Title: Automatic quality assessment of transperineal ultrasound images of the male pelvic region using deep learning

Authors: S.M. Camps[1,2], T. Houben[1], G. Carneiro[3], C. Edwards[4], M. Antico[5,6], M. Dunnhofer[7], E.G.H.J. Martens[8], J.A. Baeza[8], B.G.L. Vanneste[8], E.J. van Limbergen[8], P.H.N. de With[1], F. Verhaegen[8], D. Fontanarosa[4,5]

[1] Faculty of Electrical Engineering, Eindhoven University of Technology, Eindhoven, the Netherlands
[2] Oncology Solutions Department, Philips Research, Eindhoven, the Netherlands
[3] Australian Centre of Visual Technologies, the University of Adelaide, Adelaide, Australia
[4] School of Clinical Sciences, Queensland University of Technology, Brisbane, Queensland, Australia
[5] Institute of Health & Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia
[6] School of Chemistry, Physics and Mechanical Engineering, Queensland University of Technology, Brisbane, Queensland, Australia
[7] Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy
[8] Department of Radiation Oncology (MAASTRO), GROW – School for Oncology and Developmental Biology, Maastricht, the Netherlands

Author for correspondence
Davide Fontanarosa
School of Clinical Sciences
Queensland University of Technology
Gardens Point campus, 2 George St, Brisbane, QLD 4000
E-mail: d3.fontanarosa@qut.edu.au
Work Ph: +61 (7) 3138 2585
Mob Ph: +61 (0) 403862724

Abstract

Ultrasound guidance is not widespread in prostate cancer radiotherapy workflows. This can be partially attributed to the need for image interpretation by a trained operator during ultrasound image acquisition. In this work, a one-class regressor, based on DenseNet and Gaussian processes, was implemented to automatically assess the quality of transperineal ultrasound images of the male pelvic region. The implemented deep learning approach was tested on 300 transperineal ultrasound images and it achieved a scoring accuracy of 94%, a specificity of 95% and a sensitivity of 92% with respect to the majority vote of three experts, which was comparable with the results of these experts. This is the first step towards a fully automatic workflow, which could potentially remove the need for ultrasound image interpretation and make real-time volumetric organ tracking in the RT environment using ultrasound more appealing.

Keywords: Transperineal ultrasound imaging, deep learning, prostate, image guided radiotherapy, ultrasound, radiotherapy

Introduction

One of the curative treatment modalities for prostate cancer is radiotherapy (RT). This modality aims to irradiate tumor tissue in the prostate (sometimes including the seminal vesicles), while sparing the surrounding organs at risk (OAR) (e.g. bladder and rectum) as much as possible. The radiation dose is typically delivered to the patient in multiple treatment fractions, in accordance with a treatment plan designed based on a computed tomography (CT) scan.

It has been shown that the shape and position of the prostate might differ between treatment fractions (inter-fraction), due to changes in bladder and/or rectal filling (Roeske et al. 1995). Also during a treatment fraction (intra-fraction) the tissue distributions might change (Ballhausen et al. 2014). If the original treatment plan was delivered on the changed tissue configuration, this could potentially result in a suboptimal dose deposition in the tumor and the organs at risk could receive extra undesired dose (Fraser et al. 2010).

For this reason, several solutions have been proposed to identify the position and shape differences of the anatomical structures during the treatment course with respect to the treatment plan. This information can be used to potentially improve dose delivery precision. Most of the proposed solutions require frequent imaging during the course of the RT treatment (image guided RT, IGRT) with or without implanted fiducial markers (van der Heide et al. 2007) using X-ray, magnetic resonance imaging (MRI) (Lagendijk et al. 2008; McPartlin et al. 2016) or ultrasound (US) imaging (Camps et al. 2018; Fontanarosa et al. 2015; O'Shea et al. 2016). In the typical clinical workflow of prostate cancer patients, the treatment plan is based on a simulation CT scan. Then, prior to a treatment fraction, cone-beam CT (CBCT) imaging of the bony structures and/or of fiducial markers implemented in the prostate is used to identify and correct for inter-fraction changes. Intra-fraction changes are typically not taken into account.

In this work, we focused on the use of US imaging for intra-fraction guidance during RT. US imaging allows real-time volumetric organ tracking in the RT environment and, in addition, it is cost-effective and harmless for the patient. Currently, there is one system available on the market (Clarity

Autoscan system, Elekta, Stockholm, Sweden) that allows for intra-fraction transperineal US (TPUS) imaging of the male pelvic region for prostate motion monitoring during RT (Lachaine and Falco 2013). Despite the advantages of US imaging and its availability on the market, the use of this image modality in the RT workflow is not yet widespread. This can be partly attributed to the need for a trained operator during manual image acquisition to verify if the correct anatomical structures are visualized with sufficient quality.

To allow for intra-fraction monitoring of anatomical structures, the operator needs to position the US probe prior to treatment fraction commencement. As the operator cannot stay in the treatment room during radiation delivery, the probe would need to be fixed using either a mechanical or a robotic arm. During the treatment fraction, small motion of the patient or changes in anatomical structures can compromise image quality due to, for example, a loss of acoustic coupling and/or a sudden appearance of shadowing artifacts. In the workflow of the above mentioned Clarity Autoscan system, a quality metric is output during intra-fractional motion monitoring. However, this metric gives information on the localization quality of the target (prostate in this case) and does not take into account the OARs. In addition, it does not provide information on the overall image quality of whole 2D slices or US volumes. In order to evaluate the overall US image quality and have information on the OAR, the operator would need to be present in the control room to promptly identify this quality loss and, if necessary, take appropriate action.

The aim of this study was to develop a deep learning algorithm to automatically score 2D US images out of a 3D volume of the male pelvic region based on their quality or, in other words, on their usability during the US guided RT (USgRT) workflow. This algorithm should achieve an accuracy at least equal to the worst performing expert consulted in this study. Machine learning has been used before in the assessment of US image quality, primarily in the obstetrics field (e.g. (Rahmatullah et al. 2011; Zhang et al. 2017)). In these studies, the assessment was based on initial segmentations or on the presence of specific anatomical structures in the image. In the work of Schwaab et al. (2016) the quality of 3D US images of the breast was automatically assessed; however, this work made use of handcrafted features, such as the total 2D physical area of the breast.

Several different aspects of the image acquisition procedure as well as the subject's body composition affect US image quality. Quality deterioration can be caused by many factors including insufficient acoustic coupling between the US probe and the skin, bones causing shadowing artifacts on critical anatomical structures and insufficient penetration due to (fat) tissue distributions. This makes describing features for classification challenging. Therefore, we aimed to perform the quality assessment using solely automatically learned deep features from the image without relying on any initial segmentations or specific anatomical structure detection. In particular, we developed a novel one-class regressor, based on DenseNet (Huang et al. 2017) and Gaussian processes (GPs) (Kemmler et al. 2013). This is the first step towards a fully automated workflow that would eventually remove the need for a trained operator and therefore potentially make the use of US imaging more appealing for hospitals.

Materials and methods

Image data acquisition

In this work, datasets from three different studies (METC153017, P0223, P0053) conducted after local institutional review board or medical ethics committee approval at the MAASTRO Clinic (Maastricht, the Netherlands) were combined (Table 1). The 36 male subjects were either healthy volunteers or patients with localized prostate cancer and all provided informed, signed consent. For each subject, several 3D and 4D TPUS volumes of the pelvic region were acquired using an X6-1 xMatrix array probe (Philips Healthcare, Bothell, WA, United States) and an EpiQ7 US system (Philips Medical Systems, Andover, MA, United States).

The used datasets showed a significant variability in image characteristics due to, for example, the varying body composition (BMI: 25.6 [mean] ± std 3.6 [std] based on 32 subjects only, for four subjects the BMI was not obtained), age (62 [mean] ± 18 [std] years) and medical history of the subjects. The variability in image characteristics also resulted from the fact that the exact settings on the US system such as imaging depth and focus varied between the different studies and between the different subjects. The majority of the subjects has been imaged with the following settings: 75° by 75° viewing angles, 7.5cm focus depth, 11cm imaging depth, focus on penetration (HPEN setting) and resolution (RS setting). In addition, both the volume dimensions ([X,Y,Z] = [315 ± 22, 255 ± 25, 196 ± 16] voxels) and voxel sizes ([X,Y,Z] = [0.3678 ± 0.0352, 0.5662 ± 0.0529, 0.7372 ± 0.0718] mm) of the acquired US volumes varied, due to the different settings on the US system (as mentioned above) and due to a requirement to achieve an acceptable frame rate (about 2Hz) in the 4D sequences. Some datasets showed anatomical structure displacements, which were artificially introduced by instructing the subjects to consciously contract muscles in the pelvis area or to cough. Finally the variability in the datasets resulted from the fact that four radiation oncologists were involved in the acquisition of the volumes, each of them with their own approach to TPUS image acquisition with the prototype mechanical arm (Fig. 1) which was used to fixate the probe.

Initial image data pre-processing

Three initial pre-processing steps were necessary to prepare the datasets for processing by a deep learning algorithm. These steps were all performed using MATLAB (Version 9.3.0 (R2017b), The Mathworks Inc. Natick, MA, United States). First, the volumes were resampled to the largest voxel size present in the database (1.0292mm x 1.0292mm x 1.0292mm), which allowed easy volume comparisons and batch processing of the data in the next steps. Second, the TPUS volumes were sliced to 2D images along the sagittal direction, as this was the direction with the highest resolution before resampling. Data collection prior and during this study has shown that it is more challenging to achieve consensus of radiation oncologists on the quality score criteria in three dimensions than in two dimensions. As an example, scoring differences may arise when a specific anatomical structure of interest is only visible in a part of the 3D volume. In addition, a 2D approach is less computationally expensive which allows processing of images with higher resolution and it is less labor intensive to acquire sufficient image data (as each volume provides multiple 2D image data samples), which is required for algorithm training, in comparison with a 3D approach. Therefore, a 2D approach was chosen for the initial prototype development.

The visual inspection of the 2D images performed by research team member S.C. in the third step revealed that the anatomical structures of interest were most often located at the center of each volume. For this reason, the empirical decision was made that only the central 16 sagittal 2D images from each volume were selected for further processing, which also reduced the total computational cost. Then all 2D images were symmetrically padded with black (zero-valued) pixels to ensure that all images had the same dimensions as the largest 2D image (namely 216x180 pixels) in the entire dataset. Finally, a fixed region of interest was defined by automatically cropping the images based on geometry to primarily remove background pixels while preserving the crucial information of all anatomical structures. This resulted in 178,368 2D TPUS images overall composed of 116x100 pixels originating from 11,148 TPUS volumes.

2D US image classification

The crucial anatomical structures for prostate RT treatments are: prostate, seminal vesicles, bladder and rectum. The prostate is the target of the treatment and should therefore be always completely visible on an acquired US volume. In the ideal case, also the edges of the bladder and rectum adjacent to the prostate should be visible to potentially spare these OARs from excessive radiation exposure. As it was not possible to identify the seminal vesicles with sufficient certainty on the acquired US volumes, these were not evaluated in this study.

Based on the above-mentioned criteria, three image categories were defined which are detailed in Table 2. An example of each category is displayed in Fig. 1. Category 1 involves images that have insufficient quality to be used clinically for USgRT, as the prostate cannot be identified. The quality of Category 2 and Category 3 images was considered sufficient as at least the target (Category 2) or the target and two OARs (Category 3) are visualized and potentially can be tracked.

In order to provide the deep learning algorithm with labeled training, validation and test samples, a subset (16,000 2D images) of the available 2D TPUS images was manually and independently scored by four not medically trained members of our research team (S.C., T.H., M.A. and M.D.), who had experience with US imaging as the experts involved in this study had very limited time available. Three members received training prior to the image classification task from the fourth team member, who gained experience how to interpret the images during multiple TPUS scanning sessions of prostate cancer patients and while performing image processing tasks on these type of images. The central 16 2D images (see Section 2.2) of each volume were presented to each team member. On the images presented to the team members, the cropping as described in the previous section was not performed. Each team member could then scroll through the images of each volume and assign a score between 1 and 3, corresponding to Categories 1 to 3, respectively, to each image. Some of the 2D images were horizontally flipped, due to the fact that the probe was sometimes held upside down. This resulted in a flipped anatomical structure configuration. During the scoring process, the orientation of these images was manually corrected, to ensure that the bladder was located on the left side and the rectum on the right. The team members were instructed to only assign a score to an image if they were highly confident, so it was also possible to leave images unscored. Following this procedure, 1000 randomly selected volumes were scored by each team member.

The images that received a consistent score from at least three out of four team members were included in a database (*Database_NonBinary)* with the majority vote of the scores given by the team members assumed to be their ground-truth annotations. Subsequently, the scores of each team member were binarized, with Score 1 = 0 (poor quality) and Score 2 or 3 = 1 (good quality). Then, the same procedure of including images in the database that at least three out of four team members scored consistently was followed, resulting in a binary database (*Database_Binary)*.

Subject data split and database generation

The research team evaluated overall 16,000 2D TPUS images distributed over 1,000 volumes. In total 13,463 of these images (from 34 out of 36 subjects) received a consistent score from three out of four team members and were therefore included in *Database_Binary*. Subsequently, the data were split into training (60% = 20 subjects), validation (20% = 7 subjects) and test (20% = 7 subjects) sets. For each subject the number of classified images varied. In addition, not for all subjects images of all categories (Category 1-3) were available. For this reason, the split was performed using an optimization approach based on simulated annealing (Kirkpatrick et al. 1983). *Database_Binary* was used to train and test the algorithm. However, the subject split into training, validation and test sets was performed based on

Database_NonBinary. This was done to ensure a balance between good (Category 2) and very good (Category 3) images in the positive binary group.

First, the data were split into a test and train set by randomly assigning the subjects to one of the groups, while not exceeding the defined sizes of each group. Subsequently, in each iteration, a random subject from the test set was swapped with a random subject from the training set. The aim was to obtain similar ratios between the number of images of a certain category (1-3) in each group (test or train) with respect to the total size of that group. So, for example, if 20% of the training images were from Category 1, also about 20% of the test images should be from Category 1. In total 1000 iterations were executed, in which more weight was put on the ratios of Category 2 and 3 images. The ratios of the Category 2 and 3 images were more important, due to the fact that a one-class approach was implemented. This is explained further in Section 2.6. The same process was repeated to extract the validation set from the training set. In the end, this resulted in a distribution of poor-quality images (binary score 0) and good-quality images (binary score 1) over the train, validation and test set as shown in Fig. 2A. In Fig. 2B the distributions of the binary score 0 and 1 images per subject and per group are detailed. In the remainder of this paper, this subject distribution will be referred to as *D0*.

To allow for cross-validation of the hyper-parameters of the deep learning algorithm, nine additional subject distributions were created (*D1 – D9)*. These distributions were also created using the approach based on simulated annealing as described above. However, as the hyper-parameters were optimized based only on the validation set of *D0,* it was not necessary to perform the second step in which the training set is again split into a training set and a validation set. In addition, the distributions were chosen in such a way that each of the 34 subjects appeared at least once in the test set of a distribution. Figure 3 shows the number of test and training images in each distribution (including *D0*) and the subjects indicated by their numbers appearing in the test set are detailed in Table 3.

Quality score validation

Quality score validation was performed by an accredited medical sonographer (C.E.) and by two of the radiation oncologists (B.V. and E.L.) involved in the acquisition of the images. These experts were each presented with the same 300 2D TPUS images, which were randomly selected from the test set of *D0,* and asked to score these images between 1 and 3. Also in this case, the experts were presented with the down sampled but not cropped images. The inter-expert agreement, the test data agreement and the performance of the algorithm were then compared to the majority vote of the experts using Fleiss' kappa (Fleiss 1971)(with the interpretation of Landis and Koch (1977)), accuracy, sensitivity and specificity metrics.

Deep learning algorithm selection

As described in the introduction, several aspects of the image acquisition procedure as well as the subject's body composition affect US image quality, which makes it difficult to describe the features for classification. For this reason, we approached this problem as a one-class classification (OCC) problem. This approach involves the definition of a single class that should contain all images with "good" (according to clinical requirements) quality, while considering the images with "poor" (according to clinical requirements) quality as outliers. One-class support vector machines (OCSVM) can construct a hyper-sphere with a minimum radius, which contains all positive data points in the multi-dimensional feature space (Khan and Madden 2004). However, even though this technique is widely used, it does not perform well on noisy data (Ghahramani 2011).

In this work, the use of Gaussian processes (GPs) instead of conventional SVM was explored for OCC of US image quality. In line with Kemmler *et al.* (2013), GPs were used for regression acting as a one-class classifier. In contrast to SVMs, GPs are robust to noise, deliver probabilistic predictions and are able to automatically learn regularization and kernel parameters as well as feature importance (Ghahramani 2011). In addition, GPs seem promising in knowing when they do not know (Bradshaw et al. 2017). However, GPs lack characterization power for complex data (Bradshaw et al. 2017). For this reason, a combination of two techniques was considered: a convolutional neural network (CNN) was used as an autonomous feature descriptor. Then its output was supplied to the GP for OCC.

Architecture and implementation

In this work, DenseNet (Huang et al. 2017) was used for feature description. This CNN provides a robust architecture which reduces the chance to over-fit and for vanishing gradients, while giving state-of-the-art results on fundamental datasets, like ImageNet (Huang et al. 2017). Within the dense blocks, the characteristic elements of a DenseNet, all layers are all directly connected to all other layers and not only to the subsequent layer, as in more traditional networks. The network implemented in this work contained 2 dense blocks with 18 layers per block and a growth rate k of 12 (see Table 4) and no bottleneck layers were included. Prior to the first dense block, a convolutional operation with a 7x7 pixel filter was performed, followed by a max pooling operation. Finally, the last fully connected layer was removed and replaced by a GP regressor.

This regressor was implemented using GPflow (Matthews et al. 2017). A major advantage of GPflow is that it supports sparse GPs (Titsias 2009), which reduces computation time and memory usage (one of the main drawbacks of GPs). The regressor used a radial basis kernel function (RBF) with an initial variance of 0.1 to fit the data (see Table 4). The number of points used during the GP calculations was 150, which was 75% of the outputs from the CNN. As the GPflow library is built on TensorFlow (Abadi et al. 2016), the DenseNet was also implemented in TensorFlow to make end-to-end training possible.

Prior to providing the deep learning algorithm with the image datasets, two final processing steps needed to be performed on the data. First, all pixel values were normalized by setting the total mean to zero and the standard deviation to unity, to ensure that the training backpropagation algorithm of the CNN would work efficiently. Second, the training data was randomly permuted and then split in mini-batches to ensure subject balance in the mini-batches.

All training and testing was performed on a Linux Cluster with a NVIDIA Tesla K40 GPU with 12 GB VRAM (NVIDIA, Santa Clara, CA, USA). During training, which could take up to two hours, the one-class classifier algorithm was only provided with images with good quality (binary score 1). The optimization was done using the Adam optimizer (Kingma and Ba 2014) and a fixed learning rate, see Table 4. After the deep learning hyper-parameters were optimized (indicated with an asterisk in Table 4) using the validation set of *D0*, the training and validation sets were combined into the final training set of *D0*. Finally, scoring one image from the test set took about 1.5ms using the GPU.

Comparison with other deep learning algorithms

In addition to the one-class approach in which a CNN was combined with GPs, two additional deep learning approaches were implemented for comparison purposes. The first approach also consists of a CNN in combination with GPs, but instead of only training on the positive data (one-class), the network was trained on both the negative and positive classes. This was possible in this study as sufficient negative class data was available. The parameters used in this implementation are detailed in Table 1 of Supplementary Materials A and again the asterisks indicate the optimized parameters based on the validation set of *D0*.

The second deep learning approach consisted of a DenseNet implementation with a softmax classifier attached to it, as described in the paper by Huang *et al.* (2017). With this approach a binary classification was performed, again with the hyper-parameters optimized using the validation set of *D0* (see Table 1 of Supplementary Materials A).

Cross-validation

As described in the previous two sections, the hyper-parameters of the deep learning algorithms were optimized using the validation set of *D0*. To understand if these parameters generalize well over the available dataset, a cross-validation was performed. To this end, three algorithms were trained with these hyper-parameters using the training sets of *D0 – D9* respectively and tested using the corresponding test sets. For each distribution the accuracy, specificity and sensitivity were calculated and finally the mean and standard deviation (σ) were reported.

Workflow implementation and data analysis

The pre-processing of the image data was performed using MATLAB (Version 9.3.0 (R2017b) on a standard PC (i5 CPU, 2.5 GHz, 4 GB RAM). The subsequent implementation, training and testing of the neural network was done using Python 3.5 and TensorFlow on a Linux Cluster with a NVIDIA Tesla K40 GPU with 12 GB VRAM (NVIDIA, Santa Clara, CA, USA). Finally, the obtained results were analyzed by calculating the accuracy, specificity, sensitivity and Fleiss' kappa's. In addition, also a receiver operating characteristics (ROC) curve was generated. All these analyses were also done using MATLAB on the earlier mentioned standard PC.

Results

In Table 5 the cross-validation results are reported as per the implemented deep learning approach. These results are based on the full test sets and not just the expert validated test subset. As the training of the CNN + Softmax based on *D9* with the corresponding hyper-parameters ran out of GPU memory, only the results of *D0* – *D8* were averaged. The CNN + Softmax approach had the worst accuracy and sensitivity in comparison with the CNN + GP approaches. Both CNN + GP approaches performed comparably and the hyper-parameters seem to be able to generalize.

The Fleiss' kappa among the three experts, calculated based on the subset of 300 images randomly picked from the test set of *D0,* was equal to 0.80 (95% confidence interval (CI) [0.77, 0.83]). The kappa among the three experts and the test subset was equal to 0.79 (95% CI [0.77, 0.81]). The accuracy, sensitivity and specificity results with respect to the majority vote of the experts are detailed in Table 6. The accuracy of the test subset with respect to the majority vote was 91%, while the accuracy from the experts ranged within 92% - 97%. The test subset had the lowest sensitivity (80% compared to 90%-99%), but a specificity of 99%.

All algorithms achieved an accuracy, which was equal to or higher than the accuracy of the worst expert (Expert 3) with respect to the majority votes of the experts. The CNN + GP approaches achieved a better accuracy than the CNN + Softmax, while the one-class approach achieved a better accuracy and specificity in comparison with the two-class approach.

In Fig. 4 the ROC curve of the one-class CNN + GP approach is plotted, again with respect to the majority vote of the experts. The square indicates the highest accuracy of the algorithm (94%), which corresponded to a sensitivity of 92% and a specificity of 95% (see Table 6). The circle, cross and upside down triangle indicate the performance of the experts, while the triangle corresponds to the test subset. The Fleiss' kappa for the experts and the algorithm was equal to 0.82 (95% CI [0.80, 0.84]).

Discussion

In this work, a one-class deep learning approach was proposed that could be used to automatically assess the quality of TPUS images of the male pelvic region. For comparison purposes, two additional approaches were also implemented. The CNN + Softmax was not able to train on *D9* as it ran out of memory. This can potentially be explained by the size of the test set of *D9* and by the network depth of the CNN + Softmax network in comparison with the network depth of the CNN + GP approaches. In addition, both the cross-validation results and the validation by experts showed that better accuracy and sensitivity results was achieved using the CNN + GP approaches. It has to be noted that during the optimization of the hyper-parameters the hyper-parameter space has only been explored up to a certain extent. However, the current results seem to justify the use of GPs instead of softmax for classification.

The cross-validation results show a comparable performance of both the one-class and binary classification using CNN + GP, while the expert validation shows a higher accuracy and specificity for the one-class approach. So, the one-class regressor seems to be able to identify the not-usable 2D TPUS images well even though it was only trained on the usable images (i.e. images belonging to categories 2 and 3, which consists of 32% of the training images). This is a very promising result in cases where there is a lack of available not-usable images or it is difficult to capture the whole range of appearances of not-usable images for training purposes.

All the algorithms were trained based on a subset of a larger database and the labels used for training were generated by a small research team. The team members were asked to only assign a score when they were highly confident of their results. In addition, only images to which at least three out of four team members assigned a consistent score were included in the database. This was done to partly eliminate the inter-user variability from the database. In the ideal case, the labels would be generated by experts, but this was not feasible due to time constraints. However, the kappa values (0.80 vs. 0.79) showed a good agreement between the scores of the team-members and the experts. This agreement also shows that, even if the four team members were not fully independent, the resulting database was in agreement with the experts. The accuracy of the test subset was 91%, which is lower than the accuracy of the experts (range: 92% - 97%), but still comparable.

The accuracy, specificity and sensitivity achieved by the one-class CNN + GP algorithm are all higher than the reported results for the test subset (database in comparison with the majority vote of the experts). This could potentially imply that, even though the algorithm was trained with some incorrectly labeled training data, it managed to understand the characteristics of each image category better than the non-experts who provided the initial labels. Another explanation could be that the performance of the non-experts in classifying the test images was not representative for the classification of the training images. Future research investigating which parts of the images contributed to the category decision making and inclusion of more data provided by experts are required to further investigate this phenomenon.

The initial aim was to achieve an accuracy equal to the performance of the worst performing expert (92%). In Fig. 4 it can be observed that the algorithm is able to achieve a sensitivity and specificity that are comparable with the experts, which resulted in an overall accuracy of 94%. Calculating the Fleiss' kappa of the experts and the algorithm resulted in 0.82, which seems to imply that there is almost perfect agreement between the experts and the algorithm (according to the interpretation of Fleiss' kappa from Landis and Koch (1977)). The current performance evaluation was performed with a subset of the test set, due to limited availability of the experts. In future research, this subset will be expanded and the algorithm parameters will be optimized further in order to achieve a 96% accuracy goal, which is the performance of the second to best expert.

The scores assigned to the 2D images were binarized, as currently the quality of Category 1 was considered insufficient for use in clinical practice, while the quality of Category 2 and 3 was considered sufficient. In Category 2 images, none or just one of the OARs (bladder and rectum) is visualized. As the OARs should be spared from radiation as much as possible, in the future not only the position of the prostate should be monitored, but also the position of these organs. This would introduce the need to also make a distinction between Category 2 and 3 images. In addition, a single poor-quality 2D image does not necessarily imply that the whole volume is not able to provide useful clinical information. Therefore, the next steps should move towards the interpretation of a whole volume, for example, using recurrent neural networks which can take into account inter-slice context (e.g. (Chen et al. 2016)).

The potential of the database that was available in this work has not been fully exploited, as only 16,000 2D images of the 178,368 images were examined by the team, resulting in 13,463 images with labels. Potentially, the performance of the algorithm can be improved by using more images for training. Also, the orientation of the images that had a flipped anatomical structure configuration were manually corrected during the scoring process. However, during the actual image acquisition the probe might be held upside down as well, so the algorithm should be robust for any image orientation changes. This robustness will also be examined in future research.

Summarizing, the limitations of the presented work include the fact that the labels used for training of the algorithm were not generated by experts and that only part of the available dataset was exploited. In addition, the algorithm implementation was not memory efficient even though it focused only on 2D images instead of on whole volumes.

Finally, the focus in this work was on the use of US imaging during the RT workflow of prostate cancer patients. However, a similar approach could be adapted for use in other medical procedures in which US imaging may be beneficial for anatomical localization, but where it is not yet feasible and/or desirable to have a trained operator present at the time. These procedures could be, for example, USgRT workflows of other cancer sites (e.g. liver, bladder of cervical cancer) or US guided surgeries.

Conclusion

The purpose of this work was to propose a deep learning approach that could be used to automatically assess the quality of TPUS images of the male pelvic region. This could potentially remove the need for quality interpretation by a trained operator. The performance of the implemented one-class GP regressor was compared with three experts and the results showed that the algorithm achieves a comparable accuracy with these experts in a binary scoring scenario. Future work will involve exploring the non-binary scoring scenario, including adding additional annotated images into the database and assessing the overall quality of the TPUS volume instead of judging individual 2D images.

Acknowledgments

The computational resources and services used in this work were provided by the HPC and Research Support Group, Queensland University of Technology, Brisbane, Australia. This research forms part of a project supported by an Australia-India strategic research fund (AISRF53820). G.C. acknowledges the support received by the Australian Research Council's Discovery Projects funding scheme (project DP180103232).

References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M, Levenberg J, Monga R, Moore S, Murray DG, Steiner B, Tucker P, Vasudevan V, Warden P, Wicke M, Yu Y, Zheng X. TensorFlow: A System for Large-Scale Machine Learning. OSDI 2016. pp. 265–283.

Ballhausen H, Li M, Hegemann NS, Ganswindt U, Belka C. Intra-fraction motion of the prostate is a random walk. Phys Med Biol IOP Publishing, 2014;60:549.

Bradshaw J, Matthews AG de G, Ghahramani Z. Adversarial Examples, Uncertainty, and Transfer Testing Robustness in Gaussian Process Hybrid Deep Networks. 2017;1–33.

Camps SM, Fontanarosa D, Verhaegen F, de With PHN, Vanneste BGL. The Use of Ultrasound Imaging in the External Beam Radiotherapy Workflow of Prostate Cancer Patients. Biomed Res Int Hindawi, 2018;2018.

Chen J, Yang L, Zhang Y, Alber M, Chen DZ. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. Adv Neural Inf Process Syst 2016. pp. 3036–3044.

Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull US: American Psychological Association, 1971;76:378–382.

Fontanarosa D, van der Meer S, Bamber J, Harris E, O'Shea T, Verhaegen F. Review of ultrasound image guidance in external beam radiotherapy: I. Treatment planning and inter-fraction motion management. Phys Med Biol 2015;60.

Fraser DJ, Chen Y, Poon E, Cury FL, Falco T, Verhaegen F. Dosimetric consequences of misalignment and realignment in prostate 3DCRT using intramodality ultrasound image guidance. Med Phys Wiley Online Library, 2010;37:2787–2795.

Ghahramani Z. A Tutorial on Gaussian Processes (or why I don't use SVMs). Mlss2011CompNusEduSg 2011;

Huang G, Liu Z, Weinberger KQ, van der Maaten L. Densely connected convolutional networks. Proc IEEE Conf Comput Vis pattern Recognit 2017. p. 3.

Kemmler M, Rodner E, Wacker ES, Denzler J. One-class classification with Gaussian processes. Pattern Recognit 2013;46:3507–3518.

Khan SS, Madden MG. One-Class Classification : Taxonomy of Study and Review of Techniques. 2004.

Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2014;1–15.

Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science (80-) American Association for the Advancement of Science, 1983;220:671–680.

Lachaine, M, Falco, T. Intrafractional prostate motion management with the Clarity Autoscan system. Medical physics international, 2013; 1

Lagendijk JJW, Raaymakers BW, Raaijmakers AJE, Overweg J, Brown KJ, Kerkhof EM, van der Put RW, Hårdemark B, van Vulpen M, van der Heide UA. MRI/linac integration. Radiother Oncol Elsevier, 2008;86:25–29.

Matthews AG de G, van der Wilk M, Nickson T, Fujii K, Boukouvalas A, León-Villagrá P, Ghahramani Z, Hensman J. GPflow: A Gaussian process library using TensorFlow. J Mach Learn Res 2017;18:1–6.

McPartlin AJ, Li XA, Kershaw LE, Heide U, Kerkmeijer L, Lawton C, Mahmood U, Pos F, van As N, van Herk M, others. MRI-guided prostate adaptive radiotherapy--A systematic review. Radiother Oncol Elsevier, 2016;119:371–380.

O'Shea T, Bamber J, Fontanarosa D, van der Meer S, Verhaegen F, Harris E. Review of ultrasound image guidance in external beam radiotherapy part II: intra-fraction motion management and novel applications. Phys Med Biol 2016;61.

Rahmatullah B, Sarris I, Papageorghiou A, Noble JA. Quality control of fetal ultrasound images: Detection of abdomen anatomical landmarks using adaboost. Biomed Imaging From Nano to Macro, 2011 IEEE Int Symp 2011. pp. 6–9.

Roeske JC, Forman JD, Mesina CF, He T, Pelizzari CA, Fontenla E, Vijayakumar S, Chen GTY. Evaluation of changes in the size and location of the prostate, seminal vesicles, bladder, and rectum during a course of external beam radiation therapy. Int J Radiat Oncol 1995;33:1321–1329. Available from: http://www.sciencedirect.com/science/article/pii/0360301695002251

Schwaab J, Diez Y, Oliver A, Marti R, van Zelst J, Gubern-Mérida A, Mourri AB, Gregori J, Günther M. Automated quality assessment in three-dimensional breast ultrasound images. J Med Imaging International Society for Optics and Photonics, 2016;3:27002.

Titsias M. Variational Learning of Inducing Variables in Sparse Gaussian Processes. Aistats 2009;5:567–574.

van der Heide UA, Kotte ANTJ, Dehnad H, Hofman P, Lagenijk JJW, van Vulpen M. Analysis of fiducial marker-based position verification in the external beam radiotherapy of patients with prostate cancer. Radiother Oncol Elsevier, 2007;82:38–45.

Zhang L, Dudley NJ, Lambrou T, Allinson N, Ye X. Automatic image quality assessment and measurement of fetal head in two-dimensional ultrasound image. J Med Imaging International Society for Optics and Photonics, 2017;4:24001.

Figure captions

Figure 1. Example 2D transperineal ultrasound (TPUS) image of each quality category. (A) Category 1 with only bladder (b) identifiable; (B) Category 2 with bladder (b) and prostate (p); (C) Category 3 with bladder (b), prostate (p) and rectum (r). The dashed rectangle indicates the cropping performed during pre-processing of the images.

Figure 2. Distribution of the image data in *D0* per binary score (A) and per subject (B) in the training, validation and test set.

Figure 3. Number of training (purple) and test (yellow) images per subject distribution.

Figure 4. Receiver operating characteristics (ROC) curve (full curve and zoomed in) of the one-class CNN + GP algorithm with respect to the majority vote of the experts, where the circle, cross and upside down triangle are indicating the performance of the experts, the triangle gives the performance of the test subset and the square gives the performance of the algorithm.

Table 1. Summary of the available datasets in this study in total comprising 11,148 TPUS volumes from 36 male subjects.

Study	Subject type	# subjects	Age mean [range]	Total # volumes
Study 1	Volunteers	6	35 (range: 26 – 52)	840
Study 2	Patients	21	74 (range: 58 – 85)	1,269
Study 3	Volunteers	9	51 (range: 31 – 73)	9,039
Total	-	36	-	11,148

Table 2. Definition of three image criteria used to classify 2D TPUS images based on their quality.

Category	Criteria
Category 1	Prostate could not be identified
Category 2	Prostate alone or in combination with either a part of the bladder or the rectum could be identified
Category 3	Prostate could be identified, as well as a part of the bladder and the rectum

Table 3. Subjects in the test set of each distribution.

	Test subject numbers
D0	4, 8, 12, 19, 22, 32, 33
D1	7, 9, 13, 15, 18, 25, 31
D2	5, 13, 17, 24, 25, 26, 32
D3	7, 13, 14, 22, 23, 28, 33
D4	1, 2, 5, 18, 20, 24, 30
D5	3, 4, 5, 12, 20, 29, 34
D6	2, 15, 16, 17, 26, 28, 33
D7	5, 11, 21, 22, 26, 29, 31
D8	4, 5, 12, 13, 17, 22, 27
D9	10, 12, 14, 17, 19, 23, 29

Table 4. Algorithm parameters per implementation step with the asterisks indicating the optimized hyper-parameters.

		CNN + GP (one class)
	Parameter	**Value**
DenseNet	Number of blocks*	2
	Number of layers*	18
	Growth rate *k**	12
	Outputs*	200
GPflow	Model	Sparse GP Regression (SGPR)
	Kernel	Radial Basis Function (RBF)
	Initial kernel variance*	0.1
	Inducing points*	150
Training	Batch size*	200
	Epochs*	75
	Optimizer	Adam
	Learning rate*	1e-8
	Drop-out rate*	0.05

Table 5. Cross-validation results per deep learning approach reporting the mean and σ of the accuracy, sensitivity and specificity calculated over *D0-D8.*

	Accuracy [mean ± σ]	**Sensitivity [mean ± σ]**	**Specificity [mean ± σ]**
CNN + GP (one class)	92 ± 1.7 %	89 ± 5.8 %	93 ± 2.6 %
CNN + GP (two class)	92 ± 1.5 %	90 ± 6.1 %	93 ± 2.2 %
CNN + Softmax	90 ± 1.3 %	79 ± 7.6 %	95 ± 2.5 %

Table 6. Accuracy, sensitivity and specificity (1-false positive rate) results for the algorithms, test subset and three experts calculated with respect to the majority vote of the three experts.

	CNN + GP One class	**CNN + GP Two class**	**CNN + Softmax**	**Test subset**	**Expert 1**	**Expert 2**	**Expert 3**
Accuracy	94%	93%	92%	91%	96%	97%	92%
Sensitivity	92%	96%	87%	80%	99%	95%	90%
Specificity	95%	91%	96%	99%	93%	99%	94%

Supplementary materials A

Table 1. Parameters of the additional deep learning algorithms per implementation step with the asterisks indicating the optimized hyper-parameters.

		CNN + GP (two class)	CNN + Softmax
	Parameter	**Value**	**Value**
DenseNet	Number of blocks*	2	3
	Number of layers*	18	18
	Growth rate *k**	12	12
	Outputs*	200	-
GPflow	Model	SGPR	-
	Kernel	RBF	-
	Initial kernel variance*	0.2	-
	Inducing points*	150	-
Training	Batch size*	200	50
	Epochs*	75	75
	Optimizer	Adam	Adam
	Learning rate*	1e-8	1e-6
	Drop-out rate*	0.05	0.05

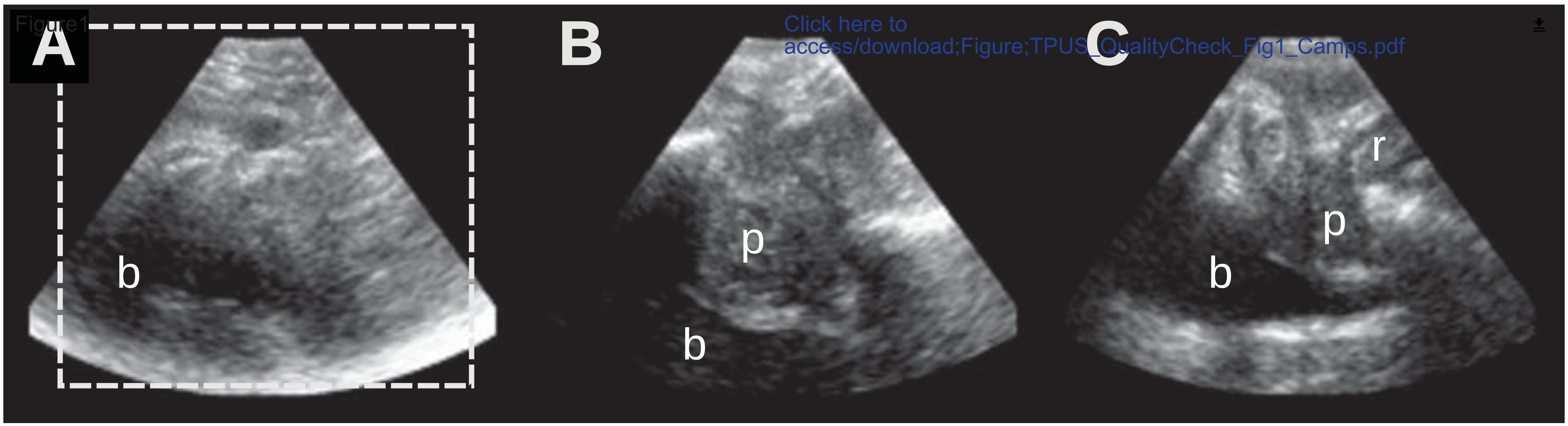
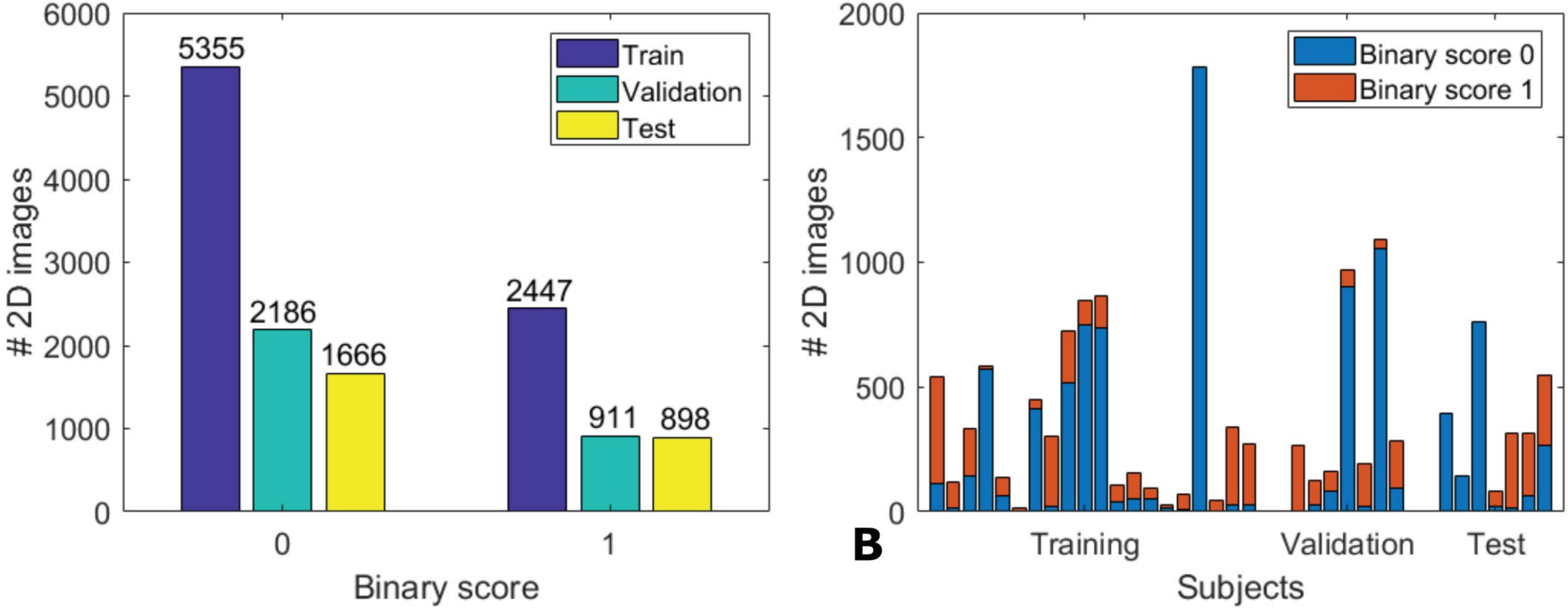


Figure1
A
b
B
p
b
C
r
p
b

Figure2



2D images
6000
5000
4000
3000
2000
1000
0
5355
2186
1666
2447
911
898
Train
Validation
Test
0
1
Binary score
B
2D images
2000
1500
1000
500
0
Binary score 0
Binary score 1
Training
Validation
Test
Subjects

Figure3

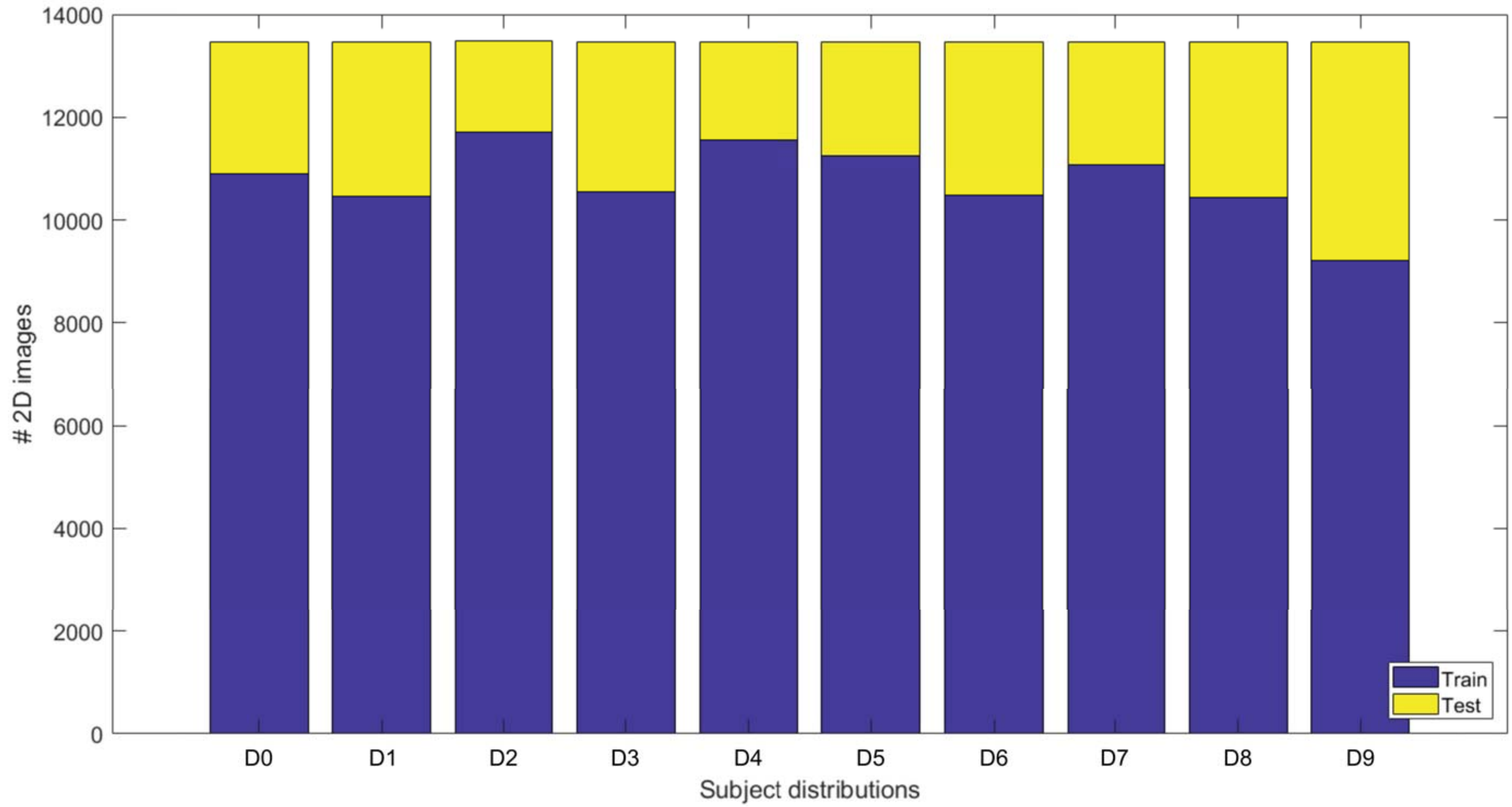
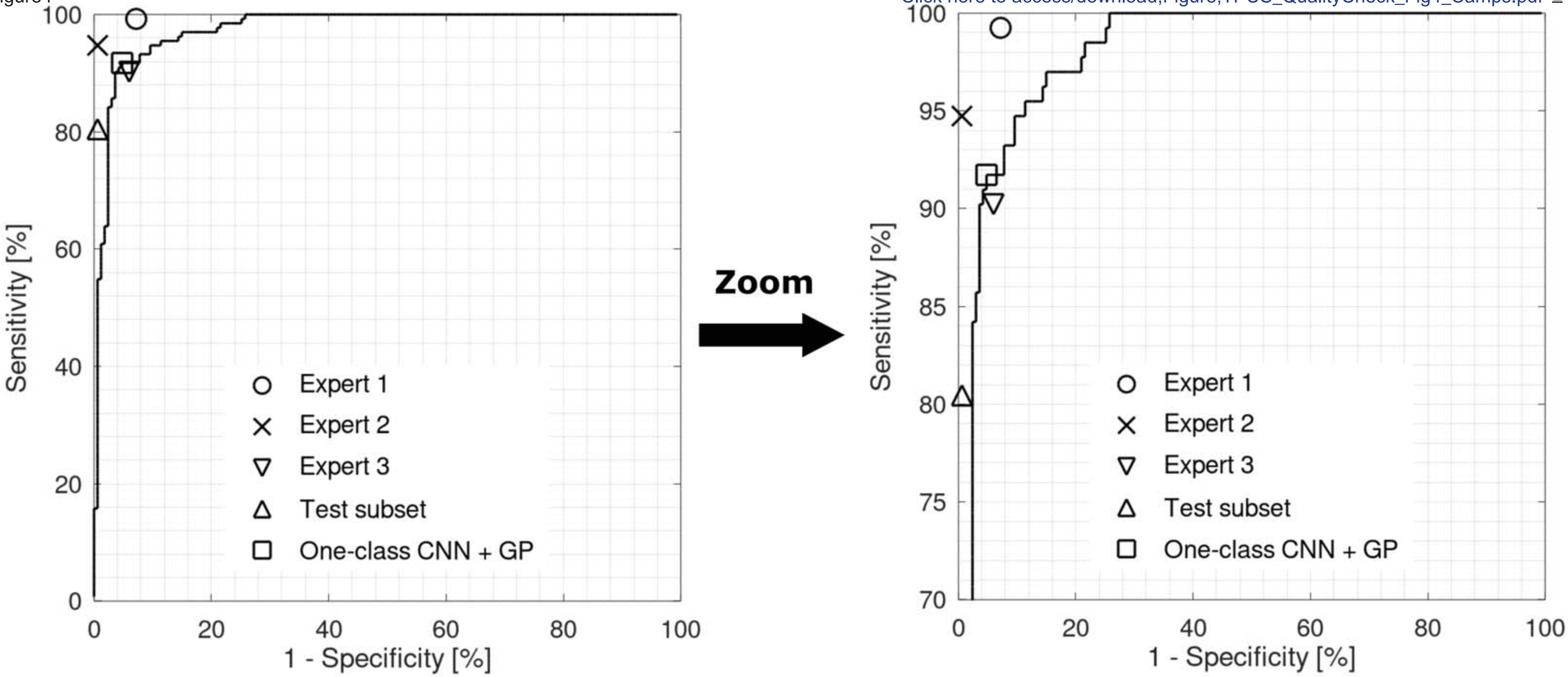

14000
12000
10000
8000
6000
4000
2000
0
2D images
D0
D1
D2
D3
D4
D5
D6
D7
D8
D9
Subject distributions
Train
Test

Figure4



Sensitivity [%]
1 - Specificity [%]
Expert 1
Expert 2
Expert 3
Test subset
One-class CNN + GP
Zoom
Sensitivity [%]
1 - Specificity [%]
Expert 1
Expert 2
Expert 3
Test subset
One-class CNN + GP