Unsupervised convolutional autoencoders for 4D transperineal ultrasound classification

Frieda van den Noort[a,*] **Claudia Manzini**[b] **Merijn Hofsteenge**[a] **Beril Sirmacek**[c] **Carl H. van der Vaart**[b] **and Cornelis H. Slump**[a]

[a]University of Twente, Technical Medical Centre, Robotics and Mechatronics, Faculty of Electrical Engineering Mathematics and Computer Science, Enschede, The Netherlands

[b]University Medical Centre Utrecht, Department of Obstetrics and Gynecology, Utrecht, The Netherlands

[c]Saxion University of Applied Sciences, School of Creative Technology, Smart Cities Group, Enschede, The Netherlands

Abstract

Purpose: 4D Transperineal ultrasound (TPUS) is used to examine female pelvic floor disorders. Muscle movement, like performing a muscle contraction or a Valsalva maneuver, can be captured on TPUS. Our work investigates the possibility for unsupervised analysis and classification of the TPUS data.

Approach: An unsupervised 3D-convolutional autoencoder is trained to compress TPUS volume frames into a latent feature vector (LFV) of 128 elements. The (co)variance of the features are analyzed and statistical tests are performed to analyze how features contribute in storing contraction and Valsalva information. Further dimensionality reduction is applied (principal component analysis or a 2D-convolutional autoencoder) to the LFVs of the frames of the TPUS movie to compress the data and analyze the interframe movement. Clustering algorithms (K-means clustering and Gaussian mixture models) are applied to this representation of the data to investigate the possibilities of unsupervised classification.

Results: The majority of the features show a significant difference between contraction and Valsalva. The (co)variance of the features from the LFVs was investigated and features most prominent in capturing muscle movement were identified. Furthermore, the first principal component of the frames from a single TPUS movie can be used to identify movement between the frames. The best classification results were obtained after applying principal component analysis and Gaussian mixture models to the LFVs of the TPUS movies, yielding a 91.2% accuracy.

Conclusion: Unsupervised analysis and classification of TPUS data yields relevant information about the type and amount of muscle movement present.

 [DOI: 10.1117/1.JMI.10.1.014004]

Keywords: urogynecology; transperineal ultrasound; convolutional autoencoder; classification; unsupervised learning.

Paper 21276GRRRR received Oct. 14, 2021; accepted for publication Jan. 23, 2023; published online Feb. 11, 2023.

1 Introduction

Pelvic floor disorders are common among the female population.[1] These disorders, which can often be linked back to vaginal delivery, tend to (re)appear after menopause. Around one-third of all women above an age of 40 years report pelvic floor disorders, such as pelvic organ prolapse, urinal, or fecal incontinence.[1] Awareness about risk factors, such as vaginal birth, and incidence of these disorders is limited even among pregnant women.[2] The clinical understanding of these

*Address all correspondence to Frieda van den Noort, f.vandennoort@utwente.nl

2329-4302/2023/$28.00

disorders has grown in recent years mostly by biomechanical analysis of the pelvic floor levator ani muscles (LAMs).[3] However, this type of analysis is often based on models obtained from magnetic resonance imaging (MRI) data of a single woman,[4–6] which limits generalization to the wider female population.

4D Transperineal ultrasound (TPUS) is another imaging modality used to assess the pelvic floor for scientific and diagnostic purposes.[7] In 4D TPUS, the forth dimension is the time dimension; ultrasound volumes are captured in time, with a framerate of ∼2 Hz. It has a few distinct advantages over MRI: it is easy to acquire, cheap and therefore widely available, making it easy to collect large datasets.[7] Furthermore, it is easy to capture pelvic floor motion, which allows for *in vivo* function assessment of the LAM (e.g., strain measurements[8]). There are two types of LAM movement that can be captured using TPUS: the first movement type is muscle contraction and the second movement type is the Valsalva maneuver (forced expiration with closed air ways). The latter maneuver raises the abdominal pressure and will therefore stretch the LAM. Strain measurements can be acquired from both movements, which allows for quantification of LAM functionality.[8] These type of measurements yield a better understanding of pelvic floor disorders and the effect of different treatments, such as surgery or physical therapy. However, strain measurements require 3D segmentation of the data, which is time-consuming and is therefore only used in a research setting.[7–10]

The current focus in TPUS literature is on manual selection of one or a few slices within a few specific frames (often at rest, maximum contraction, and maximum Valsalva), in which relevant distances and areas are measured.[7] Based on these measurements, muscle damage (avulsion) can be diagnosed and, to some extent, quantified.[11–13] Here the state of the muscle is relevant for assessment. The maximum Valsalva frame is used to investigate the amount of pelvic organ decent, which is maximized due to the raise in abdominal pressure.[14] The frame of maximum contraction has better tissue contrast and is therefore suitable to diagnose muscle avulsion.[15] Identifying the correct movement type and frame within a TPUS movie is time-consuming, as is obtaining the most basic area and distance measures. Therefore, this is often skipped in clinical practice.[16] Before we can benefit from the large datasets acquired by TPUS and move toward the use of more complex analysis such as strain measurements, automation of the image analysis is needed.

In the last decade, deep learning has proven to be a powerful tool in the automation of medical image analysis.[17–19] This also holds for the analysis of TPUS, where promising results were reported for segmentation of 2D[16,20] and 3D data.[9,21] Furthermore, most TPUS analysis starts with the selection of a single nonorthogonal slice (slice of minimal hiatal dimensions). The selection of this slice was automated on a single frame,[22,23] which is a significant step forward in automating the current clinical analysis of TPUS. However, the selection of the relevant frames from the TPUS movies is still a manual task.

The aforementioned automation studies all use supervised learning and therefore require labels for training, which are not easily obtained for TPUS. In this work, we will explore unsupervised learning[24] on TPUS data, to benefit from large TPUS datasets without the need for labeling. We will use convolutional autoencoders[25–27] to learn a low-dimensional latent feature vector (LFV) representation[28] of a TPUS frame and apply clustering algorithms[29] to find relevant data clusters. We expect the pelvic floor movement, either contraction or Valsalva, to be a prominent feature of the TPUS movie. Therefore, the goal of this study is to use unsupervised learning to correctly classify TPUS movies, to further automate TPUS analysis procedures.

2 Methods

2.1 *Data*

The data used in this study were collected as part of the Gynecological Imaging using 3D Ultrasound project. Women, who visited a tertiary urogynecological clinic with various pelvic floor disorders, were included in the dataset, from May 2018 till December 2019. The Medical Research Ethics Committee of the UMC Utrecht exempted the project from ethical approval (reference 18/215), because TPUS can be considered part of routine diagnostic procedure and

standard care. All women signed informed consent forms. This resulted in a heterogeneous dataset since patients with different pathologies were included.

A Philips Epiq-7 machine with a X6-1 matrix transducer was used for data acquisition. The volume angle was 90 deg in both azimuthal and elevational direction, postprocessing filters were set off, the volume scan rate was 2 Hz, and the scan depth was 9 cm. The transducer was covered with a 2-cm-thick gel pad, which created more a spacing between the patient and the probe, allowing for the capture of the full LAM within the scanned volume. All scans were acquired with the patient in supine position, and patients have emptied their bladder before acquisition. The patients were asked to contract their LAM and to perform a Valsalva maneuver. The maximum number of volumes that could be captured by the ultrasound system in a single movie was 22 frames. Therefore, both types of muscle movement were recorded separately. The recording ended after a full contraction cycle (from rest to contraction to rest) or at maximum Valsalva (from rest to Valsalva) since Valsalva is performed slower than contraction.

The dataset was unsorted, meaning that there are no labels to indicate whether a movie captured a contraction or Valsalva maneuver. However, the clinical examiner (author C. M.) followed a specific acquisition order. For each patient, the acquisition preferably started with contraction and was followed by Valsalva. However, some women confused Valsalva and contraction or needed more attempts to perform the maneuvers correctly, making this information not completely reliable. Based on the experience of the clinical expert (C. M.) who collected the data, it was estimated that the contraction and Valsalva video were in the expected order of the data acquisition around 90% of the time. These were the "labels" used in this study, to investigate whether or not the contraction and Valsalva movement captured is a prominent data feature that can be detected via unsupervised learning.

2.1.1 *Preprocessing and experimental data flow*

A frame of the TPUS data has a size of $277 \times 352 \times 229$ voxels but was cropped, around the volume center, to $192 \times 256 \times 192$ for training of the 3D-convolutional autoencoder (3D-CAE). The outer part of the data does not contain relevant pelvic floor information, due to the conic shape of the TPUS volume (see Fig. 1). The intensity values of the images were linearly scaled between 0 and 1. TPUS movies of 304 patients, with a variety of pelvic floor disorders, were used in this study. From these movies, the first and last frames were selected, as well as two additional randomly selected frames within the TPUS movie. The resulting dataset caused memory problems on the external server used for the training of a 3D-CAE. Therefore, 790 frames were randomly selected from this dataset to train a 3D-CAE (see Sec. 2.2). During training, a single frame was left out and used as independent validation set to check training progress. For 22 patients, none of the frames were used. These patient scans were kept separate as an independent testing set, not to be used in the training and selecting of the best performing models.

4D TPUS movies having <22 frames were excluded from the next LFV analysis step (see Secs. 2.3 and 2.4) to make the data uniform for the subsequent analysis steps, which resulted in a training set of 345 4D TPUS data of 180 patients. Since the clustering classification is unsupervised and the training set labels are not 100% correct, it was decided to use the training set both for training and validation, to have the largest possible dataset for identifying clusters. The independent testing set for classification consisted of 57 4D TPUS data from the 22 testing set patients not involved in training the 3D-CAE. The clinical expert (C. M.) checked the correct labels for the testing set, after which more than 2 TPUS data could be used for some patients. For the LFV analysis, all 4D TPUS frames were converted to LFVs using the 3D-CAE, resulting in 22×128 representations of a single 4D TPUS movie. The full work and dataflow of this paper is visualized in Fig. 1.

2.2 *3D Convolutional Autoencoder*

Figure 2(a) shows the design of the 3D-CAE used to compress a single TPUS frame into a LFV and to provide a reconstruction of the original frame as output.[30] This design is a trade-off between network depth, which improves the learning of the network, and the available GPU

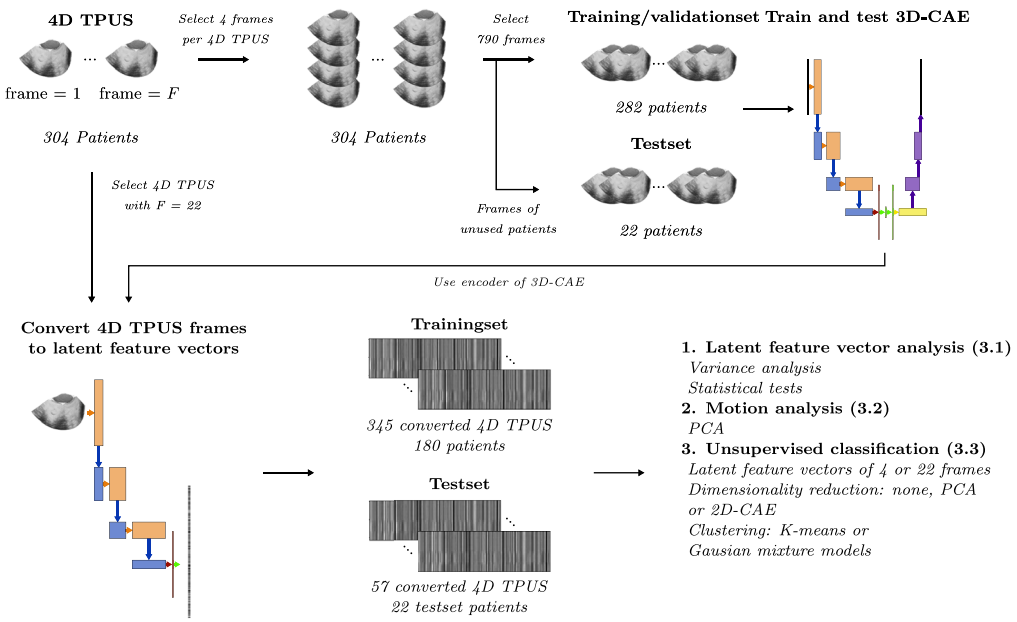

Fig. 1 An overview of the work and data flow presented in this paper. From 304 patients, multiple TPUS movies, with F frames, are available. Frames are selected from these movies to train a unsupervised 3D-convolutional autoencoder (3D-CEA) [see Fig. 2(a) for network details]. In this process, data of 22 patients are not used for network training and left as independent testing set. 4D TPUS movies of 22 frames are frame-by-frame converted to the latent feature (LF) space using the encoder of the 3D-CEA to create uniform LF space representations of 22×128. This LF representation of the 4D data is analyzed further. First, analysis of the variance of the LFs and statistical test are performed to better understand the LF representation learned by the 3D-CAE (Sec. 3.1). Second, PCA is applied to the LF vectors of single movies and analyze the change on the first principal axis to analyze the motion of an individual movie (Sec. 3.2). Finally, unsupervised classification is trained on the LF vectors of the 4D TPUS movies (Sec. 3.3). To find the best performing classification strategy, the LF vectors of either the four frames of maximum contraction/Valsalva or full 22 frames are used. Dimensionality reduction is either not applied or PCA or a 2D-CAE is used to the LF of the frames. For the final identification of data clusters, K-means clustering and GMM are used.

memory. The 3D-CAE compresses a TPUS frame to LFV of 128 elements, which reduces the dimensionality of the data with a factor $\frac{192\times256\times192}{128} = 73{,}728$. A Swish-function, with $\beta = 1$, was used as the activation function.[31] For the output layer, a clipped Rectified linear unit (ReLU) was used,[32] to scale the output between 0 and 1.

2.2.1 *Training*

For network and training, Python 3.7.6, with deep learning framework Tensorflow 2.4.0, was used. The network was trained with the following loss function (L):

$$L = \frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2 + \gamma\left|\sum_{j=1}^{M}(y_j^2) - 1\right|. \tag{1}$$

Here x_i and $\hat{x}_i$ are the i'th input- and output-voxel values, respectively, with N being the total number of voxels in the TPUS frame. The first part of the loss function is the mean squared error loss, to minimize the difference between the network input and output. The second part of the loss limits the M (128) elements of the LFV y [see Fig. 2(a)] from becoming too large, their quadratic sum is forced toward 1. This enforces the use of all elements within the LFV rather then the use of a few prominent features, whereas others are not used in the encoding of the frames. The scaling factor between the two sides of the loss is γ, which was set to be 10^{-3},

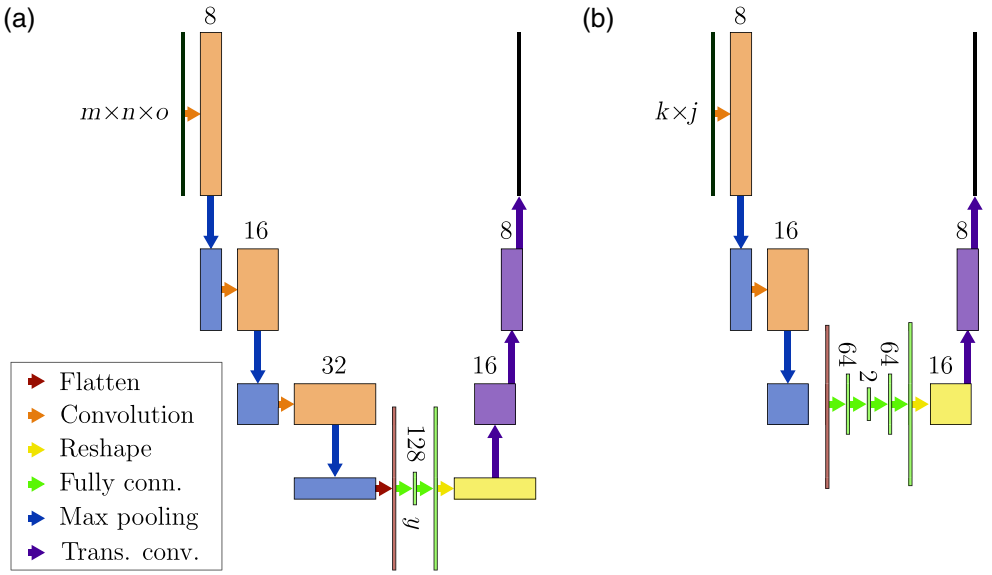

Fig. 2 The design of (a) the 3D-CAE network and (b) the 2D-CAE. The input of the 3D-CAE is an 4D TPUS frame of size $m \times n \times o$ and for the 2D-CAE an LFV representation of 4D TPUS of $k \times j$. The convolution layers apply a convolution using a $3 \times 3(3\times)$ kernel, the filter size is mentioned above the layers. The max-pooling layer reduces the spatial resolution of the data by a factor 2, applying a $2 \times 2(2\times)$ kernel with stride 2. The flatten layer transforms the 4D or 3D data into a single vector, whereas the reshape layers perform the opposite operation. The fully connected operation (Fully conn.) in the center of the network creates the LFV y of 128 elements in the 3D-CAE. In the 2D-CAE, more fully connected operations are applied, the LFV consists of two elements. The data are spatially upsampled in the decoding part of the network by transposed convolution layers (Trans. conv.) that are applied using a $3 \times 3(3\times)$ kernel and a stride of 2. The activation functions in the network are the Swish function and the clipped ReLU function for the output layer. The blocks have the same color as the preceding operation, to visualize the result of this operation. The network output is trained to reconstruct the network input and therefore has the same size.

after both training and validation loss showed that this number provided a balanced trade-off between both parts of the loss function. The Adam optimizer[33] with a learning rate of 3×10^{-4} was used for training. Due to GPU memory limitations, the network was trained with a batch size of 1 TPUS frame. Due to the large training set, the network was only trained for 20 epochs, which was sufficient for convergence and took around 15 h on our setup. The validation loss was calculated after every 30'th training step. The model with the lowest validation reconstruction loss was saved and used for further analysis.

2.2.2 *Evaluation of the latent feature space*

To evaluate the latent feature space learned by the 3D-CAE, the variance for each feature was calculated for the entire data set and for the contraction and Valsalva data separately. The variance provides a measure of the differences present within the dataset for a specific feature. The covariance matrix of the features was calculated as well to analyze the correlation between the features.

To evaluate which features play a significant role in capturing contraction and Valsalva information, some statistical tests were performed: the first test is the Wilcoxon signed rank sum test to test if the features significantly change within the movie. Therefore, the LFVs of the first frame and the frame of maximum contraction or Valsalva were selected and compared. To test for similarities of the feature means and distributions, the Wilcoxon–Mann–Whitney test and T-test were used, comparing the latent features of all contraction frames to all Valsalva frames and only the maximum contraction frames to maximum Valsalva. Since the T-test requires normally distributed data, this was tested using the Shapiro–Wilk test.

2.3 *Dimensionality Reduction*

Using the 3D-CAE, the 4D TPUS data can be compressed to 2D data with a size of 128×22. This is already a significant dimensionality reduction. To further reduce the dimensionality of the data, principal component analysis (PCA)[34] is applied. This can be beneficial for analyzing the principal variance axes within the dataset or within the movies itself.

2.3.1 *Frame difference analysis*

Focusing on the first principal axes of the LFVs of a single TPUS movie, the frame differences in the movie were analyzed. Within a single movie, these differences are mainly caused by muscle movement. To investigate this movement, PCA was applied to the 22 LFVs of the individual TPUS movies from the testing set. Each TPUS frame (frame number n) was represented in this new feature space and has a value on the first principal axis (f_n, the value of the first frame is f_0). Subsequently, each frame was represented by value p_n, where $p_n = |f_n - f_0|$, to represent its distance with respect to the first frame. p_n is to be expected to maximize at maximum contraction or Valsalva. The values of p_n were plotted to analyze the frame-to-frame changes within the TPUS movies.

2.3.2 *Dimensionality reduction before clustering*

Different types of dimensionality reduction methods were used before clustering algorithms are applied, to investigate how they influence the final classification success. To start, it is expected that the difference in the TPUS data between contraction and Valsalva is the largest at maximum contraction and Valsalva. Therefore, it was checked if only frames containing contraction and Valsalva are sufficient for correct classification. The frame with the largest p_n was identified, which likely contains the state of maximum contraction or Valsalva. This frame and its three subsequent frames were selected for further analysis since these frames likely provide a snapshot of the state of contraction and Valsalva. When the maximum contraction or Valsalva frame was within the last three frames, the last four frames were selected. The classification results using the LFVs from these 4 frames were compared to the results of the LFVs from the original 22 frames.

Next, two dimensionality reduction methods were applied to the 4 or 22 LFVs as a post-processing step before applying clustering algorithms. First, PCA was applied and the first two principal components were kept. Second, the benefit of nonlinear dimensionality reduction was investigated by training an 2D-convolutional autoencoder (2D-CAE) [see Fig. 2(b)] on the 2D latent feature data. The design of the 2D-CAE is similar to the 3D-CAE design. The input is the 22 or 4×128 LFV representation of the TPUS movie. The encoding part compresses this into a new LFV of two elements. The decoding part reconstructs the original 22 or 4×128 LFV TPUS movie representation. The mean squared error loss was used to train the 2D-CAE, which reduces the LFV-data of 4- or 22 frames to a LFV of size 2.

2.4 *Clustering*

Two clustering algorithms were used to identify clusters in the dataset that might allow classification of the contraction and Valsalva TPUS movies: K-means clustering[35] and Gaussian mixture modeling (GMM).[36] Since the difference between contraction and Valsalva TPUS is the classification task of this work, both methods were applied to identify two clusters. The default sklearn implementation in Python 3 was used for both methods. This means that the Elkan[37] implementation was used for K-means clustering and independent covariances were calculated for each cluster in GMM. The clustering algorithms were applied to the full latent feature space of 22- and 4-frames and to a reduced feature space, which was lowered in dimension by either applying PCA or the 2D-CAE. The best performing clustering models were selected based on the lowest training loss, and no separate validation set was used.

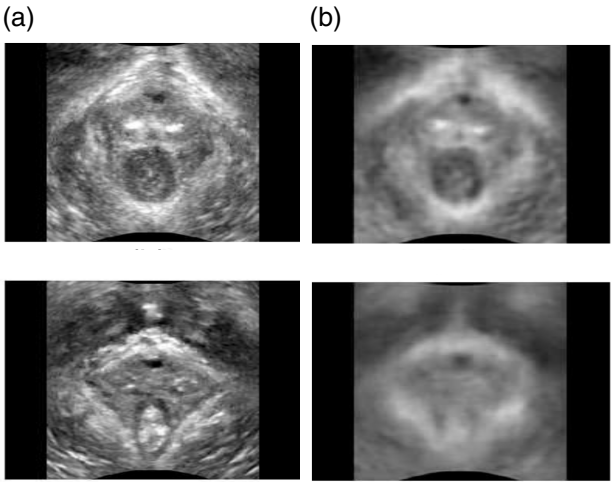

Fig. 3 Slices of (a) two original TPUS frames and their reconstruction by the 3D-CAE are shown to allow visual examination of (b) the reconstruction results. The top row shows an image from the training set and the bottom row the validation image.

3 Results and Discussion

The 3D-CAE was saved with a validation mean squared error loss of 7.72×10^{-3}. Figure 3 shows slices form a TPUS frame and its reconstruction. Most small details are lost in the reconstruction, however, the general appearance of the image is preserved. The clinical structures, such as the muscles, bone, urethra, vagina, and rectum, are still recognizable and preserve their initial shape. This information is thus successfully captured within the LFV. The conversion of a single frame into an LFV was around 2 s on a laptop (Macbook Pro 2015).

3.1 *Latent Feature Vector Analysis*

To obtain a better understanding of the latent feature space learned by the 3D-CAE, the LFVs of the training set were analyzed for their contribution in storing either contraction or Valsalva information. Therefore, the (co)variance of the features was analyzed and statistical tests were performed.

3.1.1 *(Co)variance analysis*

Figure 4(a) shows the variance of the individual LFV features within the training set. The variance of all frames, only contraction frames and only Valsalva frames are presented. The variance allows for identification of prominent features in decoding contraction and Valsalva movement. Most features show similar variance for the complete dataset as for the Valsalva and contraction data. However, the features with most variance (33, 47, and 50) also show significant difference in variance for contraction and Valsalva data. Most notable is feature 47, which shows the most prominent difference: almost all variance on this element in the dataset seems to be in Valsalva data, whereas the contraction data barely shows any variance on this element. The tests presented in Fig. 5 reinforce the idea that this feature mainly captures Valsalva information since all tests show significant difference except the Wilcoxon ranked sum test comparing the difference between rest and maximum contraction. Feature 50 shows the largest opposite difference: the Valsalva variance is approximately two-thirds of the contraction variance. The Wilcoxon ranked sum test for this feature also shows significance for contraction but not for Valsalva, reinforcing the idea that this feature is mostly involved in capturing contraction-specific changes. Feature 33 likely captures both movements since the variance of the complete dataset is larger than for both individual movements. However, the Wilcoxon ranked sum test only shows significance for Valsalva, which suggests that this feature captures Valsalva specific change.

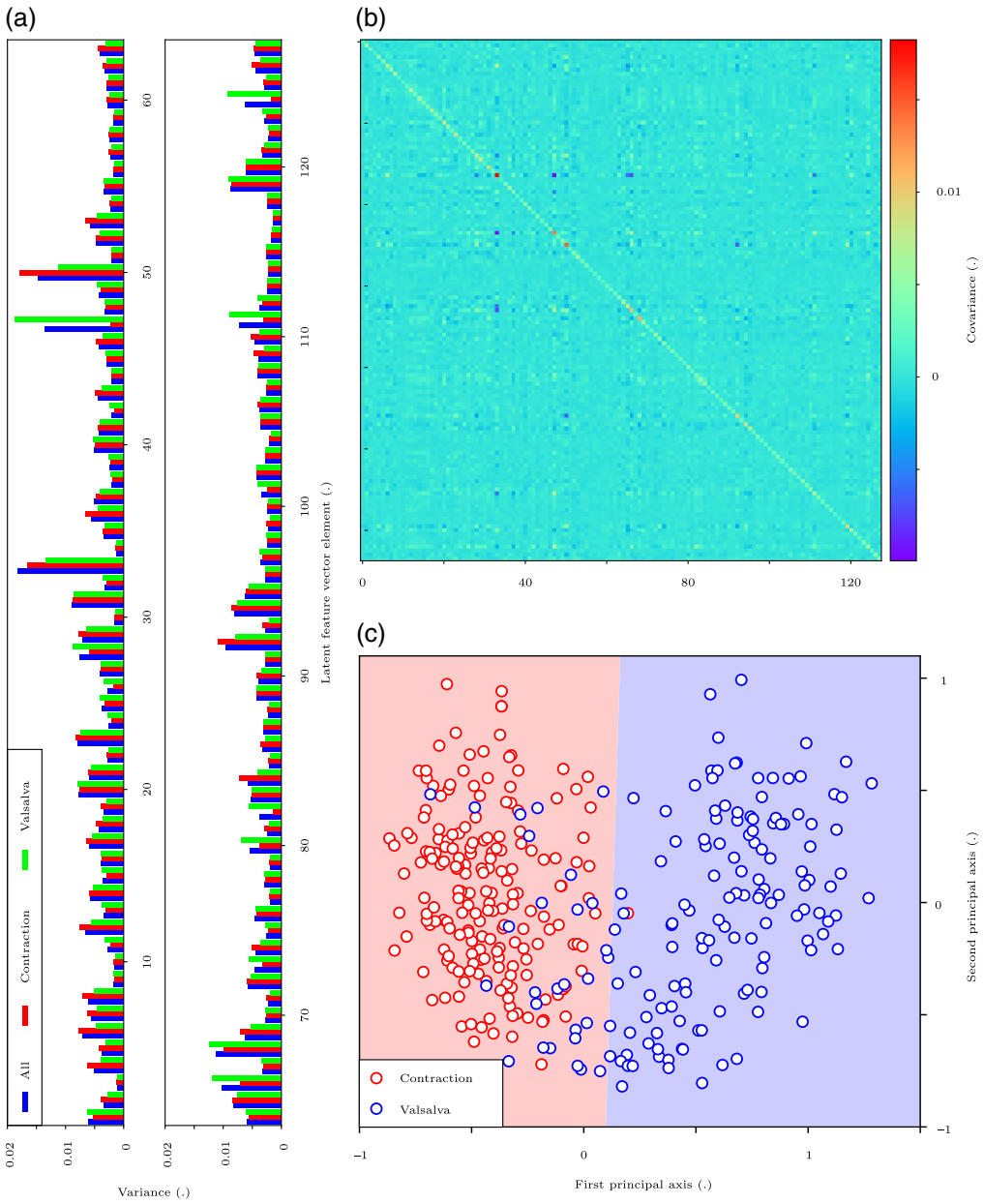

Fig. 4 (a) The variance within the latent feature elements of all frames in the complete training set (blue), the frames of contraction data (red) and of Valsalva data (green). (b) The covariance matrix of the LFVs of all the frames in the trainging set. (c) PCA and *K*-means clustering are applied to LFVs of four frames (likely of maximum contraction or Valsalva) from each TPUS movie in the training set. This plot shows the data points with respect to the first two principal axes, together with the decision boundary of the *K*-means clustering (training accuracy 91.6% and testing accuracy of 91.2%.).

In Fig. 4(b), the covariance matrix of the LFVs is presented. We will discuss the correlations with the three most prominent features for which $|\text{covariance}| > 0.015$, feature 33, 47, and 50. Feature 33 has the largest variance and is also negatively correlated with a few features that have relatively high variance (feature 47, 65, 66, and 111), there is a positive correlation with feature 6. Feature 47 has in turn positive correlation with feature 66, 111, and 124. Inspecting these features in Fig. 4(a) shows that their variation in contraction data is low, this makes it a reasonable assumption that these features (especially 47, 111, and 124) capture variance that is present in Valsalva movement. This idea is reinforced by the results of the Wilcoxon ranked sum test. Since feature 33 is negatively correlated with most of these features, it therefore also

captures Valsalva movement. Feature 50 has a positive correlation with feature 66 and 119, and a negative correlation with feature 92.

It is noteworthy that some features (47, 82, 111, and 124) are dominated by variance in the Valsalva data and barely have any variance in the contraction data. Although features like feature 50 are more dominated by variance related to contraction, they still also contribute significantly to the variance of the Valsalva data. This can be explained from a clinical understanding of the images: the LAM are a prominent element of the image and they are smoothly deformed from contraction to Valsalva (with rest in between), which appears to be captured most dominantly in feature 33. The deformation of these muscles due to movement might be closely related to differences in appearance of the muscle within different women.

A larger image transformation occurs in the case of Valsalva. The abdominal pressure is raised during Valsalva, causing the pelvic organs (e.g., bladder) to descent and enter the TPUS field of view (especially for pelvic organ prolapse[14]). Since these objects are not (as prominent) in view during rest and contraction, it is likely that the features dominated by a large Valsalva variance capture these image changes.

3.1.2 *Statistical analysis*

To analyze if the changes observed in all features are significantly different, several statistical tests were applied. The Shapiro–Wilk tests for all features resulted in p-values <0.05, so the features were nonnormally distributed. However, since the tests were performed on >8000 data points, a slight deviation from normality would already yield a significant result. Still the results of the t-test are presented since most histograms of the features show that they only slightly differ from normality. In Fig. 5, the p-values of the statistical tests are presented, the majority of the tests showed significant differences ($p < 0.05$). Most of these significances still hold, even if a correction for multiple comparisons would be applied.

The Wilcoxon signed rank sum test compares if there is a significant difference between frames in a single movie and therefore provides the best indication on features capturing Valsalva and contraction motion since patient specific features do not change within a single TPUS movie. Most features (85 for $p < 0.05$ and 73 for $p < 0.01$) are significantly involved in capturing Valsalva specific volume changes. Less features (52 for $p < 0.05$ and 34 for $p < 0.01$) are involved in capturing contraction specific changes. The majority of features (102 for $p < 0.05$ and 89 for $p < 0.01$) are involved in capturing either contraction of Valsalva changes. Nine features (17, 18, 57, 59, 79, 85, 87, 116, and 127) show no or only weak significance on all tests and are therefore likely not involved in storing contraction or Valsalva specific information.

Based on the variance and statistical analysis, it is possible to get a general impression on how the contraction and Valsalva information is stored by the 3D-CAE in the LFV. Furthermore,

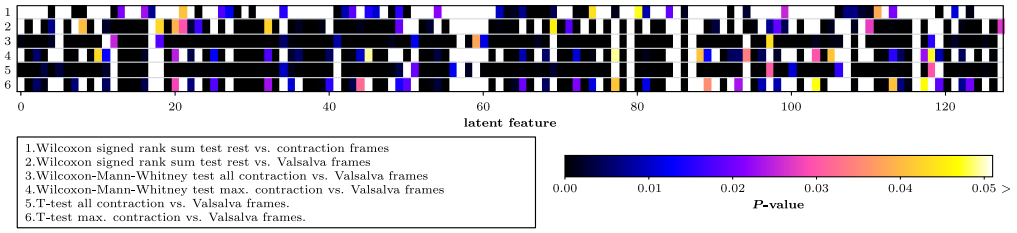

Fig. 5 A visual table of the p-values of the several statistical tests applied to the LFVs. The first and second row represent the Wilcoxon signed rank sum test of the LFVs of the resp. contraction and Valsalva data. Testing for a significant difference in a specific feature between the rest frame and either maximum (max.) contraction or Valsalva frame. The third and forth row show the results of the Wilcoxon–Mann–Whitney tests, which test the similarity of distributions for all contraction and Valsalva frames (row 3) and only max. contraction and Valsalva frames (row 4). The results of the t-test for the same groups are shown in the last two rows, to check for significant changes is means of contraction and Valsalva.

prominent features can be distinguished. As enforced by the loss function, all features show variance and therefore seem to store some relevant reconstruction information. As the statistical tests show, this results in most features significantly contributing to storing information relevant for contraction and Valsalva.

3.2 *TPUS Motion Analysis*

PCA is applied to the 22 LFVs of the individual TPUS movies from the testing set. Each TPUS frame is represented in a new feature space and has a value (f_n) on the first principal axis, which represents the most variance (on average 73%). From this, p_n is calculated and plotted per movie in Fig. 6 for contraction (a) and Valsalva (b). Contraction TPUS movies were recorded with the intent to capture the movement from rest to contraction and back to rest again since this process takes only a few seconds. If the examiner noticed that the process took longer, the movie was stopped at maximum contraction. Valsalva is performed slower and due to the maximum recording limit of 22 frames, the procedure was to stop the recording at maximum Valsalva.

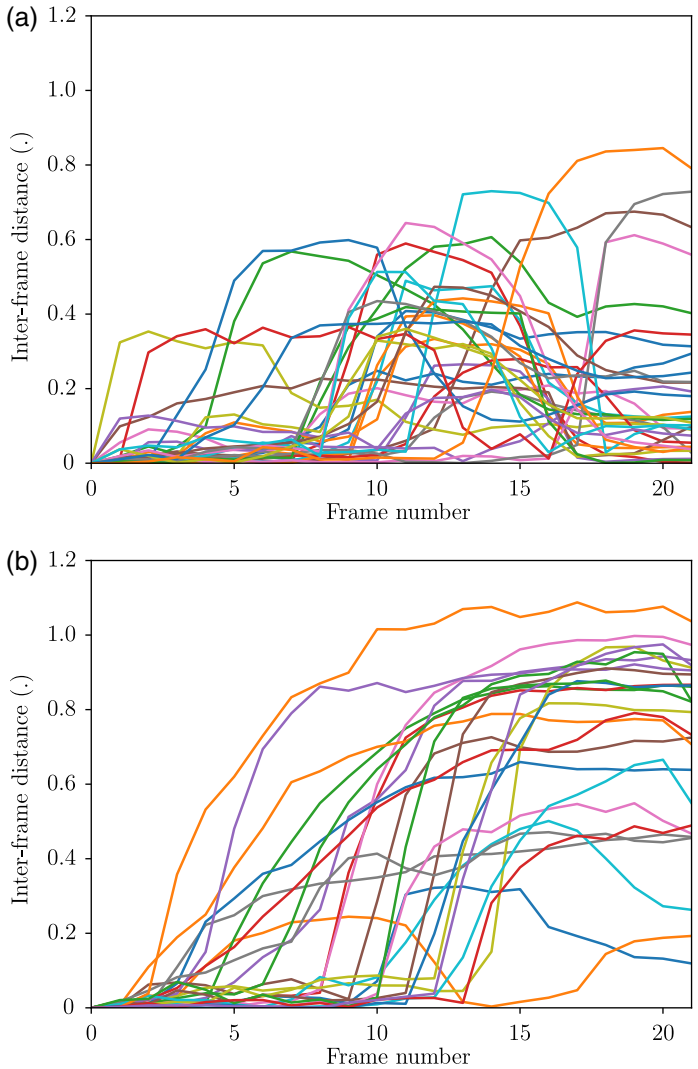

Fig. 6 Both plots display the interframe distance ($p_n = |f_n - f_0|$) from a given frame (f_n) with respect to the first frame (f_0), on the first principal axis of a TPUS movie. Each line in the plot represents a single movie of (a) contraction or (b) Valsalva in the testing set.

These described patterns are visible in these plots: the interframe distance of contraction starts increasing, peaks, and then decreases to almost zero. For Valsalva, the interframe distance slowly increases to the maximum at the end of the movie. When these results are visually compared with muscle movement within the TPUS movie, it appears to correlate very well with the clinical experts opinion when the muscle starts moving and when maximum contraction or Valsalva occurs. However, not only all patients are able to perform a proper contraction or Valsalva, but also the attempts did still generate some movement. Therefore, it is difficult to distinguish these attempts from proper contraction or Valsalva in these plots.

The possibility to identify maximum contraction and Valsalva is relevant for automating the analysis of TPUS data. The current clinical manual analysis is performed on the frames where the LAM are at rest, maximum contraction, and maximum Valsalva. In existing automating attempts, this frame still needs to be identified manually.[16,20] Recently, there has been a successful method to measure LAM strain on TPUS data.[8] Since this is still a new procedure, it is important to validate the measurement with manual identification of maximum contraction of Valsalva. When deploying this strain analysis method to a larger dataset, it can be beneficial to compare the results with our PCA movement analysis. Only in the case of a large mismatch in maximal strain measurement and this PCA identification of maximum contraction or Valsalva, an expert is necessary to relabel the data, which significantly reduces the analysis time of the total dataset for this expert.

3.3 *Unsupervised Classification*

While deploying unsupervised classification (both GMM and K-means clustering) to the training set, 24 TPUS data were consistently mislabeled. These images were examined by an medical expert (C. M.) and 14 were originally mislabeled. The other 10 were in principle labeled with the correct maneuver but were not well performed by the patient. The labels of our training set are updated based on the clinical experts judgment, making the labels more reliable, however, most labels of the training set are still based on the assumed acquisition order. As mentioned before, the labels of the testing set are verified by the expert.

Table 1(a) shows the accuracy of the unsupervised classification, of Valsalva and contraction TPUS movies, on the training and testing set, after different dimensionality reduction and clustering methods are applied to the LFVs of these movies. The best results for unsupervised classification are obtained on the LFV of the 4 frames with the most distance on the first principal axis to the first frame, with most accuracies lying around 90%. Even though the best results on both testing (91.2%) and training sets (91.9%) are obtained after PCA is applied and GMM is used for clustering, the results are comparable for all methods applied to 4-frame data. Only GMM clustering applied to the full feature space of 4-frames performs poorly, it is unclear why this is the case. Applying PCA as a dimensionality reduction method seems to be sufficient, whereas the 2D-CAE provides slightly worse results, especially on the testing set. However, most percentual differences on the 4-frame data only represent a few misclassifications, which makes it difficult to draw strong conclusions on the best methods. Since the 4-frame data consistently outperforms the 22-frame data, we can conclude that it is beneficial to select frames that likely to represent maximum contraction and Valsalva in a movie.

Table 1(b) also presents the percentages of correctly labeled contraction and Valsalva data; this can be considered analogous to the sensitivity and specificity. For the 22-frame classifications, the contraction and Valsalva data are classified correctly at similar percentages. For the 4-frame classification, the percentage of correctly labeled contraction TPUS is in almost all cases high ($>$90%), whereas the results for Valsalva are lower (80% to 85%). The 4-frame data capture only the extremes of contraction and Valsalva, reducing the noise for this classification task. The misclasssifcations are likely due to the fact that patients are not always able to perform a proper Valsalva maneuver.

In Fig. 4(c), the training data and decision boundaries are plotted after application of PCA and K-means clustering, to provide an insight in how the data distributions appear. Most misclassifications lay closely to the decision boundary, therefore these data points are likely examples of weak contractions and Valsalva. This idea is also reinforced by the experts opinion on the 24 consistently misclassified datapoints.

Table 1 (a) The accuracy of *K*-means clustering (*K*-means) and GMM in clustering contraction and Valsalva data in the training and testing set. (b) The percentages of correctly labeled contraction (Cont.) and Valsalva (Val.) TPUS data. These measures are analogous to sensitivity and specificity. The clustering algorithms were provided with data from the LFVs of, respectively, 22 and 4 frames. The clustering algorithms were either applied to the entire LFV space or to a reduced two component space when PCA or the 2D-CAE were applied. In table (a), the highest training and testing accuracies are presented in bold. In table (b), the best combinations are presented in bold.

(a) Accuracy

		22 Frames		4 Frames	
		Train (%)	Test (%)	Train (%)	Test (%)
	K-means	83.4	82.5	91.3	**91.2**
	GMM	84.3	86.0	91.3	57.9
PCA	*K*-means	82.3	82.5	91.6	**91.2**
	GMM	80.6	77.2	**91.9**	**91.2**
D-CAE	*K*-means	80.6	77.2	**91.9**	87.7
	GMM	80.9	77.2	91.0	84.2

(b) Correctly labeled

		22 Frames				4 Frames			
		Train		Test		Train		Test	
		Cont.	Val.	Cont.	Val.	Cont.	Val.	Cont.	Val.
	K-means (%)	82.4	84.7	81.3	84.0	99.4	81.5	**96.9**	**84.0**
	GMM (%)	85.1%	83.4%	96.9%	72.0%	99.4%	81.5%	100%	4.0%
PCA	*K*-means (%)	83.5	80.9	81.3	84.0	99.4	82.2	**96.9**	**84.0**
	GMM (%)	76.1	85.9	71.9	84.0%	**99.4**	**82.8**	**96.9**	**84.0**
D-CAE	*K*-means (%)	76.6	85.4	71.9	84.0	**99.4**	**82.8**	90.6	84.0
	GMM (%)	78.2	84.1	71.9	84.0	99.4	80.9	87.5	80.0

3.4 *General Discussion*

The results presented and discussed in this paper started with the training of the 3D CAE. This autoencoder was engineered based on general ideas present in the field of image segmentation (like the design of Unet,[30]) and optimized to fit on the training server. However, we did not vary the design to test which design works best for capturing relevant image features. Especially varying the size of the latent feature space will influence the quality of the reconstruction and change the amount of relevant information that can be stored. It is, however, difficult to control the learning of the network, so even training the current network from scratch will yield a different latent feature space since the initialization of the network weights is random. This will impact the LFV analysis presented in this work (Sec. 3.1), with regards to the conclusion for specific features. However, since this analysis shows that contraction and Valsalva are prominent characteristics with information stored in a majority of the features, the results of the motion analysis and classification will be reproducible. For this study, we did not investigate the question about the optimal design since the results on classification were already good and optimizing the

design can be a study in itself. The same holds for the loss function, we made the choice for a loss function that enforces the use of all features for encoding since this allows for capturing image information in more distinguishable entities. However, one could easily make the opposite claim that allowing for only a few features to be used provides more distinct features. We did not investigate which claim is correct. We consider this study a proof-of-concept that can be optimized.

During training of the network, we only used a single frame as validation set to prevent overfitting. For the clustering algorithms, we did not use a validation set but instead used the training "labels" to select the best model. Both approaches can result in overfitting, however, the results on the testing set suggest that overfitting did not occur.

For the analysis of the latent feature space, we focused on contraction and Valsalva only since that is the scope of this paper. This provides, however, limited insight into how change in value of a specific feature impact the image reconstruction.

Although our p_n representation of the frames allowed for the analysis of changes within a movie, there is no way to discriminate between movement induced by the contraction or Valsalva and other types of movement generated by the ultrasound examiner or the patient. This makes this method not completely reliable for identifying maximum contraction or Valsalva.

The strength of this study is that the method is completely unsupervised. The results of the training set are almost as indicative for the performance as the results for the testing set since the results on the latter show that overfitting did not occur. There was no specific patient selection so the heterogeneous dataset represents the variety of patients entering the urogynecological clinic. We therefore are optimistic that the results can easily be translated to be used in clinical or research practice.

3.5 *Future Work*

In this work, we have analyzed the learned TPUS features related to contraction and Valsalva. However, potentially more relevant information can be acquired from these LFVs. Diagnosing pathologies like LAM muscle avulsions might be possible based on the information stored within the LFVs. Analyzing only the first frames will help to remove the contraction and Valsalva variation within the LFVs and only show interpatient differences. Furthermore, the information stored in the features can be visually analyzed; when a value of a specific feature is varied, the TPUS frame can be reconstructed and the effect of this variation can be visually examined. This might help improve our understanding of the TPUS data itself. To improve the reliability of the TPUS movie motion analysis, a better profile of feature change during contraction and Valsalva should be established in order to discriminate this from other sources of movement.

4 Conclusion

In this work, we presented a 3D-CAE to compress tranperineal ultrasound frames into 128 element LFVs. The information stored in these LFVs proved to obtain relevant information to further analyze the muscle movement present in the 4D TPUS data. Furthermore, it allowed for unsupervised classification of 4D TPUS data into the contraction and Valsalva labels with high accuracy. Applying the methods to medical and research TPUS datasets will significantly reduce the labeling and frame selection time for experts.

Disclosures

The authors declare no conflicts of interest.

Acknowledgments

The authors would like to thank the editor and referees for the constructive suggestions. This study is part of the Gynaecological Imaging using 3D Ultrasound project funded by the Dutch Science Organization (NWO, Grant No. 15301).

References

1. G. Rortveit et al., "Urinary incontinence, fecal incontinence and pelvic organ prolapse in a population-based, racially diverse cohort: prevalence and risk factors," *Female Pelvic Med. Reconstruct. Surg.* **16**(5), 278–283 (2010).
2. A. M. Hill et al., "Pregnant women's awareness, knowledge and beliefs about pelvic floor muscles: a cross-sectional survey," *Int. Urogynecol. J.* **28**(10), 1557–1565 (2017).
3. J. O. L. DeLancey, "Mommy, how will the baby get out of your tummy? Will it hurt you?," *Am. J. Obstetr. Gynecol.* **217**(2), 110–111 (2017).
4. K.-C. Lien et al., "Levator ani muscle stretch induced by simulated vaginal birth," *Obstetr. Gynecol.* **103**(1), 31–40 (2004).
5. L. Hoyte et al., "Quantity and distribution of levator ani stretch during simulated vaginal childbirth," *Am. J. Obstetr. Gynecol.* **199**(2), 198.e1–198.e5 (2008).
6. N. Sindhwani et al., "In vivo evidence of significant levator ani muscle stretch on MR images of a live childbirth," *Am. J. Obstetr. Gynecol.* **217**, 194.e1–194.e8 (2017).
7. H. P. Dietz, "Pelvic floor ultrasound: a review," *Clin. Obstetr. Gynecol.* **60**(1), 58–81 (2017).
8. S. Das et al., "3D ultrasound strain imaging of puborectalis muscle," *Ultrasound Med. Biol.* **47**(3), 569–581 (2021).
9. F. van den Noort, B. Sirmacek, and C. H. Slump, "Recurrent U-net for automatic pelvic floor muscle segmentation on 3D ultrasound," https://arxiv.org/abs/2107.13833 (2021).
10. C. Manzini et al., "Appearance of the levator ani muscle subdivisions on 3D transperineal ultrasound," *Insights Imaging* **12**, 91 (2021).
11. K. W. M. van Delft et al., "Agreement between palpation and transperineal and endovaginal ultrasound in the diagnosis of levator ani avulsion," *Int. Urogynecol. J.* **26**(1), 33–39 (2015).
12. A. T. M. Grob et al., "Changes in the global strain of the puborectalis muscle during pregnancy and postpartum," *Ultrasound Obstetr. Gynecol.* **51**(4), 537–542 (2018).
13. M. K. Van de Waarsenburg, C. H. van der Vaart, and M. I. J. Withagen, "Structural changes in puborectalis muscle after vaginal delivery," *Ultrasound Obstetr. Gynecol.* **53**(2), 256–261 (2019).
14. H. P. Dietz et al., "Ballooning of the levator hiatus," *Ultrasound Obstetr. Gynecol.* **31**(6), 676–680 (2008).
15. H. P. Dietz, A. Pattillo Garnham, and R. Guzmán Rojas, "Is it necessary to diagnose levator avulsion on pelvic floor muscle contraction?," *Ultrasound Obstetr. Gynecol.* **49**(2), 252–256 (2017).
16. F. van den Noort et al., "Deep learning enables automatic quantitative assessment of puborectalis muscle and urogenital hiatus in plane of minimal hiatal dimensions," *Ultrasound Obstetr. Gynecol.* **54**, 270–275 (2019).
17. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, Vol. 1, pp. 1097–1105 (2012).
18. I. Goodfellow, Y. Bengio, and A. Courville, *Deep-Learning*, 1st ed., The MIT Press, Cambridge, Massachusetts (2016).
19. G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.* **42**, 60–88 (2017).
20. E. Bonmati et al., "Automatic segmentation method of pelvic floor levator hiatus in ultrasound using a self-normalizing neural network," *J. Med. Imaging* **5**(2), 021206 (2018).
21. H. Williams et al., "3D convolutional neural network for segmentation of the urethra in volumetric ultrasound of the pelvic floor," in *IEEE Int. Ultrasonics Symp.*, pp. 1473–1476 (2019).
22. H. Williams et al., "Automatic extraction of hiatal dimensions in 3-D transperineal pelvic ultrasound recordings," *Ultrasound Med. Biol.* **47**(12), 3470–3479 (2021).
23. F. van den Noort et al., "Automatic identification and segmentation of the slice of minimal hiatal dimensions in transperineal ultrasound volumes," *Ultrasound Obstetr. Gynecol.* **60**, 570–576 (2021).

24. G. E. Hinton et al., *Unsupervised Learning: Foundations of Neural Computation*, 1st ed., The MIT Press, Cambridge, Massachusetts (1999).
25. K. Fukushima, "Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.* **36**, 193–202 (1980).
26. G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science* **313**(5786), 504–507 (2006).
27. J. Masci et al., "Stacked convolutional auto-encoderss for hierarchical feature extraction," in *Artif. Neural Networks and Mach. Learn. – ICANN 2011*, pp. 52–59 (2011).
28. L. J. P. Van Der Maaten, E. O. Postma, and H. J. Van Den Herik, "Dimensionality reduction: a comparative review," *J. Mach. Learn. Res.* **10**, 1–41 (2009).
29. A. Nagpal, A. Jatain, and D. Gaur, "Review based on data clustering algorithms," in *Proc. 2013 IEEE Int. Conf. Inf. and Commun. Technol. (ICT 2013)*, pp. 298–303 (2013).
30. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
31. P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," arXiv:1710.05941 (2017).
32. V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Int. Conf. Mach. Learn.*, Madison, Wisconsin, Omnipress ed., pp. 807–814 (2010).
33. D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," in *3rd Int. Conf. Learn. Represent., ICLR 2015—Conf. Track Proc.* (2015).
34. K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philos. Mag. J. Sci.* **2**(11), 559–572 (1901).
35. S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory* **28**(2), 129–137 (1982).
36. D. A. Reynolds and R. C. Rose, "Speaker identification and verification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.* **3**(1), 72–83 (1995).
37. C. Elkan, "Using the triangle inequality to accelerate K-means," in *Proc. Twentieth Int. Conf. Mach. Learn.*, Vol. 1, pp. 147–153 (2003).

Biographies of the authors are not available.