Received February 18, 2019, accepted March 17, 2019, date of publication April 1, 2019, date of current version April 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2908002

Boundary Delineation of MRI Images for Lumbar Spinal Stenosis Detection Through Semantic Segmentation Using Deep Neural Networks

ALA S. AL-KAFRI[1], **SUD SUDIRMAN**[1], **ABIR HUSSAIN**[1], **DHIYA AL-JUMEILY**[1], **FRISKA NATALIA**[2], **HIRA MEIDIA**[2], **NUNIK AFRILIANA**[2], **WASFI AL-RASHDAN**[3], **MOHAMMAD BASHTAWI**[3], **AND MOHAMMED AL-JUMAILY**[4]

[1]Department of Computer Science, Liverpool John Moores University, Liverpool L3 3AF, U.K.
[2]Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Tangerang 15811, Indonesia
[3]Irbid Specialty Hospital, Al-Hassan Sport City, Irbid 21110, Jordan
[4]Dr. Sulaiman Al-Habib Hospital, Dubai, UAE

Corresponding author: Sud Sudirman (s.sudirman@ljmu.ac.uk)

This work was supported in part by the Indonesian Ministry of Research, Technology and Higher Education, under Grant 034/KM/PNT/2018.

ABSTRACT We propose a methodology to aid clinicians in performing lumbar spinal stenosis detection through semantic segmentation and delineation of magnetic resonance imaging (MRI) scans of the lumbar spine using deep learning. Our dataset contains MRI studies of 515 patients with symptomatic back pains. Each study is annotated by expert radiologists with notes regarding the observed characteristics and condition of the lumbar spine. We have developed a ground truth dataset, containing image labels of four important regions in the lumbar spine, to be used as the training and test images to develop classification models for segmentation. We developed two novel metrics, namely confidence, and consistency, to assess the quality of the ground truth dataset through a derivation of the Jaccard Index. We experimented with semantic segmentation of our dataset using SegNet. Our evaluation of the segmentation and the delineation results show that our proposed methodology produces a very good performance as measured by several contour-based and region-based metrics. In addition, using the Cohen's kappa and frequency-weighted confidence metrics, we can show that 1) the model's performance is within the range of the worst and the best manual labeling results and 2) the ground-truth dataset has an excellent inter-rater agreement score. We also presented two representative delineation results of the worst and best segmentation based on their BF-score to show visually how accurate and suitable the results are for computer-aided-diagnosis purposes.

INDEX TERMS Lumbar spine MRI, lumbar spinal stenosis, semantic segmentation, boundary delineation, deep learning.

I. INTRODUCTION

Lumbar Spinal Stenosis (LSS) is one of the leading causes of chronic lower back pain. It is a narrowing of lumbar spinal canal caused by inflammation of bone or soft tissues, which in turn produces pressure on spinal nerve roots. Patients will experience symptoms ranging from radicular pain to atypical leg pain to neurogenic claudication [1]. Chronic lower back pain is suffered by millions of people around the world. It is a debilitating illness that is affecting the health, social life, and employment of its sufferers.

The diagnosis of LSS is usually performed through inspection of Magnetic Resonance Imaging (MRI) scan of the patient's lumbar spine. MRI images can be used to visualize lumbar spine, slice by slice, in three view-planes namely sagittal (side), axial (top-down) and coronal (frontal) – though only the first two are typically used in lumbar spine MRI studies. Most MRI scans are performed when the patients are in supine position. Many practitioners argue that this approach has several shortcomings because the images do not often reflect the condition of the spines when the

The associate editor coordinating the review of this manuscript and approving it for publication was Yudong Zhang.

patients' body weight is bearing on them. Subsequently, there have been a number of studies that suggest improvements in the way the scans are carried out which include taking the scans while the patient is in an upright position [2] and using Lumbar Pillow [3]. Other studies propose to use different radiology techniques such as Radiographic Myelography [4] to detect the most difficult cases of LSS. While we acknowledge that these approaches can serve as better alternatives to using supine MRIs in some cases, we also argue that supine MRIs remain, in the majority of cases, the most practical way to study lumbar spine. Our argument lies on a simple fact that the equipment to carry out supine MRI is much more ubiquitous in hospitals than the others. Furthermore, scans taken in a standing or sitting up position are most likely affected by patients' movement due to discomfort and can result in bad quality images. Because our study relies on gathering and analysis a large amount of data, we therefore focus on using supine MRI scans than other types of scans. For brevity purpose, we refer them simply as MRI in the remainder of the paper.

In this paper, we are proposing a method to aid clinicians in performing lumbar spinal stenosis detection through semantic segmentation and delineation of important boundaries in axial-view MRI images. We will first provide a concise review of related methodologies in the literature before presenting the rationale of our proposed approach.

There is a wide range of algorithms for computer aided medical diagnosis depending on the type of disease they are targeting. These algorithms are often specific to a certain type of medical images, such as CT scan, X-Ray, intravascular ultrasound, or MRI. This is the case because certain types of images can capture features of certain diseases better than others. Case in point, the detection of lumen and media-adventitia (MA) borders is the key procedure to determine anomalies inside coronary arteries. This procedure most often is carried out using intravascular ultrasound (IVUS) images than any other types of images because it has been well recognized as one powerful imaging technique to evaluate the stenosis inside the coronary arteries. One of the most recent frameworks for detecting these borders was proposed by Gao *et al.* [5]. In [6], the authors then apply this framework by using an Artificial Neural Network (ANN) method as the feature learning algorithm for detecting the borders. In this method, two types of imaging information, including spatial-neighboring features, were used as the input data to one ANN that has two sparse auto-encoders as a classifier. Another ANN was used as an optimizer. The paper presented a comparison between this method's performance and the manual drawing approach on 461 IVUS images and reported a high accuracy result. A variant of the method using three types of features have also been proposed in [7] using 538 IVUS images.

Intravascular ultrasound images are not, however, the only possible source of information for detecting vascular anomalies. Recently, a regression segmentation framework to delineate boundaries of bi-ventricle from cardiac MRI was proposed [8]. In this framework, a regression model has been trained automatically on a deep belief network by using extracted DAISY feature [9] as input, and using automatically generated boundary points as labels. The method was reported to yield high performance when tested on 2,900 images taken from 145 clinical subjects.

MRI scans and Neural Networks have also been used to diagnose other illnesses. A recent study [10] shows that a Convolutional Neural Network (CNN) can be used to reliably (with a reported accuracy of 98.8% when data augmentation and stochastic pooling are applied) identify Multiple Sclerosis in brain MRI scans. The experiment was conducted on images from an open source dataset [11] (containing 676 Multiple Sclerosis slices from 38 subjects) and another [12] (containing 681 healthy slices from 26 subjects) as a control dataset.

The previous methods are some of the most recent examples on how image segmentation can be used for border detection and delineation in medical images. Specific to detecting diseases on lumbar spine, Jiang *et al.* [13] proposed a visualization and quantitative analysis framework using image segmentation technique to derive six features that are extracted from patients' MRI images, which have a close relationship with lumbar disc herniation score. These features include the distribution of the protruded disc, the ratio between the protruded part and the dural sacs, and its relative signal intensity.

A different approach proposed by Alomari *et al.* [14] [15] uses a probabilistic model for automatic disc herniation detection by combining the appearance and shape features of the lumbar intervertebral discs. The technique models the shape depending on both the T1-weighted and T2-weighted co-registered sagittal views for building a 2D feature image.

A more recent and relevant technique to our work is an unsupervised neural foramina boundary delineation framework by He et al. [16]. This framework uses Multi-Feature and Adaptive Spectral Segmentation (MFASS) algorithm to automate the delineation process of neural foramina in mid-sagittal view of a lumbar spine. MFASS utilizes a combination of region and edge features to generate spectral features that can be used to separate neural foramina and its surroundings. The separation process is controlled by adjusting the separation threshold, which is optimally and automatically estimated for each individual image. The framework is tested on 280 neural foramina MR images from 56 clinical subjects. The results are compared with manual boundary delineation performed by experienced physicians and analyzed using two metrics namely the Dice metric (a variant of the intersection-over-union metric) to measure segmentation overlap, and the Symmetric Mean Absolute surface Distance (SMAD) to evaluate the delineation accuracy [17]. The paper reported a high consistency with manual delineation results (Dice: 90.58% ±2.79%; SMAD: 0.5657 ±0.1544 mm).

We deduce from analyzing the different relevant algorithms in the literature to-date that boundary delineation through image segmentation is a very popular yet effective approach to medical image analysis. Nonetheless, we have identified three main problems in the current state-of-the-art research in this field. The first one is the size of the freely available dataset. The most comprehensive database of lumbar-spine-related medical images is hosted by SpineWeb [18], however, it only contains relatively small-sized and incomplete datasets taken from between 8 to 125 patients. To address this problem, we work together with one specialty hospital in Jordan and several physicians and radiologists around the world to gather a significant number of relevant MRI scans complete with medical annotation to develop our dataset. We made this dataset freely available to the research community.

The second problem is the lack of ground truth data and the means to assess its quality. Since the ground truth data is tied to the dataset, the argument of needing a suitably large ground truth data also applies. Furthermore, because the task of developing ground truth data by manually labelling the MRI images is a laborious one, it becomes prone to errors. In other words, the data can be inaccurate and inconsistent. We address the problem of how to measure accuracy and variability in ground truth data by developing two new metrics which are derived from the existing machine learning metrics.

And lastly, despite the rapid advances in machine learning techniques, to the best of our knowledge, there is a limited study on their application to lumbar spine image segmentation. In this paper, we show how the new advances in deep learning can be used to perform semantic segmentation of lumbar spine MRI which can subsequently be used for lumbar spinal stenosis detection.

II. MRI DATASET AND GROUND TRUTH DATA

A. LUMBAR SPINE MRI DATASET

All procedures performed in this study are in accordance with the ethical standards of both the United Kingdom and the Kingdom of Jordan and comply with the 1964 Helsinki declaration and its later amendments. The approval was granted by the Medical Ethical Committee of Irbid Speciality Hospital in Jordan. The data was collected between September 2015 and July 2016 from patients who attended the hospital who reported relevant pains. Written formal consent was obtained from each patient prior to the data collection. A personal-data cleaning process was applied to the collected data to remove any information that can be used to relate it to any patient such as the patient's name, date of birth, and date of visit. We assign each patient data with a unique identification number and only refer to each data using its assigned number in all subsequent processes. This allows the data to be accessed anonymously as stipulated by the ethical committee's condition of approval.

We collected clinical MRI studies, or a set of scans, of 575 patients with symptomatic back pains. The MRI scanning parameters used in the scans can vary depending on the sequence and view plane types. The values of the most important parameters are summarized in Table 1.

TABLE 1. MRI scanning parameters.

View Plane Types	Sagittal		Axial	
Sequence Types	T1-weighted	T2-weighted	T1-weighted	T2-weighted
Number of Echoes (ETL)	3	15 to 18	3	9 to 16
Repetition Time (ms)	330 to 926	3190 to 4000	385 to 953	1900 to 5000
Echo Time (ms)	9.2 to 12.0	67.0 to 96.0	11.0	84.0 to 96.0
Slice Thickness (mm)	3.0 to 4.0	3.0 to 5.0	4.0	3.0 to 5.0
Spacing Between Slices (mm)	3.3 to 4.8	3.3 to 6.5	4.4 to 4.4	3.3 to 6.5
Field of View (mm)	280	280	220	220
Matrix (Freq. x Phase)	100%	100%	100%	100%
Imaging Frequency (MHz)	63.6765 to 63.6828	63.6765 to 63.6828	63.6801 to 63.6828	63.6801 to 63.6828
Number of Phase Encoding Steps	288 to 540	408 to 544	295 to 336	272 to 360
Flip Angle	150	150	150	150

During the data selection stage, we had to remove 60 of the studies that do not meet the requirements set by the proposed methodology we are describing in this paper. The requirements that need to be satisfied for each MRI study are as follows:

1. The study needs to include at least the last three lumbar vertebra and their adjacent posterior elements, the last three intervertebral discs (IVD), and the topmost sacral bones.

2. The study must contain both T1-weighted and T-2 weighted scans. An image registration algorithm should be able to align both scans within a reasonable search space and duration. This means the patients should not have made any significant movement during the scanning procedure as to make the two scans completely different.

3. The study needs to have both sagittal and axial view scans and their corresponding cross-view information. This should allow us to see the direction and position of the image plane of an axial view slice on the sagittal view, and vice versa in a DICOM viewer application.

4. The study needs to have at least one axial view slice close to the center of each intervertebral disc.

5. The study should be of reasonably good quality in term of focus, sharpness and distortion.

6. The study should not contain destroyed or fused lumbar spine elements as to make manual region labelling difficult or impossible.

7. The study should be from adult patients (with minimum age of 17) to ensure common physiology of the lumbar spine throughout the dataset.

At the end of the data selection process we have the MRI study of 515 patients. To help us maintain consistency in the data numbering for all processes, we decided to keep the identification number assignment for each patient data the same as opposed to rearranging and reassigning them every time a data is removed from the dataset. This approach will also be adopted in future should we be required to remove more data. If this need arises, we will simply remove the data from the dataset but keep the unused numbers intact in the record.

Each study of the remaining 515 patients in our dataset is annotated by expert radiologists with notes regarding the observed characteristics, condition of the lumbar spine, or presence of diseases which include bone marrow disease, end plate degeneration, IVD bulges, Thecal Sac (TS) compressing, central or foraminal stenosis, annular tears, scoliosis, endplate defects (Modic type), facet joint and Ligamentum Flavum hypertrophy, and spondylolisthesis.

Each patient data can have one or more MRI studies associated with it. Each study contains slices, i.e., individual images taken from either sagittal or axial view, of the lowest three vertebrae and the lowest three IVDs. The axial view slices are mainly taken from these last three IVDs – including the one between the last vertebrae and the sacrum. The orientation of the slices of the last IVD are made to follow the spine curve whereas those of the other IVDs are usually made in blocks – i.e., parallel to each other. There are between four to five slices per IVD and they begin from the top of the IVD towards its bottom. Many of the top and bottom slices cut through the vertebrae, leaving between one to three slices that cut the IVD cleanly and show purely the image of that IVD. In most cases, the total number of slices in axial view ranges from 12 to 15. However, in some cases, there may be up to 20 slices because the study contains slices of more than the last three vertebrae. The scans in sagittal view also vary but they all contain at least the last seven vertebrae and the sacrum. While the number of vertebrae varies, each scan always includes the first two sacral links.

There are a total 48,345 MRI slices in our dataset. The majority of the slices have an image resolution of 320 × 320 pixels, however, there are slices from three studies with 320 × 310 pixel resolution. The pixels in all slices have 12-bit per pixel precision which is higher than the standard 8-bit greyscale images. Specifically for all axial-view slices, the slice thickness are uniformly 4 mm with center-to-center distance between adjacent slices to be 4.4 mm. The horizontal and vertical pixel spacing is 0.6875 mm uniformly across all axial-view slices.

The majority of the MRI studies were taken when the patients were in Head-First-Supine position with the rests were taken when they were in Feet-First-Supine position. Each study can last between 15 to 45 minutes and a patient may have one or more studies taken at different times on the same day, or a few days apart. Because of the requirement of the method we are employing, we only select studies that contain both T1- and T2-weighted images in both sagittal and axial views. The difference in acquisition time between T1- and T2-weighted scans ranges between 1 to 9 minutes. Long time difference could suggest that corresponding slices may not necessarily align and may require an application of an image registration algorithm to align them. As before, due to the requirement of the method we are employing, we removed any studies where the difference in T1- and T2-weighted scans causes the image registration process to fail or produce large number of mismatched pixels. To provide our reader with a better picture on the unsuitability of the images that we discarded, we show several examples of them in Fig. 1.

B. IMAGE LABELS AND GROUND-TRUTH DATA

The development of ground truth data for machine learning depends largely on the application. In our case, the application is automatic segmentation of MRI images for lumbar spinal stenosis detection. In this section, we will provide the rationale for the design decisions that we took during the development of our ground truth data. These decisions are based on the advice from, and the experience of, several radiologists when performing manual detection of lumbar spinal stenosis.

The detection of LSS on axial-view MRI is performed through measuring the distance between the Posterior Element (PE) and the IVD at different locations of the lumbar spine as illustrated in Fig. 2. This process requires accurate delineation of boundaries between the different regions of the lumbar spine image including the region between PE and IVD. When observed from sagittal view, this region extends from the cervical spine down to the lumbar spine. Because of a lack of a suitable medical terminology that describes this area, in this paper, we will refer to this area as AAP, which is short for *Area between Anterior and Posterior* vertebrae elements.

We have demonstrated previously in [19] that an accurate and consistent delineation of these boundaries cannot be performed just through an application of an edge detection algorithm directly on the MRI image. Instead, the image needs to be segmented beforehand. However, medical image segmentation itself possesses its own unique challenges. One of the major difficulties in medical image segmentation is the high variability in medical images which is caused by the variability in human anatomy itself, the severity of the illness, the effect of age and gender, and also the intrinsic factors of the equipment such as calibration and sensitivity. To overcome these difficulties, we use a deep neural network to perform the segmentation because of the technique's widely acknowledged ability to take into account these variabilities [20]. The ground truth data used for training and testing of the deep learning algorithm consists of labelled axial-view MRI slices of the IVDs. It is important to note that we do not use the slices of all five lumbar IVDs, but instead, we use the slice of the last three only. The rationale of this was provided in our previous work [19].

The label images in the ground truth data mark several regions of interest (RoIs). Since lumbar spinal stenosis occurs inside AAP, we only focus on parts of the MRI which contain

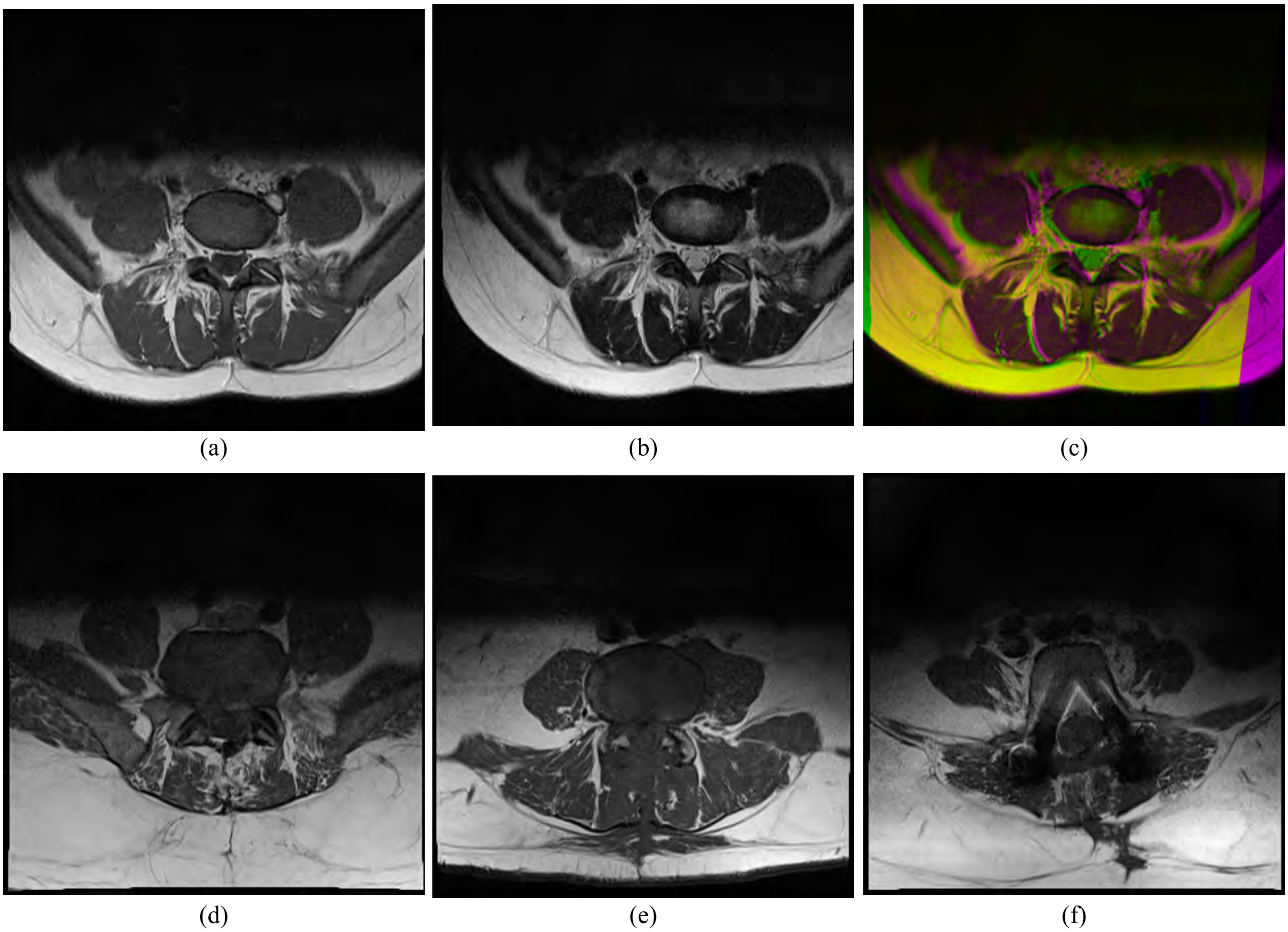


FIGURE 1. Some examples of the discarded MRI scans. Image (a) is T1-weighted and (b) T2-weighted MRI scans of the same part of lumbar spine of a patient. There is a significant difference in the position of the same organ in both images resulting in large number of mismatch pixels (purple regions) in the resulting (c) composite image. Image (d) and (e) are example cases where multiple regions are fused together making manual segmentation unreliable. And (f) is an example of a scan that contains unexplained imaging distortion or artefact.

and are around this region. Subsequently, we decided to have four RoIs which are: a) the IVD, b) the PE, c) the TS, and d) the AAP itself as illustrated in Fig. 2. Any other pixels that do not belong to any one of the above four regions are labelled as e) 'Other'.

The labelling process is carried out using the T1-weighted axial-view MRI slice of the last three IVDs. The task of manually labelling the four areas on each MRI slice is a laborious one. On average, five to ten minutes are spent to label each of the 1,545 slices. We employ the strategy we have designed previously in [21] to develop our ground truth dataset using five participant/labelers. We analyze the quality of the developed ground-truth dataset using the confidence and consistency metrics that we presented in that paper. The detail of these metrics is provided next.

C. GROUND TRUTH DATA QUALITY METRICS

Compared to other topics in computer vision, very few formal or analytic work has been published to guide the creation of ground truth data. There is some guidance [22], [23] provided by machine learning community for measuring the quality of ground truth data used for training and test datasets, but this tends to revolve only around the size of the dataset [24]. To address this problem, we propose a novel approach to assess ground truth quality not from the size of the dataset but through calculating its confidence and consistency levels to measure its accuracy and variability, respectively.

We define the confidence level of ground truth data as a sureness measure that all labelled regions contain all the pixels that should be in that class and nothing less. On the other hand, we define consistency level as how varied the ground truth data is across its sources. To measure the confidence level of the resulting labelled images we use a variant of the Jaccard Index, which is also known as the intersection-over-union metric [25]. The intersection-over-union (iou_c) of class c is calculated as the ratio between the number of correctly predicted pixels (intersection) and the sum (union) of the number of correctly and incorrectly predicted pixels.

$$iou_c = \frac{m_{cc}}{t_c + m_c - m_{cc}} \tag{1}$$

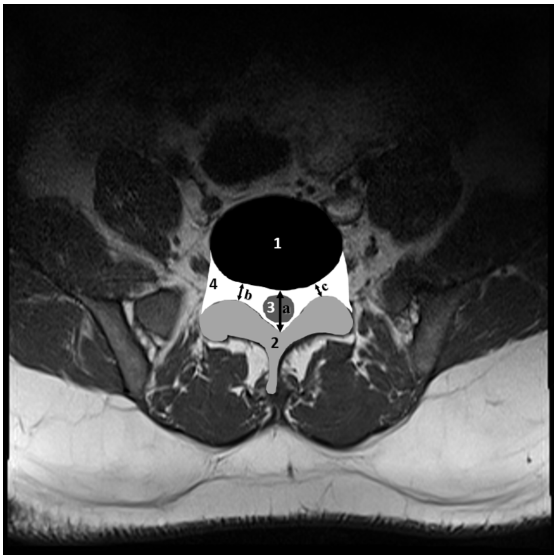


FIGURE 2. The four labelled Regions of Interest namely 1) Intervertebral Disc (IVD), 2) Posterior Element (PE), 3) Thecal Sac (TS) and 4) the Area between Anterior and Posterior (AAP) vertebrae elements. Delineation of boundaries between these regions is used to measure a) the anteroposterior diameter of the spinal canal and b) the left and c) the right foramen widths.

where m_{cc} is the number of pixels of class c correctly predicted to belong to class c, and t_c is the total number of pixels of class c – according to the ground truth, and m_c is the total number of pixels predicted to belong to class c. An ideal classifier would correctly classify all pixels that belong to class c as that class (i.e., $m_{cc} = t_c$) and only those pixels (i.e., $m_{cc} = m_c$), resulting in $iou_c = 1$. Since in this case we do not have, and are still developing, the ground truth data, the value of both m_{cc} and t_c cannot be determined. Hence, we will develop an alternative intersection-over-union metric, denoted as iou'_{cv} to estimate iou_c, which will be used as a measure of confidence of the ground truth data.

Consider a set C, defined as $C = \{1, 2, 3, 4\}$, of the four classes or RoIs. A pixel n can be labelled as l_{np}, where $l_{np} \in C$, by a participant labeller p, where, since in our case we use five participant labelers, $p \in \{1, 2, 3, 4, 5\}$. We define a vote count, k_{nc}, as the number of votes from all five participants that assign class c to pixel n, where $c \in C$.

$$k_{nc} = \sum_{p} [l_{np} = c] \tag{2}$$

where $[z]$ is the Iverson Bracket notation of logic proposition z, i.e., $[z] = 1$ if z is true or 0 otherwise. The vote count has values in the range of $0 \leq k_{nc} \leq 5$, e.g., $k_{n2} = 0$ means the pixel n receives zero vote that assigns class 2 to it and $k_{n3} = 5$ means the pixel n receives all five votes that assign class 3 to it.

Next, we define the intersection of c-labelled regions, denoted as s_{cv}, as the normalized number of pixels that receive at least v number of votes that assign class c. Here, we refer v as the vote-threshold which values are integers between 1 to the maximum number of votes, i.e., 5 in our case.

TABLE 2. Confidence and consistency values of the resulting ground truth data.

Regions	x_c	y_c
Intervertebral Disc	0.93	0.95
Posterior Element	0.82	0.87
Thecal Sac	0.81	0.87
The AAP	0.48	0.68

$$s_{cv} = \frac{1}{n} \sum_{n} [k_{nc} \geq v] \tag{3}$$

Note that s_{c1} is the normalized number of pixels that receive at least one vote that assigns class c. This represents the union of all pixels receiving a non-zero number of votes for that class. Therefore, these pixels will serve as the union, or denominator, in our alternative intersection-over-union iou'_{cv} calculation. Another important fact to consider is that for $\forall c$, the following composite inequality applies:

$$s_{c1} \geq s_{c2} \geq s_{c3} \geq s_{c4} \geq s_{c5} \tag{4}$$

Based on the above argument, we define our alternative intersection-over-union metric iou'_{cv} of class c and vote threshold v as,

$$iou'_{cv} = \frac{s_{cv}}{s_{c1}} \tag{5}$$

Substituting the equation to the above inequalities we have the following relationship:

$$1 \geq iou'_{c2} \geq iou'_{c3} \geq iou'_{c4} \geq iou'_{c5} \tag{6}$$

Hence, the closer the value of iou'_{cv} is to unity for all vote thresholds the better in-agreement the five participants are in labelling the region of class c.

The confidence metric x_c, as a sureness measure that all labelled regions contain all the pixels that should be in that class and nothing less, is defined as iou'_{cv} at the selected vote-threshold v_t.

$$x_c = iou'_{cv_t} \tag{7}$$

The consistency metric, that measures how varied the ground truth data is across its sources, is defined as the rate of change of iou'_{cv} along the vote threshold dimension.

$$y_c = 1 + 2 \times \frac{iou'_{cv_{t+1}} - iou'_{cv_{t-1}}}{v_{t+1} - v_{t-1}} \tag{8}$$

Note that the value of y_c ranges between 0 and 1, where low value suggests low consistency and high variability between the labelers, and vice versa.

Using $v_t = 3$ as our chosen vote threshold, the values of metrics of our ground truth data are shown in Table 2.

The results are consistent with the previous experiment we reported in [21]. The IVD has the highest confidence and consistency values compared to the other three. This is because it is by far the easiest region to label. The region has more consistent visual characteristics across all patients which manifest as a narrow range of pixel grey level values, smoother texture, as well as a high contrast to the surrounding tissues.

The PE and TS regions have almost identical metric values which are lower than the IVD but higher than the AAP. The latter region is the hardest region to label because it does not strictly represent any part of human tissues like the other three, but instead it represents a large osseous opening in lumbar spine structure [26]. Its shape can vary significantly depending on many factors such as the location of the slice, the patient's posture as the MRI is performed, as well as the presence of illness or defects.

D. COMPOSITING T1- AND T2-WEIGHTED MRI IMAGES

The labelling of the MRI slices used T1-weighted images because it provides us with the ability to identify and locate the TS region. However, the information in a T2-weighted image is also as important as the information contained in its corresponding T1-weighted image. The combination of the two will provide better and richer discriminating features when carrying out the segmentation process.

Despite the fact that both T1-weighted and T2-weighted images have identical dimension, not all pixels at the same location in both images correspond to the same point in an organ or tissue. We have observed a wide gap, typically between 1 to 9 minutes, between the time data recorded on T1-weighted and T2-weighted MRI scans. A large time difference suggests that corresponding T1-weighted and T2-weighted MRI slices may not necessarily align.

The process to align the two images begins by fixing one of the two images and transform the other to match the first image. We set the T1-weighted image as the fixed image because we used it when constructing our image labels. A set of affine transforms, i.e., a linear combination of translation, rotation, scaling, and shearing, are then applied to the T2-weighted image to produce transformed images and calculate the difference between them and the fixed image. The whole process is known as image registration, which is essentially an error-minimization problem over a search-space. In our experiment, we set the minimum and maximum limit to the radius of this search-space to 1.5e-6 and 13e-3, respectively with a growth factor of 1.05. To avoid a long or an indefinite search time, we limit the number of searches to 300 iterations. It is also expected that both modalities are affected by both high-frequency noise and low-frequency inhomogeneity field. To counter the latter, a parametric bias field estimation is applied before being corrected using PABIC method [27]. This method also employs a search optimization algorithm called (1+1)-Evolutionary Strategy that locally adjusts the search direction and step size while at the same time provides a mechanism to step out of non-optimal local minima.

In the event of the registration process fails to converge, we perform a manual inspection on the images. It is very likely that it is a result of large discrepancies between the two modalities which may be caused by significant movement by the patient during the MRI scan. If that is the case, we will treat the images as unusable and remove the pair from our dataset.

Once the image registration is completed, a composite 3-channel image is created from the T1-weighted and T2-weighted slices. The first channel is the original T1-weighted image whereas the second channel is constructed from the aligned T2-image. The third channel can be either a linear or non-linear combination of the two. We have experimented with a number of different functions, including Manhattan distance, Euclidean distance, Mahalanobis distance, as well as polynomials distance but settled with the simplest one which is the Manhattan distance of the images.

The registration process may also produce a set of pixels locations where T1-weighted pixels have no correspondence with any of the transformed T2-weighted pixels. To accommodate the classification of these pixels, we create a new class of pixel label in addition to the five classes we already have. We refer to this new set of pixels as 'Unregistered'. Next, we use these images for training and testing a Convolutional Neural Network (CNN) to perform automatic semantic segmentation on them.

III. BOUNDARY DELINEATION THROUGH SEMANTIC SEGMENTATION

A. SEMANTIC SEGMENTATION USING CNN

In this section, we evaluate the applicability of deep learning using CNN in performing semantic segmentation on our dataset. Prior to the advent of deep learning in computer vision, image segmentation is traditionally performed using clustering techniques such as k-means clustering [28], nearest-neighbor [29], and support vector machine [30]. With the recent increase in popularity of deep learning in image classification, so does its use in performing pixel-wise classification. This gives rise to a special type of image segmentation, namely semantic segmentation [25], which surpasses other approaches by a large margin in terms of efficiency and performance accuracy.

There are a number of proposed solutions to semantic segmentation including Fully Convolutional Networks [25], SegNet [31], DeepLab [32], and RefineNet [33]. Analysis and comparison of these solutions have already been extensively provided in the literature and are beyond the scope of this paper. Instead, we will focus on one of the most popular algorithms, namely SegNet, in carrying out the semantic segmentation of our dataset.

SegNet consists of a series convolutional layers arranged in an encoder-decoder architecture. The architecture of SegNet's encoder network is topologically identical to the first 13 convolutional layers of the VGG16 network [34]. The input image is passed on to the first layer of the encoder that performs convolution with a trainable filter bank to produce the first set of feature maps. These feature maps are then batch normalized [35] before an element-wise Rectified Linear Unit (ReLU) function [36] is applied to them. The resulting

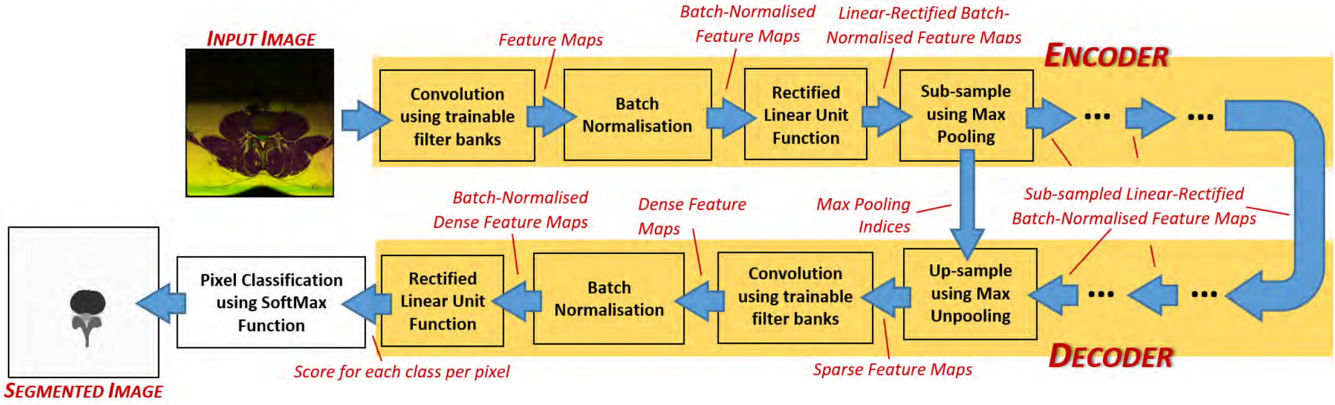

INPUT IMAGE
Feature Maps
Batch-Normalised Feature Maps
Linear-Rectified Batch-Normalised Feature Maps
ENCODER
Convolution using trainable filter banks
Batch Normalisation
Rectified Linear Unit Function
Sub-sample using Max Pooling
...
...
Batch-Normalised Dense Feature Maps
Dense Feature Maps
Max Pooling Indices
Sub-sampled Linear-Rectified Batch-Normalised Feature Maps
Pixel Classification using SoftMax Function
Rectified Linear Unit Function
Batch Normalisation
Convolution using trainable filter banks
Up-sample using Max Unpooling
...
...
SEGMENTED IMAGE
Score for each class per pixel
Sparse Feature Maps
DECODER

FIGURE 3. The processes and information flow of SegNet's encoder-decoder architecture.

signal is then applied to a max-pooling function with a 2x2 window and stride 2, non-overlapping window, before it is sub-sampled by a factor of 2. The result is then fed to the next set of convolution layer, batch-normalization layer, ReLU layer, max pooling and sub-sampling layer, and so on, until the end of the encoder section. The purpose of the encoder layers is to capture image features at varying resolutions. It is also intended to achieve translation invariance over the various size of spatial image context at the cost of increasingly lossy, or inaccurate, boundary detail in its final output signal.

The process to restore this detail starts by up-sampling that output signal using max unpooling layer in the decoder. The decoder will need the memorized indices of the corresponding max pooling process that was carried out previously in the encoder, to create sparse feature maps from the lower resolution input signal. These sparse feature maps are then convolved with a trainable decoder filter bank to produce dense feature maps. Similar to the step in the encoder, a batch normalization step is applied to the dense feature maps before applying the ReLU function. The result is then fed to the next set of max unpooling and up-sampling layer, convolution layer, batch-normalization layer, and ReLU layer and so on until the end of the decoder section. The role of the decoder is to map the low resolution, sub-sampled, feature maps to full resolution and dense feature maps, which will then are pixel-wise classified using SoftMax function [37]. The processes and information flow of the first encoder and last decoder of SegNet is illustrated in Fig. 3.

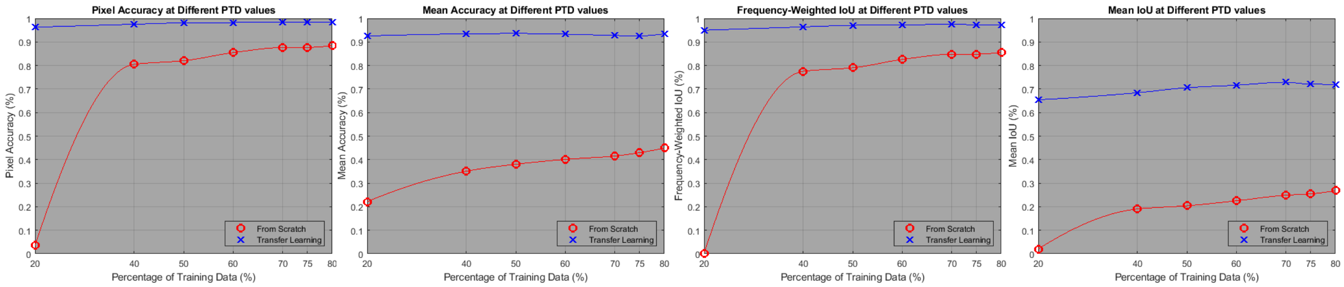
The training of SegNet adjusts the value of filter coefficients in all convolution layers as to minimize the loss function between the resulting predicted segmentation and the ground truth. In this paper, we compare the result of the segmentation of our dataset using two versions of SegNet. The first SegNet has pre-trained VGG16 coefficients which had been trained using more than a million images from the ImageNet database [38]. Since the type and number of classes in the ImageNet database do not match the type and number of classes in our case, we replace the last classification layer and retrained the SegNet with our dataset. This technique is commonly known as *Transfer Learning* hence consequently, we refer to this network as SegNet-TL. The second SegNet has an identical architecture as SegNet-TL. However, the initial values of weight and bias of the convolution layers were set using a uniform distribution random number generator between -1 and 1. This allows us to train the SegNet from scratch, and as a result, we refer to this network as SegNet-FS.

B. SEGNET TRAINING SETUP

When training both types of SegNet, we experimented using different ratios of training and testing images, namely 20:80, 40:60, 50:50, 60:40, 70:30, 75:25, and 80:20. We use the first number, referred to as the Percentage of Training Data (PTD), to append the name of each SegNet, e.g., SegNet-TL80 is a transfer-learned SegNet model using 80:20 ratio. By evaluating the output of a model when it is trained using varying values of PTD we will be able to evaluate how well increasing training data ratio improves the model's segmentation performance.

Each model is trained using the popular Stochastic Gradient Descent with Momentum algorithm [39] to update the model's weights and biases. The algorithm works by taking small steps in the direction of the negative gradient of the loss function, which is set as the cross entropy of probability distribution of each class, in order to minimize the loss function. The size of the step is modulated by a learning rate parameter that is set to 0.001. The training is performed in small batches of 40 images each, for a maximum of 100 epochs. Due to the input size requirement of VGG16 network, the input images need to be resized to 360x480 beforehand. Additionally, due to class population imbalances, we use class weighting to balance the importance of the classes. To make sure that small classes, such as the TS and AAP, are not underrepresented in our training data we set the class weighting to be inversely proportional to the class population.

The training is performed using MATLAB on a personal computer in Windows 10 with i7-7700 CPU @ 3.60GHz, 64 GB RAM, and two NVIDIA Titan X GPUs. The time



(a) (b) (c) (d)

FIGURE 4. The plot of a) Pixel Accuracy, b) Mean Accuracy, b) Frequency-weighted IoU, and d) Mean IoU of the SegNet semantic segmentation results at different training data percentages.

taken to complete the maximum number of training epochs is, as expected, linearly dependent on the number of training images used. On average it takes about 18.6 seconds per image to complete the training and it is also worth noting that there is very little difference between the average training times of each type of SegNet model.

IV. DISCUSSION AND ANALYSIS

A. SEGMENTATION RESULTS AND ANALYSIS

In this experiment, we will use a number of performance metrics to measure how well the different SegNet models perform on our dataset. These metrics include general as well as class-specific metrics. To assess how the segmentation process performs on a specific class, we calculate class-specific metrics, namely class accuracy and class intersection-over-union. The class accuracy of class c, denoted as a_c, is calculated as:

$$a_c = \frac{m_{cc}}{t_c} \tag{9}$$

The definition m_{cc} and t_c were given previously when we defined the class intersection-over-union (iou_c) in (1).

General metrics that we use to assess the overall performance of the segmentation process are pixel accuracy (a_p), mean accuracy (a_m), frequency-weighted intersection-over-union (iou_{fw}) and mean intersection-over-union (iou_m). The formula used to calculate these metrics are provided as follows:

$$a_p = \frac{\sum_c m_{cc}}{\sum_c t_c} \tag{10}$$

$$a_m = \frac{\sum_c a_c}{m_{cl}} \tag{11}$$

$$iou_m = \frac{\sum_c iou_c}{m_{cl}} \tag{12}$$

$$iou_{fw} = \frac{\sum_c (t_c \times iou_c)}{\sum_c t_c} \tag{13}$$

The pixel accuracy and mean accuracy results at different PTD values are plotted in Fig. 4a and Fig. 4b, respectively. The results show that SegNet-FS models produce very low accuracies at the low end of the PTD scale and increases in performance as the PTD increases. On the other hand, SegNet-TL models produce more accurate segmentation even at low PTD and is significantly less sensitive to the values of PTD used. The figure also shows significant differences between the two accuracy metrics. The figure shows the pixel accuracies are consistently higher than mean accuracies when the same PTD is used. Pixel accuracy measures the proportion of correctly labelled pixel in the entire pixel population in the dataset. This metric does not allow us to see how accurate the segmentation is for each class but at the same time is affected by class population imbalance. High accuracy in the largest class will significantly mask poor performance in smaller classes. This is the case in our experiment as can be seen in Fig. 4a and Fig. 4b. The fact that the mean accuracy is lower than the pixel accuracy suggests that we have one dominant class that has a significantly higher accuracy than the others.

The use of intersection-over-union metric gains popularity in the image segmentation research community because of one significant limitation of the accuracy metric. If there exists a class (or classes) of pixels that is not important to the calculation of the segmentation performance, one could design a classifier that increases the accuracy of other, more important, classes in the expense of the accuracy of the less important class, hence artificially boosts the overall segmentation accuracy. Segmentation algorithms that do this tend to produce high number of false alarms or false positives. Moving away from this limitation, many image segmentation challenges such as Microsoft COCO challenge [40] introduced intersection-over-union (IoU) as a more representative metric. As with accuracy metric, there are also two versions of IoU metric that can be used. Frequency-weighted IoU is the equivalent of pixel accuracy whereas mean IoU is the equivalent of mean accuracy. The plot of mean IoU and Frequency-weighted IoU results are shown in Fig. 4c and Fig. 4d, respectively.

When we compare the pixel accuracy and frequency-weighted IoU results, we conclude that there is not much significant difference between them. This is expected since in our case, there is no class of pixels that is not used in the calculation of the segmentation performance. In other words, the union of all pixels belonging to every class should comprise the total population of pixels under consideration. However, when we compare the mean accuracy and the mean IoU graphs, we can see that the segmentation performance

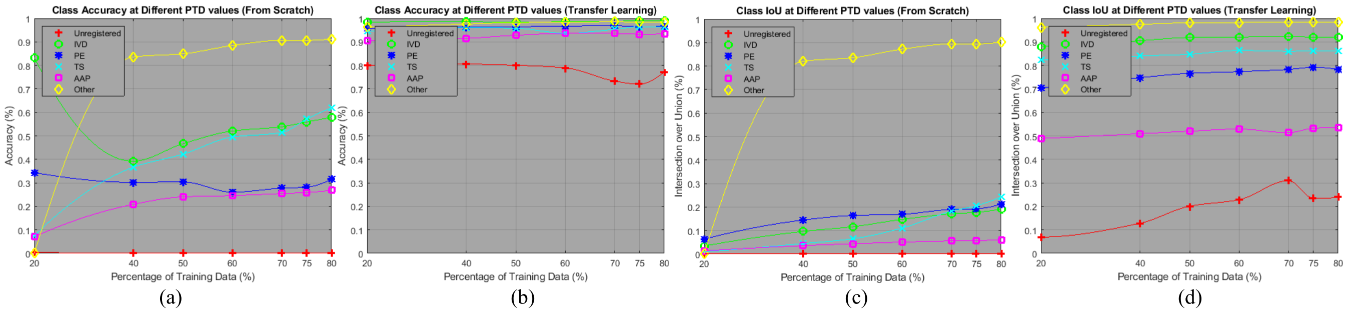


FIGURE 5. The individual class performance of the SegNet classifier when trained from scratch and using transfer learning.

measured using the latter metric is consistently lower than the former across all PTD values. To further investigate the underlying reasons for this phenomenon, we observe the individual class accuracies and class IoUs. These are given Fig. 5.

One of the most noticeable findings from our observation of the individual class performance is that the 'Other' class is consistently better than all other classes – across all PTD values and regardless whether the model is trained from scratch or using transfer learning. Using this view, we can clearly identify which of the six classes is the worst performer. We found that the 'Unregistered' class to be consistently the worse region to classify. This is expected since the population of pixels that belongs to this class is the lowest and also at the same time may not necessarily be present in all images. In fact, we believe this could be the reason behind the peculiar shape of the plot of the unregistered class shown in Fig. 5.b and Fig. 5.d. We suspect that, by chance, the randomized selection of the training data below the 75% mark had not picked up a sufficient number of unregistered pixels for training.

By comparing the performance results of SegNet-FS (Fig. 5a and Fig. 5c) and SegNet-TL (Fig. 5b and Fig. 5d), we concluded that the SegNet models trained through transfer learning produce better segmentation than the SegNet models trained from scratch. Additionally, we also concluded that the results get marginally better the higher the training percentage is used. Therefore, we decided to set the best semantic segmentation model to use to be SegNet-TL80. The segmentation performance of this network is summarized in Table 3.

By observing the range of values of the two class-specific metrics in Table 3, we concluded that the class IoU metric is the best image segmentation metric to use due to its ability to differentiate much more clearly each class individual performances. The class accuracy metric, on the other hand, produces almost identical values for the last five regions.

B. MEASURE OF AGREEMENT BETWEEN LABELS

In this section, we will discuss the performance of SegNet-TL80 model and compare it with the inherent agreement score of the ground truth data. To provide a complete comparison of the two, we calculate three types of agreement. The first one is the *inter-rater* agreement which is the degree of agreement among the five labelers who manually produced the original label images. The second one is the *rater-ground-truth* (Rater-GT) agreement which measures the degree of agreement between the labelers and ground truth data. The third one is the *model-ground-truth* (Model-GT) agreement which measures the degree of agreement between the automatically segmented results and the ground truth data. The latter, in essence, is a measure of overall performance of the segmentation model but with one significant difference compared to the result we presented in the previous section. This time, instead of using only the testing subset of the ground truth images to calculate the performance, we use the entire ground truth set. This is the case because the agreement score needs to be calculated using the entire dataset.

TABLE 3. Performance of the best semantic segmentation model (SegNet-TL80).

Regions (label/class)	a_c	iou_c
Unregistered	0.50	0.21
Intervertebral Disc	0.99	0.92
Posterior Element	0.96	0.78
Thecal Sac	0.96	0.85
AAP	0.93	0.53
Other	0.98	0.98

The strategy for the analysis is as follows. The inter-rater agreement is calculated using the statistics of the similarity score between every possible paired combination of the label images by the five labelers. Pair-wise combinations of five labelers yield ten pairs of combination, each of which has 1,545 images. We therefore have the statistics of 15,450 inter-rater agreement scores to analyze. The Rater-GT agreement is calculated between the ground truth and each of the five labelers' outputs. This provides us with 7,725 Rater-GT agreement scores. Likewise, the Model-GT agreement is calculated between the ground truth and the automatically segmented images. This provides us with 1,545 Model-GT agreement scores.

It is important to note that since the manual annotation labels do not contain 'Unregistered' class whereas the automatically segmented labels do, we merge this class with the

TABLE 4. The performance of SegNet-TL80 in comparison with the inherent agreement scores of the ground truth data.

	min	max	mean
Inter-Rater κ	0.92	0.99	0.98
Rater-GT κ	0.93	0.99	0.99
Model-GT κ	0.96	0.99	0.98
Inter-Rater x_{fw}	0.62	0.97	0.92
Rater-GT x_{fw}	0.64	0.98	0.95
Model-GT x_{fw}	0.79	0.95	0.91

'Other' class when dealing with the predicted label images. We believe this is a reasonable solution to overcome the discrepancies between the numbers of classes in the three cases. Furthermore, our inspection of the images that have some 'Unregistered' pixels shows that these pixels are mainly present around the image edges and are often part of the 'Other' classes prior to the registration process.

We use two metrics that are widely used for this purpose. The first one is the unweighted Cohen's kappa coefficient [41], denoted as κ, and the other is the frequency-weighted average of our own confidence metric, denoted as x_{fw}, calculated as the frequency-weighted intersection over union. Cohen's kappa coefficient is a statistical measure of inter-rater reliability. It is generally thought to be a more robust measure than simple percentage agreement calculation since it takes into account the agreement occurring by chance. The frequency-weighted confidence metric, as we have argued previously, is derived from the Jaccard Index which is one of the most widely used similarity metrics for segmentation. Using these two metrics, at the end we have 30,900 inter-rater agreement scores, 15,450 Rater-GT agreement scores and 3,090 Model-GT agreement scores. For each category, we decided to take their minimum, maximum and mean values as representative values and present them in Table 4.

The minimum and maximum inter-rater agreement values represent pairs of label images that are hardest and easiest to annotate, respectively. From the results shown in Table 4, we can show that the performance of the SegNet-TL80 model is within the range of the Inter-Rater and Rater-GT agreement scores. By comparing the minimum and maximum agreement scores, we show that the model performs better than manual labelling at segmenting the hardest image and on-par when it comes to segmenting the easiest images. In all cases, all κ scores are above the 80% threshold for top tier Cohen's kappa band [42]. The frequency-weighted confidence metric also shows similar conclusion but with slightly lower score for the hardest-to-segment case. These results present two very important findings, first it shows that the segmentation performance is on-par with the manual labelling performance, and secondly, it also shows that the ground-truth dataset has an excellent inter-rater agreement score.

As a side note, we would also like to draw attention to the range of scores given by the two metrics. It is in our opinion that frequency-weighted confidence metric serves as a better and more discriminative metric than unweighted Cohen's kappa because it produces a wider range of agreement values.

TABLE 5. Averaged per-class contour-based score using SegNet-TL80 at different distance error tolerance values.

Regions (label/class)	$d_T = 1$	$d_T = 2$	$d_T = 3$
Unregistered	0.08	0.10	0.11
Intervertebral Disc	0.21	0.60	0.81
Posterior Element	0.19	0.54	0.72
Thecal Sac	0.35	0.83	0.94
AAP	0.26	0.61	0.72
Other	0.50	0.75	0.85

At this point, we would like to remind the reader that our objective is to achieve delineation of important boundaries in the MRI images as illustrated Fig. 2 as opposed to just segmenting them. Therefore, we also need to analyze the delineation results to measure how the selected best SegNet perform in that regard.

C. BOUNDARY DELINEATION RESULTS AND ANALYSIS

To assess the quality of the semantic segmentation results along region boundaries, we use the semantic contour-based metric suggested in [43]. The metric's calculation involves determining precision and recall for each class c, denoted as P_c and R_c, respectively.

$$p_c = \frac{1}{|B_{pc}|} \sum_{z \in B_{pc}} \left[d\left(z, B_{gc}\right) < d_T \right] \tag{14}$$

$$R_c = \frac{1}{|B_{gc}|} \sum_{z \in B_{gc}} \left[d\left(z, B_{pc}\right) < d_T \right] \tag{15}$$

where B_{pc} and B_{gc} are the sets containing the coordinates of the contour of the region of class c from the predicted and ground-truth segmentation images, respectively. The function $d(z, B)$ denotes the shortest Euclidian distance between point z and all the points in set B, and d_T denotes the distance error tolerance. The semantic contour-based score, also known as the BF-score, denoted as F_c, of class c is then calculated using the F1 score [44], i.e.,

$$F_c = \frac{2 \times P_c \times R_c}{P_c + R_c} \tag{16}$$

We present the per-class contour-based score of the segmentation result using SegNet-TL80 model, averaged over the whole dataset in Table 5.

The table shows low scores when the distance error tolerance is set to 1, but the performance picks up rapidly as the value is increased to 2 and 3 pixels. The poor contour accuracy at the lowest distance error tolerance might at first appear to be a significant disadvantage of the application of our approach to medical image segmentation. However, the level accuracies attained is comparable to other techniques when applied to other non-synthetic datasets [43]. The fact of the matter is, this metric calculates the score from all region boundaries and low boundary accuracies are

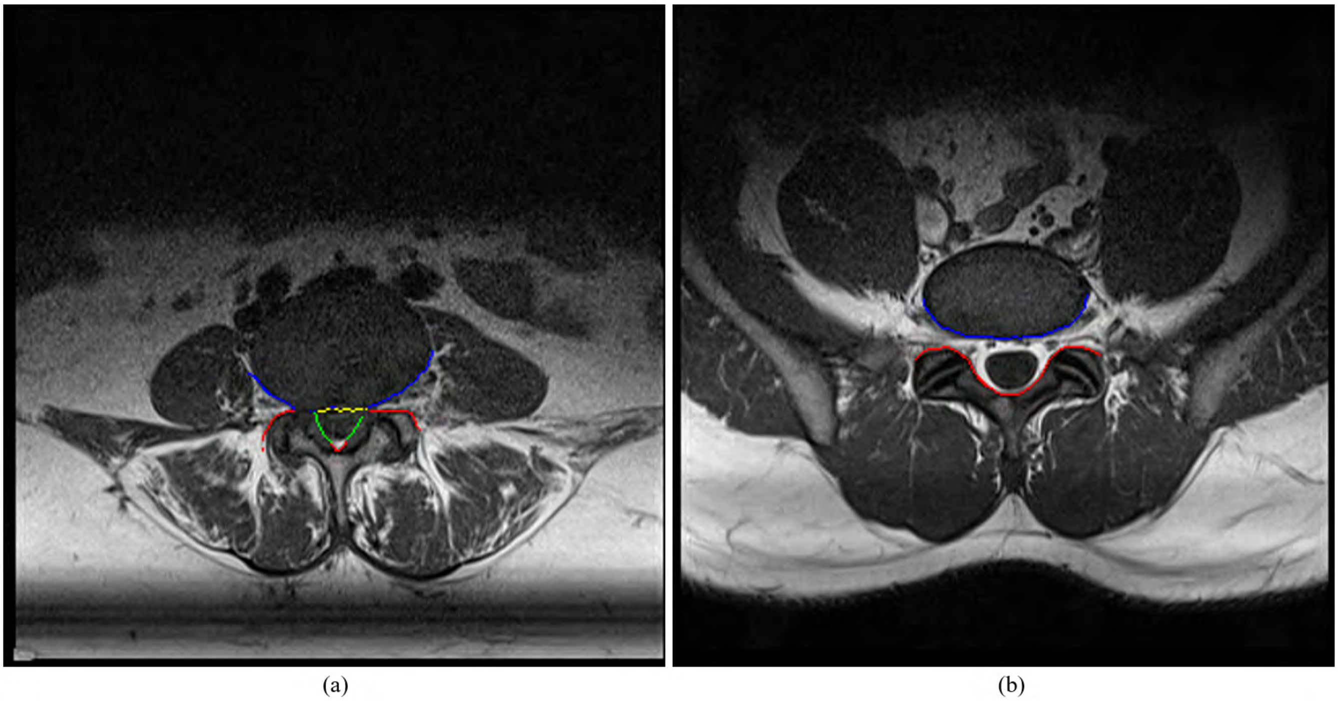


FIGURE 6. The delineation of important boundaries on a) the worst and b) the best-segmented image according to the semantic contour-based metric. Each boundary is color coded as follows: AAP-PE (Red), AAP-IVD (Blue), TS-PE (Green) and TS-IVD (Yellow).

very common. Furthermore, as we had argued previously, we are only concerned with the delineation accuracy along the important region boundaries as shown Fig. 2. Therefore, the ability of the segmentation model in getting the right delineation along those boundaries is more important.

D. VISUAL INSPECTION OF THE DELINEATION RESULTS

So far, we have discussed quantitatively the performance of the segmentation algorithms using various metrics. We have yet to show visually the final delineation results of the proposed approach along the important boundaries shown Fig. 2. Due to the large number of the data that we use, it is not practical to show all of them in this paper. Therefore, we opt to show two representative examples based on the per-image contour-based score of the segmentation result. We do this by, first, calculating that metric for all images segmented using SegNet-TL80 model. We then calculate the average BF-score of the four important regions which are IVD, PE, TS and AAP for each image. The reason why we left out the other classes is because these four regions are the regions we identified at the start to be essential in determining presence of LSS. Afterwards, we select two images with the worst and the best of the average BF-scores. We then detect the common boundaries between IVD, PE, TS and AAP regions on these two images using a Sobel edge detector. We apply a logic operation to pixels at the same location from two different edge images to detect boundaries between certain two regions. We particularly interested in the following boundaries AAP-PE, AAP-IVD, TS-PE, and TS-IVD. We visualize these boundaries superimposed on top of the original T1-weighted MRI slice using a distinct set of colors.

The results are displayed in Fig. 6. As can be seen from the figures, the delineation of the important boundaries is visually accurate even on the image with the worst semantic contour-based score of the AAP region. The presence of stenosis could be detected by locating the key points on these boundaries, as illustrated in Fig. 2, and measuring the distances between them.

V. CONCLUSION

We have presented a method to aid clinicians in performing lumbar spinal stenosis detection through delineation of important boundaries in MRI images. The method starts by applying semantic segmentation to the MRI images to locate four regions of interest, namely the Intervertebral Disc (IVD), Posterior Element (PE), Thecal Sac (TS), and Area between Anterior and Posterior (AAP) elements. We proposed to use SegNet, one of the best performing deep neural networks in the literature to date, as the pixel classifier.

Due to the limitation in size and suitability of the currently existing open-access lumbar spine dataset, we decided to develop our own dataset. Our dataset contains clinical lumbar spine MRI study of 515 patients with symptomatic back pains. Each study is annotated by expert radiologists with notes regarding the observed characteristics, condition of the lumbar spine, or presence of diseases, these include bone marrow disease, end plate degeneration, IVD bulges, Thecal Sac compressing, central or foraminal stenosis, annular tears, scoliosis, endplate defects (Modic type), facet joint and Ligamentum Flavum hypertrophy, and spondylolisthesis.

From this dataset, a ground-truth label image dataset is developed. It can be used to train and test an image segmentation model. Due to the lack of appropriate methodologies in the literature to assess the quality of ground-truth datasets, we developed two novel metrics to assess the accuracy and variability of a ground truth dataset, namely the confidence and consistency metrics, respectively. These metrics are derived from the widely used intersection-over-union metric to measure accuracy of image segmentation algorithms.

We trained a number of SegNet models using a combination of different training setups. The first setup trained the model from scratch, i.e., using random valued initial weights. The second setup uses pre-trained initial weights provided by the VGG16 network. We experimented using a variety of training-to-testing-percentage ratios on each of the above setups.

We analyze the results of the semantic segmentation and the delineation results using a comprehensive set of contour-based and region-based performance metrics including accuracy, intersection-over-union, and BF score. Our experiment shows that the different pixel classifiers produce varying levels of performances, but in general the model that uses the VGG16 pre-trained initial weights, as opposed to initial random weights, is the best. We also found that using 80:20 ratio of training to testing percentages provides the best performance across the board. Therefore, we concluded that the SegNet-TL80 to be the best segmentation model to use.

The performance of this model is also analyzed in two ways. First, by measuring and comparing inter-rater agreement, rater-ground-truth agreement, and model-ground-truth agreement. We concluded that 1) the model's performance is within the range of manual labelling performance and 2) the ground-truth dataset has an excellent inter-rater agreement score. Qualitatively, we presented two representative boundary delineation results. The results are selected from the entire dataset based on the worst and best contour-based metric score because they provide an indication of the range of quality of the boundary delineation results. Through visual inspection of these results, we can confidently claim that our proposed method is sufficiently accurate, and the results are suitable for computer-aided-diagnosis purposes.

The finding presented in this paper is part of our overall approach to develop a computer-assisted diagnosis of chronic lower back pain which was detailed in our previous publication [45]. The MRI study dataset, the ground-truth label dataset, and the MATLAB source code used in this research are made available freely to the benefit of the research community. The MRI images are stored in Siemens IMA (DICOM) format, the label ground truth and all extracted slices are stored as greyscale images in PNG format with lossless compression to preserve the quality.

APPENDIX

The information on how to download our dataset including the MRI studies, radiologists notes, manual label data, ground truth label data and the MATLAB source code, can be found on our official website [46].

REFERENCES

[1] K. P. Botwin and R. D. Gruber, "Lumbar spinal stenosis: Anatomy and pathogenesis," *Phys. Med. Rehabil. Clinics North Amer.*, vol. 14, no. 1, pp. 1–15, Jan. 2003.

[2] H. S. Nguyen *et al.*, "Upright magnetic resonance imaging of the lumbar spine: Back pain and radiculopathy," *J. Craniovertebral Junction Spine*, vol. 7, no. 1, pp. 31–37, 2016.

[3] B. B. Hansen *et al.*, "Conventional supine MRI with a lumbar pillow—An Alternative to Weight-bearing MRI for diagnosing spinal stenosis?: A cross-sectional study," *Spine*, vol. 42, no. 9, pp. 662–669, 2017.

[4] M. Merkle *et al.*, "The value of dynamic radiographic myelography in addition to magnetic resonance imaging in detection lumbar spinal canal stenosis: A prospective study," *Clin. Neurol. Neurosurg.*, vol. 143, pp. 4–8, Apr. 2016.

[5] Z. Gao *et al.*, "Automated framework for detecting lumen and media–adventitia borders in intravascular ultrasound images," *Ultrasound Med. Biol.*, vol. 41, no. 7, pp. 2001–2021, 2015.

[6] S. Su, Z. Hu, Q. Lin, W. K. Hau, Z. Gao, and H. Zhang, "An artificial neural network method for lumen and media-adventitia border detection in IVUS," *Comput. Med. Imag. Graph.*, vol. 57, pp. 29–39, Apr. 2017.

[7] S. Su, Z. Gao, H. Zhang, Q. Lin, W. K. Hau, and S. Li, "Detection of lumen and media-adventitia borders in IVUS images using sparse autoencoder neural network," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 1120–1124.

[8] X. Du *et al.*, "Deep regression segmentation for cardiac bi-ventricle MR images," *IEEE Access*, vol. 6, pp. 3828–3838, 2018.

[9] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.

[10] S. Wang *et al.*, "Multiple sclerosis identification by 14-layer convolutional neural network with batch normalization, dropout, and stochastic pooling," *Front. Neurosci.*, vol. 12, p. 818, Nov. 2018.

[11] C. P. Loizou, V. Murray, M. S. Pattichis, I. Seimenis, M. Pantziaris, and C. S. Pattichis, "Multiscale amplitude-modulation frequency-modulation (AM–FM) texture analysis of multiple sclerosis in brain MRI images," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 1, pp. 119–129, Jan. 2011.

[12] Y.-D. Zhang, C. Pan, J. Sun, and C. Tang, "Multiple sclerosis identification by convolutional neural network with dropout and parametric ReLU," *J. Comput. Sci.*, vol. 28, pp. 1–10, Sep. 2018.

[13] H. Jiang *et al.*, "Quantitative evaluation of lumbar disc herniation based on MRI image," in *Abdominal Imaging. Computational and Clinical Applications*. Berlin, Germany: Springer, 2012, pp. 91–98.

[14] R. S. Alomari, J. J. Corso, V. Chaudhary, and G. Dhillon, "Lumbar spine disc herniation diagnosis with a joint shape model," in *Computational Methods and Clinical Applications for Spine Imaging*. Cham, Switzerland: Springer, 2014, pp. 87–98.

[15] R. S. Alomari, J. J. Corso, V. Chaudhary, and G. Dhillon, "Automatic diagnosis of lumbar disc herniation with shape and appearance features from MRI," in *Proc. SPIE*, vol. 7624, Mar. 2010, Art. no. 76241A.

[16] X. He, H. Zhang, M. Landis, M. Sharma, J. Warrington, and S. Li, "Unsupervised boundary delineation of spinal neural foramina using a multi-feature and adaptive spectral segmentation," *Med. Image Anal.*, vol. 36, pp. 22–40, Feb. 2017.

[17] K. O. Babalola *et al.*, "An evaluation of four automatic methods of segmenting the subcortical structures in the brain," *NeuroImage*, vol. 47, no. 4, pp. 1435–1447, Oct. 2009.

[18] Digital Imaging Group. (2018). *SpineWeb*. Accessed: Jul. 8, 2018. [Online]. Available: http://spineweb.digitalimaginggroup.ca

[19] A. S. Al Kafri *et al.*, "Segmentation of lumbar spine MRI images for stenosis detection using patch-based pixel classification neural network," in *Proc. IEEE Congr. Evol. Comput.*, Jul. 2018, pp. 1–8.

[20] G. Wang *et al.*, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1562–1573, Jul. 2018.

[21] F. Natalia *et al.*, "Development of ground truth data for automatic lumbar spine MRI image segmentation," in *Proc. IEEE 20th Int. Conf. High Perform. Comput. Commun.; IEEE 16th Int. Conf. Smart City; IEEE 4th Int. Conf. Data Sci. Syst.*, Jun. 2018, pp. 1449–1454.

[22] V. Vapnik, E. Levin, and Y. Le Cun, "Measuring the VC-dimension of a learning machine," *Neural Comput.*, vol. 6, no. 5, pp. 851–876, 1994.

[23] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 2013.

[24] S. Krig, "Ground truth data, content, metrics, and analysis," in *Computer Vision Metrics: Survey, Taxonomy, Analysis*. Berkeley, CA, USA: Apress, 2014, pp. 283–311.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[26] R. V. Gilchrist, C. W. Slipman, and S. M. Bhagia, "Anatomy of the intervertebral foramen," *Pain Physician*, vol. 5, no. 4, pp. 372–378, 2002.

[27] M. Styner, C. Brechbuhler, G. Szckely, and G. Gerig, "Parametric estimate of intensity inhomogeneities applied to MRI," *IEEE Trans. Med. Imag.*, vol. 19, no. 3, pp. 153–165, Mar. 2000.

[28] N. Dhanachandra, K. Manglem, and Y. J. Chanu, "Image segmentation using K -means clustering algorithm and subtractive clustering algorithm," *Procedia Comput. Sci.*, vol. 54, pp. 764–771, Jan. 2015.

[29] H. A. Vrooman *et al.*, "Multi-spectral brain tissue segmentation using automatically trained k-nearest-neighbor classification," *NeuroImage*, vol. 37, no. 1, pp. 71–81, 2007.

[30] M. Keshani, Z. Azimifar, F. Tajeripour, and R. Boostani, "Lung nodule segmentation and recognition using SVM classifier and active contour modeling: A complete intelligent system," *Comput. Biol. Med.*, vol. 43, no. 4, pp. 287–300, May 2013.

[31] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[33] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, vol. 1, no. 2, pp. 1925–1934.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

[35] S. Ioffe and C. Szegedy. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift." [Online]. Available: https://arxiv.org/abs/1502.03167

[36] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[37] A. de Brébisson and P. Vincent. (2015). "An exploration of softmax alternatives belonging to the spherical loss family." [Online]. Available: https://arxiv.org/abs/1511.05042

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[39] S. Ruder. (2016). "An overview of gradient descent optimization algorithms." [Online]. Available: https://arxiv.org/abs/1609.04747

[40] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[41] D. V. Cicchetti, A. Klin, and F. R. Volkmar, "Assessing binary diagnoses of bio-behavioral disorders: The clinical relevance of Cohen's Kappa," *J. Nervous Mental Disease*, vol. 205, no. 1, pp. 58–65, 2017.

[42] D. G. Altman, *Practical Statistics for Medical Research*. Boca Raton, FL, USA: CRC Press, 1990.

[43] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?" in *Proc. 24th Brit. Mach. Vis. Conf.*, vol. 27, pp. 1–11, Sep. 2013.

[44] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

[45] A. S. Al Kafri *et al.*, "A framework on a computer assisted and systematic methodology for detection of chronic lower back pain using artificial intelligence and computer graphics technologies," in *Intelligent Computing Theories and Application* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2016, pp. 843–854.

[46] F. Natalia, H. Meidia, N. Afriliana, A. S. Al Kafri, and S. Sudirman. (2019). *Computer Assisted and Systematic Methodology for Detection of Chronic Lower Back Pain (CLBP) using Artificial Intelligence and Computer Graphics Technologies*. Accessed: Mar. 26, 2019. [Online]. Available: https://research.umn.ac.id/web/clbp/homepage

ALA S. AL-KAFRI received the bachelor's degree in computer science from Al Yarmouk University, in 2004, and the master's degree (Hons.) in computer science from the New York Institute of Technology, in 2005. He is currently pursuing the Ph.D. degree with Liverpool John Moores University. He has worked as a Lecturer for a variety of computer science subjects with United Arab Emirates University, UAE, and A'Sharqiya University, Oman, before he started the Ph.D. degree, in 2015. His areas of interests include image processing and machine learning. He has published a number of papers in the area of healthcare, image segmentation, and machine learning.

SUD SUDIRMAN received the B.Eng. degree (Hons.) from The University of Sheffield, in 1998, and the Ph.D. degree in computer science from Nottingham University, in 2003. He is currently a Senior Lecturer with the Department of Computer Science, Liverpool John Moores University, U.K. He has worked previously as a Research Assistant with the University of Derby and University of Leeds, in 1999 and 2000, respectively, before joining Liverpool John Moores University, in 2002. He teaches computer programming and computer graphics at undergraduate level and supervises a number of doctorate candidates in various computer science fields. His research interests include image analysis, image watermarking, machine learning, and agriculture technology.

ABIR HUSSAIN is currently a Professor of machine learning, and is the Head of the Applied Computing Research Group, Faculty of Engineering and Technology. She is also a Ph.D. supervisor and an External Examiner for research degrees, including Ph.D. and M.Phil. She is one of the initiators and chairs of the Development in e-Systems Engineering (DeSE) series, most notably illustrated by the IEEE technically sponsored DeSE International Conference Series. She has worked with higher order and recurrent neural networks and their applications to financial, physical, e-health, and image compression techniques.

DHIYA AL-JUMEILY is currently a Professor of artificial intelligence, and the Associate Dean of External Engagement for the Faculty of Engineering and Technology. He has extensive research interests covering a wide variety of interdisciplinary perspectives concerning the theory and practice of Applied Computing in medicine, virtual and augmented reality, human biology, and health care. He has published well over 180 peer reviewed scientific publications, seven books and five book chapters, in multidisciplinary research areas, including: technology-enhanced learning; AI-based clinical decision-making; medical knowledge engineering; and intelligent medical information systems. He is a member of the editorial board and a review committee for a number peer reviewed international journals, and is on Program Committee or as a General Chair for a number of international conferences.

FRISKA NATALIA received the Ph.D. degree in industrial engineering from Kyungsung University, Busan, South Korea. She was a Postdoctoral Fellow with Pusan National University, Busan. Her teaching and research interests include system and analysis design, logistics, E-commerce, computer programming, modelling and simulation, business intelligence, and big data. She is currently the Head of the Research Centre at Multimedia Nusantara University, and also a Lecturer with the Department of Information Systems.

HIRA MEIDIA received the B.Eng. and Ph.D. degrees from The University of Sheffield, U.K. She has worked as a Research Assistant with Arizona State University and Delaware University, USA. She has been a Senior Lecturer with Universitas Multimedia Nusantara, Indonesia, since 2008, where she is currently the Vice Rector for Academic Affairs. Her research interests are in the simulation, modelling of smart drug delivery systems, materials, and nanotechnology. She has reviewed a number of scientific papers for international conferences and regional journals. She is also a Reviewer for research project in Indonesia. She has led and co-led a number of collaborative research projects between Universitas Multimedia Nusantara and other universities.

NUNIK AFRILIANA received the bachelor's degree in computer science from President University, Indonesia, in 2005, and the master's degree in information systems from Bina Nusantara University, Indonesia, in 2013. She is currently a Lecturer with the Faculty of Engineering and Informatics, and also the Head of the Academic Administration Bureau, Universitas Multimedia Nusantara, Indonesia. Her current research interest includes machine learning. She is currently contributing to research in neural networks, which is a collaborative research between Universitas Multimedia Nusantara, Ministry of Higher Education, Indonesia, and Liverpool John Moore University.

WASFI AL-RASHDAN graduated as a Dentist from the College of Dentistry, University of Baghdad, and worked as a dentist (1976–2012). He was the President of the Jordan Dental Association (JMA) (2005–2009). He is the Founder of the Ibn Al Nafis Hospital, Irbid, and has worked as the Chair of the Board of Executive Director of Ibn Al Nafis Hospital (1989–1991). He has been the Founder and the Chair of the Board Executive Director of Irbid Specialty Hospital, Irbid, since 1992. He is a member of the Board Executive Director of Jadara University, and a member of the Board Director of the Jordan University of Science and Technology Health Centre.

MOHAMMAD BASHTAWI received the bachelor's degree in medicine and surgery from the Jordan University of Science and Technology, in 1998. He achieved his Higher Speciality in Medicine–Diagnostic Radiology, in 2003. He has been the Head of the Radiology department in Irbid Speciality Hospital, Jordan, since 2003. He participates in the supervision of a number of master's and Ph.D. research students around the world.

MOHAMMED AL-JUMAILY received the M.B.Ch.B. degree (Hons.) from Al Nahrain University, Iraq, in 1995, and the master's and Ph.D. degrees in neurobiology from the Universite de Montpellier, France, in 2002 and 2007, respectively. He is currently a Consultant Neurosurgeon with Dr. Sulaiman Al Habib Hospital, and the Lead for Clinical Research, Liverpool John Moores University. In 2010, he started as a Fellow of the Royal College of Surgeons of Edinburgh in Neurosurgery. He achieved a certificate of Completion of Training in Neurosurgery, General Medical Council, in 2011. He research interest includes minimally invasive spine interventions that preserve the mobility of the spine as well as relieving the patient's symptoms.