Development of Ground Truth Data for Automatic Lumbar Spine MRI Image Segmentation

Friska Natalia
Department of Information Systems
Universitas Multimedia Nusantara
Tangerang, Indonesia
friska.natalia@umn.ac.id

Hira Meidia
Department of Electronic Engineering
Universitas Multimedia Nusantara
Tangerang, Indonesia
hira.meidia@umn.ac.id

Nunik Afriliana
Department of Informatics
Universitas Multimedia Nusantara
Tangerang, Indonesia
nunik@umn.ac.id

Ala S. Al-Kafri
Department of Computer Science
Liverpool John Moores University
Liverpool L3 3AF, UK
a.s.alkafri@2015.ljmu.ac.uk

Sud Sudirman
Department of Computer Science
Liverpool John Moores University
Liverpool L3 3AF, UK
s.sudirman@ljmu.ac.uk

Andrew Simpson
Department of Computer Science
Liverpool John Moores University
Liverpool L3 3AF, UK
a.simpson@ljmu.ac.uk

Ali Sophian
International Islamic University
Kuala Lumpur, Malaysia
ali_sophian@iium.edu.my

Mohammed Al-Jumaily
Dr Sulaiman Al Habib Hospital
Dubai Healthcare City
Dubai, UAE
maljumaily@yahoo.fr

Wasfi Al-Rashdan
Irbid Speciality Hospital
Irbid, Jordan
drwasfi1@hotmail.com

Mohammad Bashtawi
Irbid Speciality Hospital
Irbid, Jordan
mohdbakhiet@yahoo.com

***Abstract*— Artificial Intelligence through supervised machine learning remains an attractive and popular research area in medical image processing. The objective of such research is often tied to the development of an intelligent computer aided diagnostic system whose aim is to assist physicians in their task of diagnosing diseases. The quality of the resulting system depends largely on the availability of good data for the machine learning algorithm to train on. Training data of a supervised learning process needs to include ground truth, i.e., data that have been correctly annotated by experts. Due to the complex nature of most medical images, human error, experience, and perception play a strong role in the quality of the ground truth. In this paper, we present the results of annotating lumbar spine Magnetic Resonance Imaging images for automatic image segmentation and propose confidence and consistency metrics to measure the quality and variability of the resulting ground truth data, respectively.**

Keywords—Ground Truth, Confidence Metric, Consistency Metric, Lumbar Spine MRI, Image Segmentation

I. INTRODUCTION

Fully automated computer systems have never replaced, nor are they expected to replace in the foreseeable future, experts performing medical diagnoses. Instead, Computer Aided Diagnosis (CAD) systems have been researched and developed to improve the effectiveness and efficiency of such procedures. For example, a CAD system is used to locate regions of interest in medical images for a radiologist to focus on, or to provide a "second opinion" before a physician makes their final decision as opposed to deciding what the diagnosis is automatically.

CAD has been a major research topic in medical imaging and diagnostic radiology for over five decades. One of the earliest attempts at computerised analysis of medical images was made in 1960s to help the diagnosis of bone tumours [1]. Initially, research in this field focused on achieving automated computer diagnosis which focus on developing computer algorithms to carry out full medical diagnosis. However, realisation on the limitation and risks of using such systems slowly transformed the research focus to computer-aided-diagnosis. The concept of CAD is different to that of the former because the system is expected to complement the physicians' ability rather than replacing it [2]. Both types of system, however, involve modelling and embedding of medical knowledge of sort and thus incorporate a design of artificial intelligence through machine learning.

Two learning paradigms in machine learning, namely supervised and unsupervised learning, have always been a popular subject of research, comparison and analysis by researchers. Arguably, the majority of practical machine learning algorithms, including those used in CAD systems, are trained using supervised learning. In supervised learning, the algorithm is trained to map the input variables to the output variables using pairs of known input and output values called training data set. The resulting algorithm, which can manifest as a mapping function, a decision tree or a neural network, can then be tested for performance using another set of known input and output values, called test data set. Both training and test data sets consist of *Ground Truth Data* that are developed by either manually assigning labels to the input data, or collected by taking measurements from real world experiments. The process of how the ground truth data is obtained depends on the task of

which the machine learning is set to do. Furthermore, due to the nature by which they are obtained, ground truth data can have some degree of inaccuracy.

Automatic image segmentation is one of the fundamental steps in medical image analysis. Here the ground truth data is obtained by manually assigning labels to each pixel in the image by experts. However, the resulting output labels can also be subjective and their quality can vary depending on the expert's opinion or analysis.

In this paper, we are presenting a discussion on the process of manual annotation of pixels when developing ground truth data. We will use one of our earlier works on lumbar spinal stenosis detection in lumbar spine Magnetic Resonance Imaging (MRI) images [3] as a case study. We will present our argument on the challenges in developing ground truth data by drawing from the experience we have in doing so. As one of the main contributions of this paper, we are presenting two unique metrics to measure the confidence level and consistency level of the produced ground truth data.

II. Lumbar Spine and Lumbar Spinal Stenosis

The lumbar spine refers to the lower back of our vertebral body, a set of interlocking bones that forms the spinal column. Lumbar spine consists of five spinal segments that connect the second upper part of the spine called thoracic spine, and the lower part called the sacral spine. Because of its placement in our spinal column, lumbar spine bears more weight than other parts of the spine. In fact, the lower the vertebra is in the spinal column, the more weight it must bear hence the more prone it is to degradation and injury.

When such degradation or injury occurs, a person will experience pain. A chronic type of this pain, referred to as Chronic Lower Back Pain (CLBP), has symptoms ranging from radicular pain to atypical leg pain to neurogenic claudication [4]. CLBP is a debilitating illness that is affecting the health, social life, and employment of millions of people around the world. In the UK, the cost of treating patients with CLBP is estimated to be around £500 million annually to the National Health Service (NHS) [5]. This is on top of other economic costs resulting from the loss of productivity and other informal care – which is estimated to reach around £10,668 million [6].

MRI is the preferred method of medical scans for detecting the causes of back pain. MRI images can be used to visualise lumbar spine, slice by slice, in three view-planes namely sagittal (side), axial (top-down) and coronal (frontal) – typically only the first two are used in lumbar spine MRI. A mid-sagittal MRI view shown in Fig. 1 shows the five vertebrae of the lumbar spine. and adjacent ones are separated by an Intervertebral Disc (IVD) labelled D1 to D4. The last IVD, D5, separates L5 and the large triangular shaped bone at the bottom of the spine, called the *sacrum*.

Fig. 1 also shows a long white opening the anterior arch and the posterior arch. The part of this opening that is visible in this mid-sagittal cut is *Thecal Sac* (TS) which contains *Cerebrospinal Fluid* (CSF), the same type of fluid that can be found inside the brain. The back of the opening, which borders with the anterior of the posterior arch, is covered with *Ligamentum Flavum* (LF).

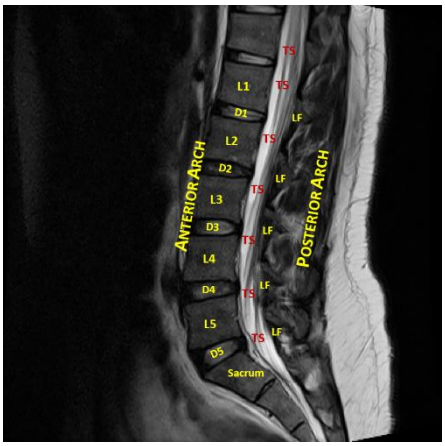

Fig. 1. Sagittal view MRI of a lumbar spine

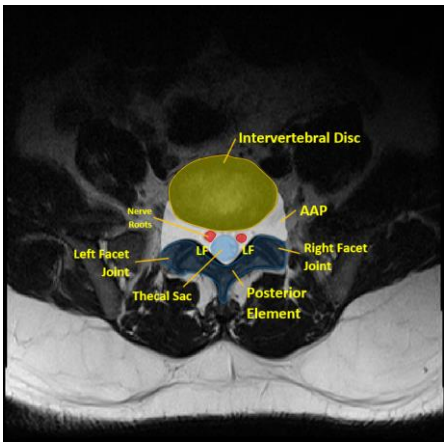

Fig. 2. Axial view MRI of D4

The axial view of the spine manifests as different slices of the MRI images across each vertebra or IVD. Such view, as illustrated in Fig. 2, can show more information on the various tissues surrounding the vertebra, the IVD and the posterior element (PE) of the vertebral body. In this view, we can also clearly see the area between the anterior and posterior element of the vertebral body which contains the thecal sac and the nerve roots in both lateral recesses. This area extends from the cervical spine down to the lumbar spine. For the lack of a better word, in this paper we will refer this area as AAP, which is a short for Area between Anterior and Posterior elements.

Lumbar spinal stenosis is a narrowing of AAP which in turn produces pressure on spinal nerve canal or roots. An abnormal compression of either of them would exert pressure and create a sensation of pain. The stenosis could occur in any part of AAP and could be caused by different types of defect such as posterior/posterolateral disc herniation, osteoarthritic thickening of the posterolateral vertebral body, or hypertrophy of LF. In all of these cases, clinicians will measure three distances in the AAP, namely the anteroposterior diameter of the spinal canal and the left and right width of the foramen. This process starts by manual delineation of the boundaries between AAP and the IVD, between AAP and the left and right facet joints, and

between the AAP and LF. The three distances and the three boundaries are illustrated in Fig. 3.

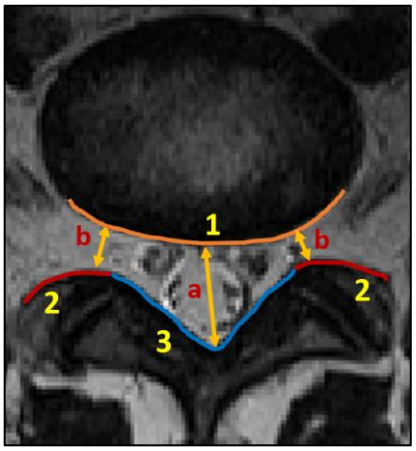

Fig. 3. The three important boundaries for stenosis detection between AAP and 1) IVD, 2) Facet Joints, and 3) LF to measure a) the anteroposterior diameter of the spinal canal and b) the left and right width of the foramen

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.

In our previous paper [3], we proposed a deep neural network solution to lumbar spine MRI image segmentation. The technique employs a patch-based approach by decomposing the input image into 25x25 overlapping patches which are then fed into a convolutional neural network that serves as pixel classifier. As with other supervised learning solutions, our technique also relies on the availability and accuracy of manually labelled MRI images that serve as ground truth data for network training and testing purposes. In the following sections, we will describe the process of developing ground truth data and evaluate the quality of the data to provide a better perspective of its effect on the quality of the trained network.

III. Research Methodology and Dataset

Compared to other topics in computer vision, little formal or analytic work has been published to guide the creation of ground truth data. There is some guidance [7], [8] 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 as opposed to other quality metrics such as accuracy and variability. This guidance states that, provided the right data is used, the larger the better [9].

Our experience in trying to obtain a reasonably large number of high quality lumbar spine MRI images had proven to be quite a challenge. The most comprehensive database of lumbar spine related medical images is hosted by SpineWeb [10], however it contains relatively small-sized and incomplete datasets taken from between 8 to 125 patients. As a result, we worked together with several speciality hospitals around the world to gather a sufficient amount of MRI scans for our dataset. Our dataset consists of complete clinical lumbar MRI scans of 568 symptomatic back pain patients, each come with a diagnosis report by an expert radiologist. The patients are mixed gender, height and age.

Each patient had one or more MRI studies, or a set of series of scans, associated with it. Each set of scans contains *slices*, i.e., images taken from sagittal and axial views, of the lowest three vertebrae and lowest three discs. Some studies contain slices of more than three vertebrae. The number of slices ranges from 12 to 20 in many cases. The majority of the slices have an image resolution of 320x320 pixels, however there are small number of slices with varying lower resolution as low as 256x256 pixels. The slices have pixel precision of 12-bit per pixel which is higher than standard greyscale images.

Each MRI study is taken at a given time with the patient lying on their back, in supine position. A patient may have one or more of these sets taken at a different time, a few days apart. The scans contain both T1-weighted and T2-weighted MRI images. These are different images of the same organs but with different contrasts and pixel intensities due to the different relaxation times of tissues when excited by magnetic field. For any further and more detailed information on MRI and its uses as medical imaging technology, interested readers can refer to [11] or any other relevant textbooks in this area.

In our previous work [3], we have provided rationale on why axial view MRI slices on the last three IVDs provide the best image for lumbar spine stenosis detection. Hence, in this paper we will be concentrating only on the development and evaluation of ground truth data using axial-view slice of the last three IVDs. The total number of slices we had to work on is therefore 1704.

Fig. 4. The four labelled RoIs namely 1) Intervertebral Disc, 2) Posterior Element, 3) Thecal Sac and 4) the AAP

The ground truth used for training and testing of an image segmentation machine learning algorithm consists of labelled images marking a number of regions of interest (RoIs). Since lumbar spine stenosis occurs inside AAP, i.e., the area between IVD and PE, we focused on the area of the MRI which contains those regions. Subsequently, we decided to have four RoIs which are a) the IVD, b) the PE, c) the TS, and d) the AAP. Any other pixels that do not belong to any one of the above four

regions are labelled as e) other. The labelling of these four regions is illustrated in Fig. 4.

The task of manually labelling the four areas on each of the 1704 slices is a laborious one. On average, five to ten minutes are spent to label each slice. This provides a significant challenge for us if we were to use the highly valuable expert's time to perform it. As an alternative, we opted to use the expert's time to label several MRI images as examples and use them to guide non-experts when labelling the dataset. We also use the examples that the expert has created to select the best results from the output. The steps to develop the ground truth data are detailed below:

1. We use a dual-view setup on a DICOM/MRI viewer illustrated in Fig. 5. On the mid-sagittal view, we observe the cut line (yellow line) of the corresponding axial-view slice on the right. The best axial-view slice of a disk, defined as one which cuts closest to the half-height of the disc, is selected.

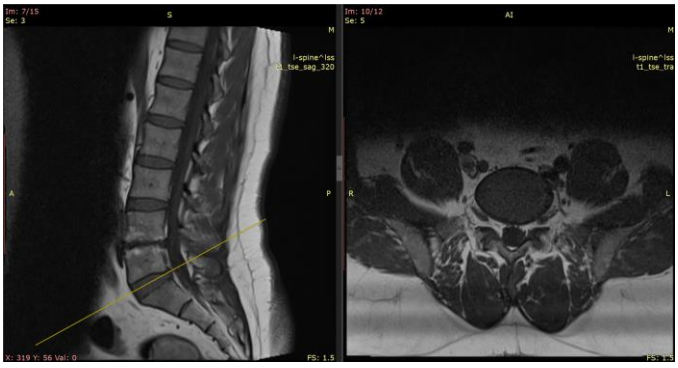

Fig. 5. Dual-view of MRI images showing the mid-sagittal view (left) and the corresponding axial-view slice (right) at the cut line.

2. We then extract three best T1-weighted MRI slices, one from each of the lowest three IVD, i.e., D3, D4 and D5, from each patient's MRI study. For brevity, we will refer to these selected slices as the images from now on. The total number of slices is therefore 1704.
3. We asked our expert to label the four RoIs on ten patients (30 slices in total) as examples. The ten selected images are chosen in such a way to provide as best representation to the rest of the dataset as possible.
4. We use these example images to train several undergraduate students, as project participants, to perform the same task. The training runs twice with a test at the end to assess their suitability. At the end, fourteen participants were recruited.
5. We split the database into two groups and assigned seven participants to label all images in each dataset group independently. Their results are regularly checked by the expert to ensure good level of consistency and accuracy. At the end, we have seven batches of labelled images per group.
6. From each group, the five best batches were selected by the expert to be used for subsequent stages.
7. The confidence and consistency metrics are calculated to measure the quality and variability of the resulting labelled images.

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 *intersection over union* metric [12]. The intersection over union (iou_c) of a class c can be calculated as $iou_c = \frac{n_{cc}}{t_c + n_c - n_{cc}}$, where n_{cc} is number of pixels of class c correctly predicted to belong to class c, and t_c is the total number or pixels of class c – according to the ground truth, and n_c is the total number of pixels predicted to belong to class c. However, since in our case we do not have yet the ground truth, therefore by definition, the value of both n_{cc} and t_c cannot be determined. We will now develop an alternative intersection over union metric iou'_{cv} as an estimate to iou_c

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

$$k_{nc} = \sum_p f_{npc}$$

$$f_{npc} = \begin{cases} 1 & if\ l_{np} = c \\ 0 & otherwise \end{cases}$$

The vote count has value range of $0 \leq k_{nc} \leq 5$, e.g., $k_{n1} = 0$ means the pixel n receives zero vote that assigns class 1 to it.

Next, we define the intersection of c-labelled regions as s_{cv}, the normalised number of pixels that receive at least v number of votes that assign class c. We refer v as the vote-threshold.

$$s_{cv} = \frac{1}{n} \sum_n g_{ncv}$$

$$g_{ncv} = \begin{cases} 1 & if\ k_{nc} \geq v \\ 0 & otherwise \end{cases}$$

Note that since we only consider pixels that have at least one vote that assigns class c, this means $g_{ncv} \geq 1$ for $\forall n$. Therefore, these pixels will also serve as the union (denominator) in our iou'_{cv} calculation. One important fact to consider is that for $\forall c$, the following composite inequality applies:

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

We define our estimate intersection over union metric iou'_{cv} of class c and vote threshold v as,

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

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

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

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.

IV. Analysis of Image Labelling Results

The MRI scans in our dataset show varying degrees of severity. Some scans, for example, show strong evidence of intervertebral disc collapse while some others show non-existent gap between the IVD and PE. The condition in these cases is so severe that we could not reliably label the images with high degree of confidence. In total, there are 53 cases in this category.

We calculate the values of iou'_{cv} for the four regions of interest using the remaining 515 MRI scans. The results are shown in Table 1. We will next discuss and analyse the result of each individual region.

TABLE I. Intersection over Union Values of Different Vote-Threshold Groups

Regions (label/class)	iou'_{c2}	iou'_{c3}	iou'_{c4}	iou'_{c5}
Intervertebral Disc (1)	0.96	0.93	0.91	0.87
Posterior Element (2)	0.89	0.82	0.76	0.66
Thecal Sac (3)	0.87	0.81	0.74	0.66
The AAP (4)	0.66	0.48	0.34	0.21

A. Individual Region Analysis

The IVD is by far the easiest region to label. The region has more consistent characteristics across all patients which manifest as narrow range of pixel grey level values, smoother texture, as well as high contrast to the surrounding tissues. This fact is reflected by the high iou'_c values the region has compared to the other three as shown in Table I. The image shown in Fig. 6 (a) shows how sharp the heat map of this region is which indicates high level of confidence in the data for this region.

The PE has a unique shape, similar to the letter Y, as exemplified in Fig. 6 (b). It has a relatively wider range of pixel grey level values compared to IVD. Furthermore, it has lower contrast to the surrounding tissue than the latter, making it more difficult to determine its edges especially towards its lower end. This fact is reflected in its lower iou'_c values than the IVD region.

In a healthy patient, the TS will appear distinct to its surrounding in T1-weighted MRI and has a round shape as depicted in Fig. 6 (c) . However, when a central spinal stenosis occurs, the spinal canal may be squashed between the IVD and the PE. This in turn could make accurate identification of its edges more difficult. Furthermore, because of its small size, the ratio between its edges and inner pixels is large hence lowering its iou'_c values.

By far, the hardest region to label is the AAP. One of the reasons for this is 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 [13]. Its shape can vary significantly depending many factors such as the location of the slice, patient's posture as the MRI is performed, as well as the presence of illness or defects.

Usually, the top and bottom boundaries of AAP with the IVD and the PE respectively, are pronounced. However, in cases where central or lateral stenosis occurs, these boundaries narrow and become unclear. The AAP region also contains many spinal nerves, spinal arteries and veins whose locations vary significantly depending on which part of lumbar spine the AAP is. This results in inconsistent and different pixel values and texture in MRI scans. Furthermore, there seems to be very loose definition on where the region should end on each side of the vertebrae. All these reasons result in poor values of iou'_c across all vote thresholds as shown in Table I.

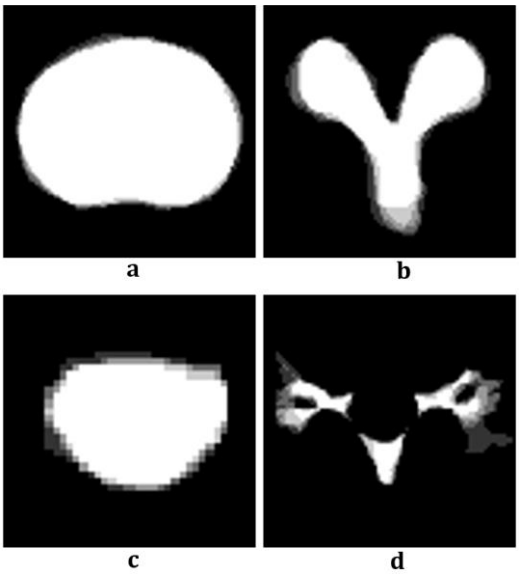

Fig. 6. Visualizing vote counts that each pixel in and around each RoI receives as heat map, where the brighter pixels have more votes than darker pixels. The RoIs are a) IVD, b) PE, c) TS and d) the AAP

B. Confidence and Consistency Metrics

At the end of the experiment, we have five groups of labelled images for each class. We need to choose one from each class that best represent the groups and to be selected as the final ground truth data. From them we will calculate the confidence metric and consistency metric, denoted as x_c and y_c respectively.

Each of the five groups is developed from its corresponding vote threshold. There is a question of which vote threshold value to be chosen as the selected vote-threshold, denoted as v_t, that will provide the best and most appropriate ground truth. The most liberal option would be to pick $v_t = 1$ as it provides the highest iou'_c value. However, this presents one problem, being it is very likely that pixels which have only one vote for one class also have at least one vote for other classes, hence we need to reconcile the different votes before assigning those pixels a label. Another alternative is to pick the most conservative

group, i.e., choosing only those pixels that have all the votes. One of the issues with this is the lowest iou'_c that the group has compared to all other options, hence it provides the least confidence level.

One intuitive, yet the best, answer to the question is to pick the majority vote. In our case, since there are in total five participants we will select $v_t = 3$ as our chosen vote threshold. This is a compromise solution that avoids both the problem of having the lowest iou' value and having to perform vote reconciliation. Therefore, a pixel n will be assigned a final classification c when $g_{nc3} = 1$. The confidence metric, x_c, in general is set as:

$$x_c = iou'_{cv_t}$$

We have previously defined consistency level as how varied the ground truth data is across its sources. One way to measure this, is by calculating the rate of change of iou'_c along the vote threshold dimension. Using the mean first derivative of the class' confidence level at the chosen vote-threshold, v_t, we calculate the consistency metric y_c as:

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

Note that the value of y_c ranges between 0 and 1, where low value suggests low consistency and high variability between sources, and vice versa. The final values of confidence and consistency values of the ground truth data are presented in Table II.

TABLE II. CONFIDENCE AND CONSISTENCY VALUES OF THE RESULTING GROUND TRUTH DATA

Regions (label/class)	x_c	y_c
Intervertebral Disc (1)	0.93	0.95
Posterior Element (2)	0.82	0.87
Thecal Sac (3)	0.81	0.87
The AAP (4)	0.48	0.68

V. CONCLUSION AND FURTHER WORK

In this paper, we have presented a method and the results of developing ground truth data for lumbar spine MRI image segmentation. We did this by first providing a rationale for developing the ground truth data, describing the anatomy of the lumbar spine and the process carried out by clinicians when analysing lumbar spine MRI images. We then detailed the method used to develop the image labels and carried out a statistical analysis of them. Critical analysis and discussion on the metrics that can be used to measure the suitability of the labelled images to be used as ground truth data were presented.

In conclusion, the contribution of this paper can be summarised as follows:

1. Novel confidence and consistency metrics to measure the suitability of the label images. These metrics are derived from the widely used intersection over union metric to measure accuracy of a machine learning algorithm.
2. Labelled images that serve as ground truth data for automatic lumbar spine MRI image segmentation.

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 [14].

ACKNOWLEDGMENT

This work is supported by PKLN funding from Indonesian Ministry of Research, Technology and Higher Education.

REFERENCES

[1] G. S. Lodwick, C. L. Haun, W. E. Smith, R. F. Keller, and E. D. Robertson, "Computer Diagnosis of Primary Bone Tumors," *Radiology*, vol. 80, no. 2, pp. 273–275, 1963.

[2] K. Doi, "Computer-Aided Diagnosis in Medical Imaging: Historical Review, Current Status and Future Potential," *Comput. Med. Imaging Graph.*, vol. 31, no. 4–5, pp. 198–211, Mar. 2007.

[3] A. S. Al-Kafri *et al.*, "Segmentation of Lumbar Spine MRI Images for Stenosis Detection using Patch-based Pixel Classification Neural Network," in *IEEE Congress on Evolutionary Computation*, 2018, p. to appear.

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

[5] G. J. Macfarlane, E. Thomas, P. R. Croft, A. C. Papageorgiou, M. I. V Jayson, and A. J. Silman, "Predictors of early improvement in low back pain amongst consulters to general practice: the influence of pre-morbid and episode-related factors," *Pain*, vol. 80, no. 1, pp. 113–119, 1999.

[6] N. Maniadakis and A. Gray, "The economic burden of back pain in the UK," *Pain*, vol. 84, no. 1, pp. 95–103, Jan. 2000.

[7] 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.

[8] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[9] S. Krig, "Ground Truth Data, Content, Metrics, and Analysis," in *Computer Vision Metrics: Survey, Taxonomy, and Analysis*, Berkeley, CA: Apress, 2014, pp. 283–311.

[10] "SpineWeb: Collaborative Platform for Research on Spine Imaging and Image Analysis," 2016. [Online]. Available: http://spineweb.digitalimaginggroup.ca/. [Accessed: 27-Apr-2018].

[11] S. W. Atlas, *Magnetic resonance imaging of the brain and spine*, 5th ed., vol. 1. Lippincott Williams & Wilkins, 2016.

[12] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *CoRR*, vol. abs/1411.4, 2014.

[13] 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.

[14] 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 *Lecture Notes in Computer Science*, 2016, pp. 843–854.