankaanpää P, Ruusuvuori P, McNerney G, Karjalainen M, Marjomäki V (2012) Application independent greedy particle tracking method for 3D fluorescence microscopy image series. In: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp 672–675, DOI 10.1109/ISBI.2012.6235637

[50] Massey FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association 46(253):68–78

[51] Sorokin DV, Peterlík I, Ulman V, Svoboda D, Nečasová T, Morgaenko K, Eiselleová L, Tesařová L, Maška M (2018) FiloGen: A Model-Based Generator of Synthetic 3D Time-Lapse Sequences of Single Motile Cells with Growing and

Branching Filopodia. IEEE Transactions on Medical Imaging 37(12):2630–2641, DOI 10.1109/TMI.2018.2845884

[52] Do MN, Vetterli M (2000) Texture similarity measurement using Kullback-Leibler distance on wavelet subbands. In: Proceedings 2000 International Conference on Image Processing, vol 3, pp 730–733, DOI 10.1109/ICIP.2000.899558

[53] Venturini GAM (2015) Statistical distances and probability metrics for multivariate data, ensembles and probability distributions [online]. Doctoral thesis, Universidad Carlos III de Madrid. Departamento de Estadística

[54] MacKay DJC (2003) Information Theory, Inference and Learning Algorithms. Cambridge University Press

[55] Gretton A, Borgwardt K, Rasch M, Schölkopf B, Smola A (2007) A kernel method for the two-sample-problem. In: Schölkopf B, Platt J, Hoffman T (eds) Advances in Neural Information Processing Systems, MIT Press, vol 19, pp 513–520

[56] Kang H, Park JS, Cho K, Kang DY (2020) Visual and quantitative evaluation of amyloid brain pet image synthesis with generative adversarial network. Applied Sciences 10(7), DOI 10.3390/app10072628

[57] Kwon G, Han C, shik Kim D (2019) Generation of 3d brain mri using auto-encoding generative adversarial networks. 1908.02498

[58] Viola P, Wells III WM (1997) Alignment by maximization of mutual information. International Journal of Computer Vision 24(2):137–154, DOI 10.1023/A:1007958904918

[59] Maes F, Collignon A, Vandermeulen D, Marchal G, Suetens P (1997) Multimodality image registration by maximization of mutual information. IEEE Transactions on Medical Imaging 16:187–198

[60] Zhu J, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp 2242–2251, DOI 10.1109/ICCV.2017.244

[61] Russakoff DB, Tomasi C, Rohlfing T, Maurer CR (2004) Image similarity using Mutual Information of Regions. In: Computer Vision - ECCV 2004, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 596–607

[62] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015) Rethinking the inception architecture for computer vision. 1512.00567

[63] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255, DOI 10.1109/CVPR.2009.5206848

[64] Wang Z, Lin Y, Cheng KT, Yang X (2020) Semi-supervised mp-mri data synthesis with stitchlayer and auxiliary distance maximization. Medical Image Analysis 59:101565, DOI 10.1016/j.media.2019.101565

[65] Fréchet MM (1906) Sur quelques points du calcul fonctionnel. DOI 10.1007/bf03018603

[66] Eiter T, Mannila H (1994) Computing discrete frechet distance. Tech. rep., Technische Universit'at Wien

[67] Xu L, Zeng X, Zhang H, Li W, Lei J, Huang Z (2020) Bpgan: Bidirectional ct-to-mri prediction using multi-generative multi-adversarial nets with spectral normalization and localization. Neural Networks 128:82–96, DOI 10.1016/j.neunet.2020.05.001

[68] Lee LH, Noble JA (2020) Generating controllable ultrasound images of the fetal head. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp 1761–1764, DOI 10.1109/ISBI45749.2020.9098578

[69] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, NIPS'17, p 6629–6640

[70] Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315(5814):972–976, DOI 10.1126/science.1136800

[71] Lehmussola A, Ruusuvuori P, Selinummi J, Rajala T, Yli-Harja O (2008) Synthetic images of high-throughput microscopy for validation of image analysis methods. Proceedings of the IEEE 96(8):1348–1360

[72] Wilk MB, Gnanadesikan R (1968) Probability plotting methods for the analysis for the analysis of data. Biometrika 55(1):1–17

[73] Svoboda D, Kozubek M, Stejskal S (2009) Generation of digital phantoms of cell nuclei and simulation of image formation in 3D image cytometry. Cytometry Part A 75A(6):494–509, DOI 10.1002/cyto.a.20714

[74] Nieuwenhuizen RPJ, Lidke KA, Bates M, Puig DL, Grünwald D, Stallinga S, Rieger B (2013) Measuring image resolution in optical nanoscopy. Nature Methods 10(6):557–562, DOI 10.1038/nmeth.2448

[75] Venkataramani V, Herrmannsdörfer F, Heilemann M, Kuner T (2016) SuReSim: simulating localization microscopy experiments from ground truth models. Nature Methods 13(4):319–321, DOI 10.1038/nmeth.3775

[76] Banterle N, Bui KH, Lemke EA, Beck M (2013) Fourier ring correlation as a resolution criterion for super-resolution microscopy. Journal of Structural Biology 183(3):363–367, DOI 10.1016/j.jsb.2013.05.004

[77] Nečasová T, Svoboda D (2019) Visual and quantitative comparison of real and simulated biomedical image data. In: Laura Leal-Taixé SR (ed) Computer Vision – ECCV 2018 - Bioimage Computing Workshop, Springer, Munich, Germany, pp 385–394, DOI 10.1007/978-3-030-11024-6

[78] van der Maaten L, Hinton G (2008) Visualizing high-dimensional data using t-sne. Journal of Machine Learning Research 9(nov):2579–2605, pagination: 27

[79] Hinton G, Roweis S (2002) Stochastic neighbor embedding. In: Proceedings of the 15th International Conference on Neural Information Processing Systems, MIT Press, Cambridge, MA, USA, NIPS'02, p 857–864

[80] Sun L, Wang J, Huang Y, Ding X, Greenspan H, Paisley J (2020) An adversarial learning approach to medical image synthesis for lesion detection. IEEE Journal of Biomedical and Health Informatics 24(8):2303–2314, DOI 10.1109/JBHI.2020.2964016

[81] Xue Y, Ye J, Zhou Q, Long LR, Antani S, Xue Z, Cornwell C, Zaino R, Cheng KC, Huang X (2021) Selective synthetic augmentation with histogan for improved histopathology image classification. Medical Image Analysis 67:101816, DOI 10.1016/j.media.2020.101816

[82] Diaz-Pinto A, Colomer A, Naranjo V, Morales S, Xu Y, Frangi AF (2019) Retinal image synthesis and semi-supervised learning for glaucoma assessment. IEEE Transactions on Medical Imaging 38(9):2211–2218

[83] West BT (2006) Linear Mixed Models. Chapman and Hall/CRC, DOI 10.1201/9781420010435

[84] Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38(4):963, DOI 10.2307/2529876

[85] Molenberghs G, Verbeke G (2001) A review on linear mixed models for longitudinal data, possibly subject to dropout. Statistical Modelling 1(4):235–269, DOI 10.1177/1471082X0100100402

[86] Svoboda D, Nečasová T, Tesařová L, Šimara P (2018) Tubular network formation process using 3D cellular potts model. In: A G, O G, Oguz I BN (eds) Simulation and Synthesis in Medical Imaging - Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings, Springer, LNCS 11037, pp 90–99, DOI http://dx.doi.org/10.1007/978-3-030-00536-8_10

[87] Gould DJ, Vadakkan TJ, Poché RA, Dickinson ME (2011) Multifractal and lacunarity analysis of microvascular morphology and remodeling. Microcirculation 18(2):136–151, DOI 10.1111/j.1549-8719.2010.00075.x

[88] Smith T, Lange G, Marks W (1996) Fractal methods and results in cellular morphology — dimensions, lacunarity and multifractals. Journal of Neuroscience Methods 69(2):123–136, DOI 10.1016/S0165-0270(96)00080-5

[89] Rabiner L, Juang BH (1993) Fundamentals of Speech Recognition. Prentice-Hall, Inc., USA

[90] Müller M (2007) Information Retrieval for Music and Motion, Springer, Berlin, Heidelberg, chap 4 (Dynamic Time Warping), pp 69–84. DOI 10.1007/978-3-540-74048-3_4

[91] Svoboda D, Nečasová T (2020) Image-based simulations of tubular network formation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp 1608–1612, DOI 10.1109/ISBI45749.2020.9098736