Title

Eigenrank by committee: Von-Neumann entropy based data subset selection and failure prediction for deep learning based medical image segmentation.

Authors

Gaonkar, Bilwaj
Beckett, Joel
Attiah, Mark
et al.

Publication Date

2021

DOI

10.1016/j.media.2020.101834

Peer reviewed

HHS Public Access

Author manuscript

Med Image Anal. Author manuscript; available in PMC 2022 January 01.

Published in final edited form as:

Med Image Anal. 2021 January ; 67: 101834. doi:10.1016/j.media.2020.101834.

Eigenrank by Committee: Von Neumann Entropy Based Data Subset Selection and Failure Prediction for Deep Learning Based Medical Image Segmentation

Bilwaj Gaonkar[a], **Joel Beckett**[a], **Mark Attiah**[a], **Christine Ahn**[a], **Matthew Edwards**[a], **Bayard Wilson**[a], **Azim Laiwalla**[a], **Banafsheh Salehi**[b], **Bryan Yoo**[b], **Alex Bui**[b], **Luke Macyszyn**[a]

[a]Department of Neurosurgery, University of California Los Angeles

[b]Department of Radiology, University of California Los Angeles

Abstract

Manual delineation of anatomy on existing images is the basis of developing deep learning algorithms for medical image segmentation. However, manual segmentation is tedious. It is also expensive because clinician effort is necessary to ensure correctness of delineation. Consequently most algorithm development is based on a tiny fraction of the vast amount of imaging data collected at a medical center. Thus, selection of a subset of images from hospital databases for manual delineation - so that algorithms trained on such data are accurate and tolerant to variation, becomes an important challenge. We address this challenge using a novel algorithm. The proposed algorithm named 'Eigenrank by Committee' (EBC) first computes the degree of disagreement between segmentations generated by each DL model in a committee. Then, it iteratively adds to the committee, a DL model trained on cases where the disagreement is maximal. The disagreement between segmentations is quantified by the maximum eigenvalue of a Dice coefficient disagreement matrix a measure closely related to the Von Neumann entropy. We use EBC for selecting data subsets for manual labeling from a larger database of spinal canal segmentations as well as intervertebral disk segmentations. U-Nets trained on these subsets are used to generate segmentations on the remaining data. Similar sized data subsets are also randomly sampled from the respective databases, and U-Nets are trained on these random subsets as well. We found that U-Nets trained using data subsets selected by EBC, generate segmentations with higher average Dice coefficients on the rest of the database than U-Nets trained using random sampling ($p<0.05$ using t-tests comparing averages). Furthermore, U-Nets trained using data

Send all correspondence to: bilwaj@gmail.com.

Author Statement

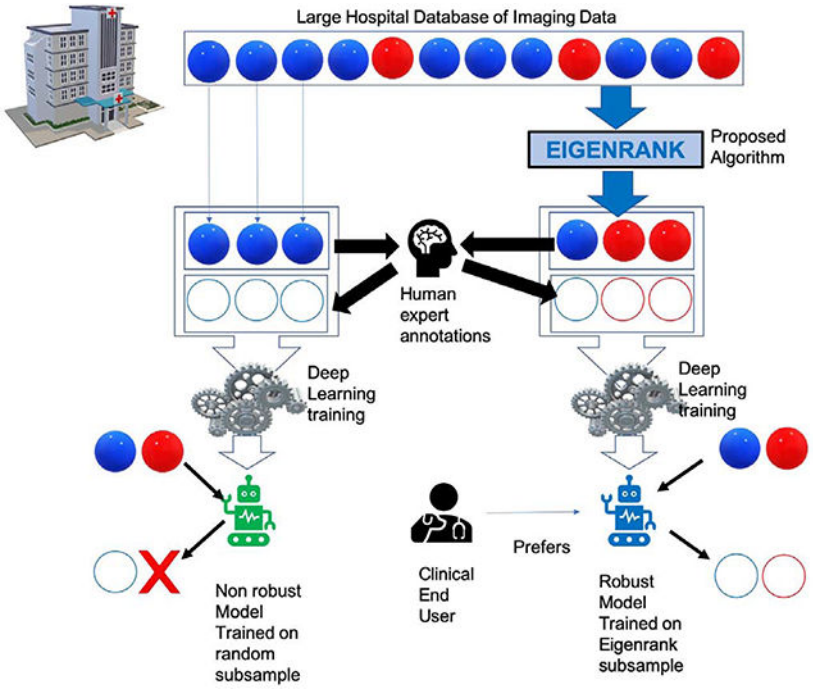
Bilwaj Gaonkar Conceptualization, Data Curation, Formal analysis, Investigation, Methodology, Software, Writing Original and Review and Editing Joel Beckett - Data curation Mark Attiah - Data curation Christine Ahn - Data curation Matthew Edwards - Data curation Bayard Wilson Writing review and editing Azim Laiwalla - Data curation Banafsheh Salehi - Data curation Bryan Yoo - Data curation Alex Bui - Writing, Review and editing, Formal analysis, Investigation Luke Macyszyn Data curation, Supervision, Validation, Writing Original and Review and Editing, Investigation, Methodology, Funding acquisition

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

subsets selected by EBC generate segmentations with a distribution of Dice coefficients that demonstrate significantly ($p<0.05$ using Bartlett's test) lower variance in comparison to U-Nets trained using random sampling for all datasets. We believe that this lower variance indicates that U-Nets trained with EBC are more robust than U-Nets trained with random sampling.

Graphical Abstract



Keywords

Active Learing; Deep Learning; Data subset selection; Failure Deep Learning

1. INTRODUCTION

1.1. Significance

Deep learning methods have become a mainstay of fully automatic medical image segmentation. These methods play a key role in the development of quantitative imaging biomarkers for a number of pathologies. However, training and deploying deep learning segmentation in practice is beset by a number of challenges. Two significant but related challenges are:

- Data subset selection (DSS) - the development of robust segmentation tools by using human annotation e$_{ff}$orts in the most efficient possible manner
- Failure Prediction (FP) - the ability to predict on which cases a deep learning based segmentation model will fail.

Both problems are significant in medical image segmentation, more than natural image segmentation, as the availability of expert annotated data for training medical image segmentation models is severely constrained. These models need to be perform adequately despite natural and pathologic variation, even when trained using datasets much smaller than those regularly used in natural image segmentation competitions. While standard machine learning improves generalizability by training on increasingly larger sets of training data, the cost of annotation is much higher in medical imaging. Given the limited availability of physician effort, it is important that manual annotation efforts be utilized in the most efficient manner when creating a new training set aimed at segmenting a specific anatomical region. We must be able to optimally choose a training subset of images for manual annotation from within the vast store of imaging data available in a standard hospital picture archiving and communication system (PACS). Moreover, this subset must be selected without the availability of any manual segmentations. This is the data subset selection (DSS) problem of medical image segmentation, which we address in this work. A related problem that emerges when one attempts to incorporate automatic medical image segmentation algorithms into clinical workflows. An algorithmic framework is not expected to be perfect. However, an algorithm that is imperfect and that can alert the attending physician to its imperfections is far more valuable than an algorithm which fails silently. The majority of existing algorithms for medical image segmentation fail silently. The DSS framework we propose induces a scan specific score, which may help predict where a deep U-Net algorithm will perform poorly or fail. We expect that our DSS framework as well as the ability to predict the possibillity of failure is critical to enable deployment of DL segmentation algorithms for clinical imaging.

1.2. Related Work

1.2.1. Data subset selection and active learning—Typical DSS aims to choose a training subset from a large dataset, such that models trained on the subset incur minimal loss compared to models trained on the complete dataset (Wei et al., 2015; Schreiber et al., 2019). Active learning on the other hand involved the ability to interactively query the user during the training process (Settles, 2009). DSS and active learning have been a part of machine learning literature for more than three decades (Settles, 2009; Rubens et al., 2015; Das et al., 2016; Zhou, 2017). Consequently, there exists substantial literature on data subset selection, active learning, as well as weakly supervised learning, all of which cannot be reviewed here. However, we note that the majority of standard DSS algorithms are designed to work with binary classification and focus on preserving classification accuracy. The closest work to ours in literature comes from pathology (Yang et al., 2017; di Scandalea et al., 2019) where uncertainty at the voxel level is used to trigger a query to the human expert to segment a patch. This strategy of using voxel-level disagreement to drive human annotator attention to specific regions of images has also been used with deep ensembles constructed by bootstrap sample selection (Dolz et al., 2017; Deng et al., 2018). A disciplined framework that defines manual annotation minimization as a linear program is described by Bhalgat (Bhalgat et al., 2018). The authors suggest that mixed supervision where weak annotation using landmarks and bounding boxes is combined with relatively few full annotations could be used to improve segmentation quality. They define an active learning based semi-automatic segmentation technique using Fisher information to optimize manual

segmentation efforts to differentiate tissue type in infant brains (Sourati et al., 2018). Our method is similar in that it is based on a Von Neumann information paradigm, but different in the sense that we operate at a whole scan level. Tangentially related work includes multilevel networks (Zhao et al., 2018; Gaonkar et al., 2016), that are used with one stage detecting a bounding contour while the second stage segments. While these approaches are neither DSS nor active learning, they do reduce the amount of human effort needed for segmentation.

The aforementioned methods have mainly been designed to improve semi-automatic segmentation and improve the throughput of manual segmentation. Hence, the aforementioned literature aims to alleviate manual work by focusing on problematic regions via active learning at the pixel/voxel level. In contrast, we approach subset selection at the subject/patient level rather than a pixel or a patch level. Our work defines and measures uncertainty between segmentations produced by multiple models at a subject level. The driving motivation in this work is to make automated segmentation-based biomarkers a part of the radiological workflow, where majority of the work may be done by the automation, while identifying cases which will need human attention in the clinic, and then using such cases to improve the automation itself. A second aspect which is not addressed widely in previous literature is that of ‘robustness’. If clinical workflow automation is the goal, robustness is as important as accuracy. Note that we define robustness as the ability of a trained model to consistently segment anatomy and quantify it using the variance of the distribution of Dice coefficients comparing automated segmentations and manually generated segmentations over a large dataset of scans. A method which performs consistently, with a slightly lower accuracy is better than a method which segments inconsistently at a high ‘average’ accuracy. The latter method may be non robust in that it may achieve higher accuracy, by segmenting ‘easy’ cases with a high Dice score but generate extremely poor segmentations on a few challenging cases. Our approach selects subsets which lead to the creation of DL models which are both more accurate and more robust than random selection. We study subset selection from a robustness point of view as opposed to an accuracy point of view. This is another philosophical difference between current art and the work proposed here.

1.2.2. Failure Prediction—Failure is a topic of research that has gained wide-attention in deep learning as well as machine learning. Deep learning systems based on convolutional networks can attain human level performance on narrow tasks yet seem to fail due to incomprehensible reasons, while maintaining ‘high confidence’ in the accuracy of prediction (Nguyen et al., 2015; Goodfellow et al., 2014). The problem of quantifying ‘model uncertainty’, that is having the model ‘know’ when it fails has been addressed by the machine learning community in multiple ways. Traditionally, uncertainty estimation is done using Bayesian Neural Networks (Neal, 2012) which aim to learn the distribution of a network’s weight parameters. Theoretically, this can then enable the computation of a distribution over the network outputs and associated uncertainty estimates. However, Bayesian inference is computationally intractable in modern deep neural networks, given their size. Thus, several recent efforts have focussed on approximating Bayesian Neural Networks (BNNs) rather than training them directly. Perhaps, the most notable of these is

the use of Dropout to approximate Bayesian inference (Gal and Ghahramani, 2016). Other notable attempts at approximating BNNs include the use of Stochastic Batch Normalization (Atanov et al., 2019) and Multiplicative Normalizing Flows (Louizos and Welling, 2017). All of these approximations produce uncertainty estimates using a large number of forward passess through the network at runtime. This makes inference computationaly intensive. Deep ensembles (Lakshminarayanan et al., 2017) provide an alternative which computes variance in prediction by training many models and recording the variance of their predictions. Yet this requires inference on many models. Some authors have proposed direct learning for uncertainty estimation to reduce dependence on sampling - a pardigm that obviates the need for sampling. (Kendall and Gal, 2017; DeVries and Taylor, 2018a).

Some of these ideas have permeated to the medical image segmentation lliterature. However, their application has mainly been to predict segmentation quality. Pixel level uncertainty may be estimated using any of the previously described techniques operating under the pretext that image segmentation is a pixel classification task (DeVries and Taylor, 2018b; Jungo and Reyes, 2019). But image segmentation, especially as applied in clinical practice is not a pixel level task but a scan level task. To automate clinical workfow in spine imaging, we require that an image segmentation algorithm should either confidently and correctly segment anatomy on a scan or leave diagnosis to the physician entirely. In this work we present a novel metric which operates at the scan level rather than at the pixel level. Our metric quantifies the degree of disagreement in segmentations produced by several DL models using the maximum eigenvalue of an associated matrix. The matrix is constructed to capture the disagreement amongst multiple deep learning segmentation models. The framework we present can incorporate various ‘segmentation’ specific metrics to generate the disagreement matrix and address the clinically relevant problem of ‘picking out’ scans which might be problematic. This is different from picking out ‘pixels’ where segmentation uncertainty might lie. We validate our approach on actual clinical data and demonstrate its effectiveness.

1.2.3. Model Stability—Model stability is an important related concept from machine learning literature. Model stability is quantified by consistency of model predictions despite perturbations in training data (Yu et al., 2013; Yu and Kumbier, 2019). The concept proposed here uses the ‘inconsistencies’ between trained deep learning models to identify challenging cases in the data. While stability has not been studied in detail in the context of deep learning, early work in machine learning links higher stability to better generalization for a large class of empirical risk minimization algorithms (Bousquet and Elisseeff, 2002). Later, the link between stability and generalization performance was proven for a much larger class of algorithms (Poggio et al., 2004; Kutin and Niyogi, 2012). Consequently, it is natural to prefer stable deep learning models. In this work, we propose a concrete criterion for choosing training data that leads to the creation of more stable deep learning models for medical image segmentation. Based on previous work in machine learning, we can expect these stable algorithms to generalize better as well.

1.3. Contributions

The main contribution of this work is to propose a novel iterative algorithm for data subset selection and failure prediction in medical image segmentation. Our approach iteratively selects challenging cases from a large dataset and archives models trained on cases selected in each iteration to generate an ensemble of deep learning models. In the next iteration, challenge cases are selected based on the degree of disagreement between all models. The degree of disagreement is defined by the maximum eigenvalue of a matrix whose entries are the Dice scores comparing segmentations generated by different models in the ensemble. We discuss how this measure is closely connected to the Von Neumann information metric and validate the proposed algorithm in clinical MRI segmentation tasks related to the spine. In broad strokes, the proposed algorithm can be seen as an extension of the query-by-committee framework (Seung et al., 1992) to medical image segmentation using Von Neumann Information metric. Using spinal canal and intervertebral disk segmentation on magnetic resonance imaging (MRI), we validated our algorithm. Our experiments show that our algorithm:

1. Chooses a subset of ‘challenging’ cases for initial training
2. Yields trained deep learning models more robust and more accurate than models trained using random selection
3. Accurately identifies entire scans in the data, which are challenging with respect to the defined segmentation task, thus enabling failure prediction

Our work presents a new way to select training data for creating novel segmentation models using deep learning. It also presents a systematic approach to identify scans that are most likely to require human attention by preempting algorithmic failure. These are fundamental challenges in medical image segmentation and addressing them makes deep learning based segmentation both more attractive and defensible for deployment in clinical workflows.

2. Methods

The central aim of the investigations presented here is to convince the reader of the value of our novel algorithmic framework for data subset selection and failure prediction in deep learning based medical image segmentation. Normally, large annotated data sets are thought of as prerequisites for training deep learning methods (Greenspan et al., 2016). In this work, we show that data selection using our framework can help create robust and accurate deep learning models with fewer data. Further, we show that with our algorithmic framework, it is possible to preemptively identify scans where a deep learning model will fail.

2.1. Data collection and preprocessing

The data used as a part of this work was obtained by querying the University of California at Los Angeles (UCLA) PACS for individuals who had undergone any form of spine imaging using the corresponding CPT (Current Procedural Terminology) codes (Terminology, 1970) corresponding to lumbar MRI. The search yielded a large number of accession numbers, of which we selected cases for the purposes of experiments detailed here. This data was obtained under the IRB 16–000196. Images were downloaded from PACS, anonymized and

resampled in the axial or sagittal plane to 256×256-px. Subsequently, each image was converted to the NIFTI (Li et al., 2016; Larobina and Murino, 2014) format and linearly histogram matched to a template image using the SimpleITK (Lowekamp et al., 2013) package. Template image intensities were scaled to lie between a maximum of 1 and a minimum of 0. Linear histogram matching ensured that the same was true of each image used in this study.

2.2. Manual segmentation

Manual segmentation of spinal canals was performed by two medical students using ITK-SNAP (Yushkevich et al., 2006) and validated by an attending physician. The manual segmentation data were used as ground truth for all experiments presented here. The tasks we focus on consists of image segmentation of spinal canals on 200 axial lumbar MRI scan series and intervertebral disks on 100 sagittal lumbar MRI scans data. We have previously published (Gaonkar 2019a, Gaonkar 2019b) on this task and enumerated challenges involved in the process. Examples of intervertebral disk segmentations and spinal canal segmentations are shown in Figure 2.

2.3. Model architecture and parameterization

We use a standard model architecture called the residual U-Net. The U-Net which was first proposed for cellular image segmentation, has become a standard methodology for medical image segmentation (Ronneberger et al., 2015). It was further modified by the addition of residual layers in (Zhang et al., 2018). For experiments presented here, we use the architecture shown in Figure 1. Implementation used the Keras (Chollet et al., 2015) interface to the Tensorflow (Abadi et al., 2016) library.

Our network was designed to operate on 128×128 pixel patches of imaging data. In our experiments, we generated image patches from axial slices extracted from 3D data using 64 px strides for spinal canals. For disks, we used sagittal slices and performed the same patch extraction. Input patches were collected from pre-processed input scan(s) and output patches were collected from corresponding manual segmentation(s). Before training the model, patches extracted from images were augmented by transforming each patch (and the corresponding segmentation) by a randomly picked combination of a translation, rotation, and scaling. Specifically, for each patch, the augmentation algorithm randomly picked an angle between +/ − 20°, a scaling factor between [0.8, 1.2], and x-translation and y-translation limited by +/ − 50 px. For training models used in the EBC selection process, each patch is augmented 20 times since these models are based on small data subsets (Gaonkar et al., 2018). For training models, which are used to validate the EBC selection procedure (see Results), each patch is augmented twice.

2.4. Terminology

- We denote the training set as $\mathcal{T}$, the set of pairs $\{(I_1, S_1),(I_2, S_2), \cdots (I_j, S_j) \cdots (I_N, S_N)\}$ with I_j representing j^{th} patient scan and S_j representing the corresponding segmentation image.

- For any subset $\mathcal{S} \subset \mathcal{T}$, we define $\mathcal{D}_{\mathcal{S}}$ as the deep learning model trained using a chosen deep learning segmentation algorithm denoted by $\mathcal{D}$.
- Further, we denote $S_j^{\mathcal{D}S}$ as the segmentation image obtained by applying the model $\mathcal{D}_{\mathcal{S}}$ to the image $I_j \in \mathcal{T}$.
- The Dice coefficient of overlap is denoted by operator $\Delta(.,.)$, so the Dice coefficient comparing $S_j^{\mathcal{D}S}$ annd S_j would be $\Delta(S_j^{\mathcal{D}S}, S_j)$.
- Further, we remind the reader of the set difference notation. Given sets $\mathcal{A}$ and $\mathcal{B}$, the set $\mathcal{A}\backslash\mathcal{B}$ contains all elements of $\mathcal{A}$ that are not in $\mathcal{B}$.

2.5. The Eigenrank by Committee Algorithm

Figure 3 shows a gestalt representation of the EBC algorithm and the pseudocode below explicitly presents the algorithm itself. As shown in the figure, EBC has an initialization step and an iterative step. The initialization step of EBC closely follows the query-by-committee (QBC) (Seung et al., 1992) paradigm, although with the modification that Dice coefficients used to affirm model agreement are real numbers rather than binary labels. In the initialization phase, the algorithm randomly selects two subsets of size k from $\mathcal{T}$ and trains deep learning segmentation models on them. Then, it compares segmentations generated using one model to the other on the remnant of the training images using the Dice score. Note that this compares segmentations generated by one model to another and does not need ground truth. Images corresponding the lowest 'k' Dice coefficients are used to 'select' the next subset to train on. The second step of EBC is the 'iterative' step, in which we have to compare segmentation results from an increasing number of deep learning models. A direct generalization of the Dice coefficient, if used to compare three or more segmentations, yields a metric which becomes zero if just one model in the ensemble $\{\mathcal{D}_{\mathcal{S}_1}, \cdots \mathcal{D}_{\mathcal{S}_t}\}$ fails, or generates a segmentation which has no overlap with any of the other segmentations. This presents a unique problem which we solve by generating a Dice matrix and using its maximum eigenvalue as a measure of 'disagreement'. At the t^{th} iteration, t models are available, each trained on a distinct $k-$ subset of $\mathcal{T}$. We use these models to construct a $t \times t$ matrix, whose elements are Dice scores comparing segmentations derived from each model with every other model. The principal eigenvalue of this matrix serves as a measure of disagreement among these t models. This principal eigenvalue is representative of the Von Neumann entropy of the Dice matrix, a connection further elucidated in a later section. Note that we allow t to increase to a preset T which represents the total number of iterations in EBC. Selecting images corresponding to the minimum k principal eigenvalues of the Dice matrix takes us to the $t + 1^{th}$ iteration. We formally present the algorithm next.

Initialization

From $\mathcal{T}$ randomly select subsets $\mathcal{S}_1, \mathcal{S}_2 \subset \mathcal{T}$

we require $k = |\mathcal{S}_1| = |\mathcal{S}_2| < < |\mathcal{T}|$

and define $\mathcal{S} \doteq \mathcal{S}_1 \cup \mathcal{S}_2$

Train U-Nets $\mathcal{D}_{\mathcal{S}_1}$ and $\mathcal{D}_{\mathcal{S}_2}$ and define $\mathcal{L} = \{\}$

For all $I_j \in \mathcal{T} \backslash \mathcal{S}$,

– Compute segmentations $S_j^{\mathcal{D}S_1}$ and $S_j^{\mathcal{D}S_2}$

– Compute Dice score $\Delta_j^{S_1, S_2} \doteq \Delta(S_j^{\mathcal{D}S_1}, S_j^{\mathcal{D}S_2})$

– $\mathcal{L} = \mathcal{L} \cup \{\Delta_j^{S_1, S_2}\}$

Use images corresponding to the k–smallest values in $\mathcal{L}$ to construct $\mathcal{S}_3$

Set $\mathcal{S} = \mathcal{S} \cup \mathcal{S}_3$

Iterations

- For t in $\{3, \cdots, T\}$
  - Train model $\mathcal{D}_{\mathcal{S}_t}$ and set $\mathcal{L} = \{\}$
  - Using $\left\{\mathcal{D}_{\mathcal{S}_1}, \cdots \mathcal{D}_{\mathcal{S}_t}\right\}$ compute:
    - For all $I_j \in \mathcal{T}, I_j \notin \mathcal{S}$

      Compute $D_j^{pq} \doteq \Delta_j^{\mathcal{S}p, \mathcal{S}q} = \Delta(S_j^{\mathcal{D}S_p}, S_j^{\mathcal{D}S_q})$

      $\forall p, q \in \{1, \cdots, t\}$

      Define: $\mathbf{D}_j \doteq [D_j^{pq}] \in \mathbb{R}^{t \times t}$

      Compute $\lambda_j^{max} = max\left(eig\left[\mathbf{D}_j\right]\right)$

      $\mathcal{L} = \mathcal{L} \cup \{\lambda_j^{max}\}$
  - Use images corresponding to the k–smallest values in $\mathcal{L}$ to construct $\mathcal{S}_t + 1$
  - Set $\mathcal{S} = \mathcal{S} \cup \mathcal{S}_t + 1$

Output

Output selected data subset $\mathcal{S}$

2.6. EBC with Alternate Metrics

The Dice coefficient is not the only available metric for comparing two segmentations. The Hausdorff distance (Hauss.) and the average surface/contour distance (Surf.) are both established metrics. Unlike the Dice metric, these metrics are distances instead of similarities. Hence, we use the regularized inverse of the Haussdorff distance and the surface distance to drive EBC.

Concretely, when using the Haussdorff distance as a metric, we replace the Dice matrix $\mathbf{D}_j$ with a matrix $\mathbf{U}_j$ defined as:

$$\mathbf{U}_j \doteq [U_j^{pq}] = \left[\frac{1}{\mathscr{H}_j^{pq} + \varepsilon}\right] \in \mathbb{R}^{t \times t} \tag{1}$$

where $\mathscr{H}_j^{pq}$ is the Haussdorff distance between segmentations $S_j^{\mathscr{D}S_p}$ and $S_j^{\mathscr{D}S_q}$.

Similarly, we define the matrix $\mathbf{V}_j$ as:

$$\mathbf{V}_j \doteq [V_j^{pq}] = \left[\frac{1}{\mathscr{A}_j^{pq} + \varepsilon}\right] \in \mathbb{R}^{t \times t} \tag{2}$$

where $\mathscr{A}_j^{pq}$ is the average surface distance between segmentations $S_j^{\mathscr{D}S_p}$ and $S_j^{\mathscr{D}S_q}$. We use an ε value of 0.001 for all our experiments.

We used both these matrices to drive EBC selection using 100 spinal canal cases and the results are tabulated in table 5

2.7. Von Neumann Information- Why EBC works

In EBC, we expect each of the t models to generate unique segmentations and expect $\mathbf{D}_j$ to be at least positive semi-definite (see Appendix for proof) so that we define the associated Von Neumann entropy as:

$$H_j = -\sum_{r=1\cdots t} \lambda_r log(\lambda_r) \tag{3}$$

where $\{\lambda_1 > \lambda_2 > \cdots > \lambda_t\}$ are the ordered eigenvalues of $\mathbf{D}_j$.

In the experiments described here it generally turns out that, $\lambda_1 log(\lambda_1)$ dominates H_j because:

$$\lambda_1 > > \lambda_2, \ \lambda_3, \ \lambda_4, \cdots \lambda_t \tag{4}$$

Thus, we intuit that EBC effectively looks for cases with the highest Von Neumann entropy. To understand why we can generally expect (4) to be true, consider the two extremes of $\mathbf{D}_j = \mathbf{I}$ with $\mathbf{I} \in \mathbb{R}^{t \times t}$ and $\mathbf{D}_j = \mathbf{J}$ with $\mathbf{J} \in \mathbb{R}^{t \times t}$. Both of these cases never occur in practice but correspond to specific fictional scenarios. When $\mathbf{D}_j = \mathbf{I}$, each model agrees with itself but disagrees completely with every other model. When $\mathbf{D}_j = \mathbf{J}$, all models fully agree with each other. In the first case, all eigenvalues of $\mathbf{D}_j$ are unity. In the second case, we can analytically work out $\lambda_1 = t$ and $\lambda_2, \cdots \lambda_t = 0$ and (4) will be true and the maximum eigenvalue dominates the Von Neumann information. In most cases of practical interest, we would expect the various models involved to *mostly but not completely* agree and the maximum eigenvalue remains an effective metric.

The analysis is presented to give the reader a perspective into why the proposed metric for ordering scans works by linking it to a well established information theoretic concept, at least as long as the Dice coefficient is used. However, it is important to note that EBC works well with other potential metrics as well. As shown in the results, this is true for both the Haussdorff distance and the average surface distance. It is important to note that the definition of the matrices $\mathbf{U}_j$ and $\mathbf{A}_j$ also leads to their being positive semi-definite. Although, this cannot be explicitly proven, given the definitions of $\mathbf{U}_j$ and $\mathbf{V}_j$, it is easy to see that their diagonal elements are bound to be constant because $\mathscr{H}_j^{pp} = \mathscr{A}_j^{pp} = 0$ which makes:

$$U_j^{pp} = V_j^{pp} = \frac{1}{\varepsilon} \tag{5}$$

As, the diagonal is constant, we may apply Conjecture 1 from (Nader et al., 2019), if we can show that:

$$U_j^{pq} + U_j^{qr} \geq U_j^{rp} + \frac{1}{\varepsilon} \tag{6}$$

$$V_j^{pq} + V_j^{qr} \geq V_j^{rp} + \frac{1}{\varepsilon} \tag{7}$$

This is easy to enforce by choosing:

$$\varepsilon < < \mathscr{H}_j^{pq} \tag{8}$$

$$\varepsilon < < \mathscr{A}_j^{pq} \forall\, j, p, q \tag{9}$$

In our case, we chose $\varepsilon = 0.001$ while $\mathscr{H}_j^{pq} > 1$ and $\mathscr{A}_j^{pq} > 0.1$. Thus both $\mathbf{U}_j$ and $\mathbf{V}_j$ are generally positive semi-definite with properties similar to $\mathbf{D}_j$. This explains the effectiveness of both metrics in EBC (see results).

2.8. EBC for Failure Prediction

The intuition that drives EBC also provides a framework for failure prediction. If multiple models lack strong agreement over segmenting a particular scan, such a case is best referred to a human expert. This process closely mimics what human trainees do. The theory behind EBC is based on a framework that can quantify how much a group of image segmentation models disagree on a particular scan. Thus, it can be used to select scans which are likely to challenge deep learning based anatomy segmentation, and refer such cases to human experts. This is particularly important if we are to ever deploy machine learning based techniques to the clinic. It is no secret that even the best machine learning based segmentation can fail on a particularly difficult case. Cases which contain hitherto unseen pathology or atypical

anatomy are prime examples of such failure. Such failure is accepted from students in training. Even experts consult each other when faced with a challenging case. During training, some physicians can often identify cases where their own judgement may not be accurate. These cases are most often refer upto their mentors. While this ability to identify errors is impressive, it is imperfect and even a senior attending physician might make a mistake confidently. However, it shows us that humans associate a measure of confidence with each diagnostic judgement. Yet deep learning algorithms, when naively trained, do not replicate this ability. The maximum eigenvalue criterion used in EBC provides a simple measure of case difficulty. The more multiple models agree upon a case the higher the eigenvalue and the higher the confidence that such a case will be segmented well. This unlocks a potential deployment workflow where the algorithm performs adequately on cases which it is 'confident' about and refers the more complex cases to human experts. We have illustrated this in Figure 4. Furthermore, since EBC quantifies degree of model disagreement on the basis of the maximum eigenvalue of a symmetric positive definite matrix which is a continuous real variable, it could also be used to prioritize workflows for clinicians themselves.

3. Results

We present both quantitative and qualitative results comparing residual U-Net models trained using data subsets selected with EBC, to comparable datasets selected using random sampling. We use spinal canal segmentation datasets with 200, 150 and 100 MR scan series segmented by physicians and run EBC with $k = 2$, $k = 3$ and $k = 4$. While $k = 2, 3, 4$ may seem small for training a deep neural network - we use heavy data augmentation to make training plausible. The data is augmented by adding random rotations, translations and data flips 20 times per image (Gaonkar et al., 2018).

3.1. EBC by the iterations

The first set of experiments used 200 axial T2-MRI scans of the lumbar spine, on which manual segmentations of the spinal canal are available. The scans used were randomly selected from a clinical imaging database. They contained artifacts due to variation in acquisition, pathology, and metallic implants and surgical hardware often used in treating spine related conditions. We expect a robust segmentation algorithm to achieve accurate segmentation despite the presence of these artifacts. Thus, a robust algorithm will have both high average Dice score and a lower standard deviation in Dice scores. The more the robustness, the better the applicability to a clinical scenario.

At each iteration we train a residual U-Net model using the data selected. We use the model to segment all remaining cases out of the 200 and report the mean Dice coefficient and the standard deviation of the distribution of these Dice coefficients. At each iteration we also randomly sample a corresponding number of cases randomly - train a residual U-Net on the randomly selected data and use this model to segment the remaining cases.

Thus, for example at iteration $t = 5$, EBC run with $k = 3$ will have cumulatively selected $5 \times 3 = 15$ scans. A residual U-Net trained with these 15 scans is used to segment the remaining 200–15 = 185 scans which exclude the cases selected by EBC. Note that during this training

process patches extracted from the images are augmented only twice as opposed to 20 times during the EBC selection process. Dice coefficients are then computed for each of those 185 scans and the mean and standard deviation of these coefficients is tabulated.

To run a comparative analysis, we also randomly sample 15 cases from the original 200, train a residual U-Net using those 15 cases and compute Dice coefficients on the remaining 185 scans which exclude the cases selected by random sampling only. The mean and the standard deviation of these Dice coefficients is recorded and compared to the corresponding values for EBC using a t-test for means and a Bartlett's test for variances. We have presented these results for spinal canal segmentation using a dataset of 200 axial T2-MRI in Table 1, for iterations $t = 5$ to $t = 7$.

3.2. EBC for on datasets of various sizes

The first set of experiments used $|\mathcal{T}| = 200$, $|\mathcal{T}| = 150$ and $|\mathcal{T}| = 100$ axial T2-MRI scans of the lumbar spine, on which manual segmentations of the spinal canals are available. The scans used were randomly selected from a clinical imaging database.

They contained artifacts due to variation in acquisition, pathology, metallic implants and surgical hardware used in treating spine related conditions. We expect a robust segmentation algorithm to achieve accurate segmentation despite the presence of these artifacts. Thus, a robust algorithm will have both high average Dice score and a lower standard deviation in Dice scores. The more the robustness, the better the applicability to a clinical scenario. We ran EBC for 7 iterations with $k = 3$, thus choosing $7 \times 3 = 21$ scans to train a residual U-Net. We also selected 21 scans randomly and trained a separate residual U-Net using the randomly sampled subsets. For each dataset containing $|\mathcal{T}|$ scans, the trained residual U-Nets (both EBC based and random sampling based) were used to segment spinal canals on the remaining, 179, 129 and 79 scans - and Dice coefficients comparing residual U-Net segmentations to human generated segmentations were computed. The mean of such Dice coefficients as well as their standard deviations are presented in Table 2. Bartlett's test are use to compare standard deviations and t-tests to compare means. These statistical tests indicate a significant increase in accuracy and decrease in standard deviation for the Dice score distributions achieved by the algorithm trained on EBC data subsets as compared to those trained on random sampling data subsets. We also observe that EBC seems to be more effective with larger datasets.

3.3. EBC characterization using different k values

To understand how the choice of k changes the performance of EBC we ran the algorithm for 7 iterations with $k = 2$, $k = 3$ and $k = 4$ using the dataset of $|\mathcal{T}| = 150$ scans with segmented spinal canals. Thus, for $k = 2$, $k = 3$ and $k = 4$, residual U-Net models were trained using 14, 21 and 28 scans respectively. These models were used to segment the remaining 136, 129 and 122 cases respectively. Comparable analyis was done using residual U-Nets trained on randomly sampled datasets as well. The average Dice scores and the associated standard deviations are recorded in Table 3. The table shows that models trained on data subsets picked using EBC are significantly more accurate and robust as compared to models trained using data subsets picked using random sampling for all values of k.

3.4. EBC for data subset selection using a different anatomy

All the experiments presented so far have utilized datasets containing axial MR images and corresponding segmentations of spinal canals. We applied EBC to a dataset containing 103 sagittal MR scans wherein intervertebral disks had been segmented. Table 4 presents a comparison between the performance of a U-Net trained on 21 scans (7- iterations) selected by EBC and one trained on a comparable data subset picked using random sampling. The mean Dice score achieved by the residual U-Net trained on EBC subset continues to be signifcantly better than the mean Dice score achieved by the residual U-Net trained on the randomly sampled subset, despite the change in anatomy targeted by the segmentation.

3.5. EBC characterization using metrics other than the Dice coefficient to compare segmentations

We use the Hausdorff distance and the average surface distance to evaluate the effectiveness of models trained using EBC based selection as outlined in section 2.6. The experiment we conducted used 1) a dataset of 100 axial MRI scans with spinal canal segmentations and 2) a dataset of 103 sagittal MRI scans with intervertebral disk segmentations. We ran EBC to 7 iterations in both datasets with $k = 3$ and with matrices $\mathbf{D}_j$, $\mathbf{U}_j$ as well as $\mathbf{V}_j$ used to drive selection. The results are presented in Table 5. It can be seen that EBC based data subset selection increases average segmentation quality when the metric used is either the Dice coefficient or the average surface distance. However, the Hausdorff distance does not replicate this performance. The use of the raw Hausdorff distance as a metric in EBC leads to relatively higher standard deviations and lower accuracies. However all metrics lead to some improvement over random selection, as evaluated using the Dice coefficient. While a thorough investigation of each metric is out of scope in the present manuscript, we note that the proposed technique remains applicable for these three common metrics used to compare image segmentations.

3.6. EBC - Qualitative visualization

To help the reader gain a qualitative picture of how EBC selects cases, three cases selected by the first iteration of an EBC run - on a dataset of 100 axial T2-MRI scans with $k = 3$ is illustrated in figure 5. Three cases selected via random sampling are presented alongside for comparison. It can be seen from the Figure 5 that EBC selects cases which are much more complex as compared to random sampling. One of these cases has abnormally scoliotic pathology, the second has screws, and the third has a relatively lower contrast, perhaps due to an MRI acquisition issue. By contrast, the variation in both intensity, pathology, and instrumentation within cases picked randomly is distinctively lower. Similarly, figure 6 presents 3 cases chosen by the first iteration of EBC run with $k = 3$ on the dataset of 103 sagittal T2-MRI scans. Again scans selected by EBC contain either instrumentation or pathology. Scans selected randomly do not demonstrate these problems.

3.7. EBC - Failure Characterization

EBC can serve as a method for predicting ‘failure’ of deep learning models. In this section, we present an experiment which explores this possibility. This experiment uses 1) the dataset containing 100 axial lumbar MRIs with spinal canal segmentations and 2) the dataset

containing 103 sagittal lumbar MR scans with intervertebral disk segmentations. We train residual U-Net models on a random subset of 15 scans selected from each of these datasets. Let's denote these models as D_{canal} and D_{disk} for the respective anatomies. This leaves 85 axial MR images with spinal canal segmentations and 87 sagittal MR images with intervertebral disk segmentations for the rest of the experiment. We eliminate scans from these 85 and 87 in batches of $k = 3$ using EBC.

For the canal segmentation data after the first iteration, 82 validation cases remain and 3 cases are eliminated. After two iterations, 6 cases are eliminated and 79 cases remain and so on. Similarly, for the disk segmentation data, after the first iteration, 85 validation cases remain and 3 cases are eliminated. After two iterations 6 cases are eliminated and 82 cases remain and so on. We use D_{canal} and D_{disk} to segment both: the set of eliminated cases and the set of remaining validation cases - at each iteration for both anatomies. Dice score means and standard deviations across these sets are documented for canal segmentation in Table 6 and for disk segmentation in Table 7.

Both tables indicate that as EBC eliminates the complex cases, the average Dice score on the remaining cases increases and the standard deviation of the Dice scores decreases. This indicates that EBC can preemptively detect cases which D_{canal} and D_{disk} will find challenging - even though the data they were trained on were not part of the evaluations presented in the Tables 6 and 7.

4. Discussion

We have presented our algorithm from a utilitarian point of view. In this section, we first present intuitions which drove the design of EBC. Then, we discuss alternative metrics which could be used in EBC, in place of the eigenvalue measure proposed. We also discuss in detail why we consider EBC a better alternative to traditional data subset selection in medical image segmentation. We also highlight how EBC is related to QBC and note some of the mathematical problems which emerge from our experiments.

4.1. Relationship to Query-by-Committee

The query-by-committee (QBC) framework of active learning, first presented by Seung (Seung et al., 1992), motivated EBC. QBC operates on a framework similar to EBC, where multiple models are trained on current labels, and new candidates for training are picked based on where "the committee" disagrees the most. QBC was first proposed from an information theoretic perspective , and further developed in it (Freund et al., 1997). Later, other authors extended QBC with kernels (Gilad-Bachrach et al., 2006) and studied its theoretical properties (Buus et al., 2003). The premise of QBC-based DSS is that a data instance in which two machine learning models label differently, is more informative for the training subset. In the standard classification setting where labels are either binary or discrete, this premise is straightforward to apply. However, in deep learning based medical image segmentation, applying QBC directly presents several challenges unless one is applying it at a pixel level where standard deviation of segmentation intensities provide a simple metric to quantify disagreement. Applying QBC at the scan level requires that

comparisons of outputs of multiple models are made at the global segmented image level. EBC presents one paradigm in which this may be done.

4.2. Comparison to traditional data subset selection

Traditional data subset selection methods (as well as active learning) algorithms have most often been designed for either classification or regression problems. Their applicability in medical image segmentation is thus fairly limited. Further most traditional data subset selection techniques tend to operate independently of the algorithm. For instance, facility location based submodular dataset selection, would select the same subset whether we were segmenting a spinal canal or spinal vertebrae or some other anatomical structure. EBC, on the other hand, has the potential adapt selection and selection strategy to the specific anatomical substructure of interest. This is true of active learning in general. Yet, the majority of literature on active learning for medical image segmentation focuses on identifying variance between models at a local pixel/voxel level rather than a global entire image level. It is unclear whether such disagreement at the pixel or patch level translates to overall disagreement at the scan level. Moreover, it is easy to imagine scenarios where local disagreement does not translate to global disagreement. For instance, the existence of an unusually bright pixel, may cause certain models to fail locally causing local variance, yet globally a single pixel being mis-segmented hardly matters. To the best of our knowledge EBC based data subset selection is unique in quantifying and utilizing inter-model variance on an full image basis for data subset selection and active learning. This global variance quantification using Von Neumann entropy places EBC uniquely in the space of active learning methods used in medical image analysis.

4.3. A note on model selection

In this work, we have used a specific instance of a residual U-Net model to both construct and validate our framework. Perhaps a completely different model, patch generation and data augmentation scheme could be used. As such, hyperparameter optimization, learning rate optimization, batch normalization, architecture optimization, and all the other techniques which can improve deep networks could be used to create better models. Hyperparameter selection is an area of research unto itself. Our aim in this work is not to focus on model optimization, but rather to highlight the effectiveness of EBC for data subset selection and failure prediction, rather than delving into parameter or network selection theory. Hence, we have used a relatively straightforward architecture with fixed hyperparameters, patch generation, and augmentation schema.

4.4. Mathematical aspects

It is useful to understand the positive semi-definiteness of $\mathbf{D}_j \succeq 0$ from a geometric standpoint. Specifically, we explore the implications of this for comparing three segmentations to each other.

In the case presented by Figure 7, it is possible to visualize why $\mathbf{D}_j$ might be positive semi-definite for the relative of three models shown in Figure 7. In the case of three models, if two models agree with the third one, they cannot disagree among themselves. This unviable situation would lead to a non-positive definite matrix

$$\mathbf{D}_j = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

with eigenvalues [2.42, 1, −0.42]. Thus, the intersection of the cube and the cone of positive semi-definite $\mathbf{D}_j$ forms a region of space where feasible Dice matrices arise. $\mathbf{D}_j \succeq 0$ also leads to an elegant relationship between Dice coefficients arising out of mutual comparisons of segmentations generated by a trio of deep learning models. If D_j^{pq}, D_j^{qr} and D_j^{rp} are Dice scores comparing segmentations generated on image I_j by a trio of models $\mathcal{D}_{\mathcal{S}_p}$, $\mathcal{D}_{\mathcal{S}_q}$, $\mathcal{D}_{\mathcal{S}_r}$. Then,

$$[D_j^{pq}]^2 + [D_j^{qr}]^2 + [D_j^{rp}]^2 - 1 < 2D_j^{pq}D_j^{qr}D_j^{rp} \tag{10}$$

This follows from the fact that the Schur complement of positive semi-definite matrix is positive semi-definite under the appropriate conditions. This can be used as an efficient testing criterion for simulating viable Dice matrices. We use it to test the following conjecture:

Conjecture: As the number of models t increases, the Shannon information of the maximum eigenvalue $\mathbf{D}_j$ *dominates the Shannon information of all other eigenvalues.*

If $\lambda 1 > \lambda 2 > \lambda 3 \cdots \lambda t$ were sorted eigenvalues of $\mathbf{D}j$ then this conjecture can be expressed as:

$$\lim_{t \to \infty} \frac{\lambda_1 log(\lambda_1)}{\sum_{r=1}^{t} \lambda_r log(\lambda_r)} = 1 \tag{11}$$

In Figure 8, we provide the results of simulations performed using randomly generated positive semi-definite matrices confirming diagonal elements being equal to ‘1’ and off-diagonal elements modeled as $1 - \delta^{pq}$. Trios of $1 - \delta^{pq}$ are constrained by (10). δ^{pq} is randomly selected from the interval $[0, \epsilon]$ with ϵ set to various values. These simulations support the conjecture and this conjecture is the link connecting EBC and information theory. Specifically it justifies the use of the maximum eigenvalue measure. Future work to ascertain the exact conditions under which it remains true, will be necessary to understand the limits of the proposed algorithm.

5. Conclusion

In conclusion, we have proposed a method for addressing both data subset selection and failure prediction for deep learning based image segmentation. We have also demonstrated the effectiveness of the proposed paradigm in two medical image analysis datasets. Our technique can help select subsets of images from large databases, in a manner such that accurate and more importantly, ‘robust’ deep neural networks can be trained for anatomical

segmentation. It can also accurately identify challenging cases from a given dataset, where human attention is most likely needed. This gives deep learning based segmentation algorithms the ability to prioritize challenging cases within automated clinical image analysis workflows, thereby enabling better integration between human and machine in the future.

Acknowledgment

We thank the National Institutes of Health for support through the grant R21EB026665 and The University of California Los Angeles for supporting this research.

7.: Appendix

Proof of Positive semi-definiteness of the Dice matrix D_j

The Von Neumann entropy for the symmetric positive definite matrix and is defined to be the sum of the Shannon entropy of its eigenvalues. Thus, in our paper the Dice matrix $\mathbf{D}_j$ can be proved to be positive semi-definite (Nader et al., 2019). The proof follows from the fact that we can express:

$$D_j^{pq} = \frac{2\mathbf{s}_p \cdot \mathbf{s}_q}{|\mathbf{s}_p| + |\mathbf{s}_q|} \tag{12}$$

where

$$\mathbf{s}_p, \mathbf{s}_q \in \{0,1\}^{|I_j|} \tag{13}$$

are the vectorized representations of images $S_j^{\mathcal{DS}_p}$, $S_j^{\mathcal{DS}_q}$. Thus, the Dice matrix itself can be thought of as a Hadamard product of an inner product matrix and a Cauchy matrix. That is:

$$\mathbf{D}_j = 2\mathbf{K}_j \circ \mathbf{C}_j \tag{14}$$

where we define:

$$K_j^{pq} = \mathbf{s}_p \cdot \mathbf{s}_q \tag{15}$$

and

$$C_j^{pq} = \frac{1}{|\mathbf{s}_p| + |\mathbf{s}_p|} \tag{16}$$

Thus, $\mathbf{K}_j$ is an inner product matrix - which are always positive semi-definite. The Cauchy matrix is positive semi-definite (Bhatia, 2009) because it can be expressed as an inner product matrix in Hilbert space:

$$\frac{1}{|\mathbf{s}_p| + |\mathbf{s}_q|} = \int_0^\infty e^{-(|\mathbf{s}_p| + |\mathbf{s}_q|)t} dt \tag{17}$$

and

$$\int_0^\infty e^{-(|\mathbf{s}_p| + |\mathbf{s}_q|)t} dt = \int_0^\infty e^{-|\mathbf{s}_p|t} \cdot e^{-|\mathbf{s}_q|t} dt \tag{18}$$

Given that both $\mathbf{K}_j$ and $\mathbf{C}_j$ is positive semi-definite, the Schur product theorem then ensures that $\mathbf{D}_j$ is positive semi-definite as well. The proof presented here is based on previous work by Nader (Nader et al., 2019) and Bhatia (Bhatia, 2009), which the reader should refer to for more details.

References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al., 2016 Tensorflow: a system for large-scale machine learning., in: OSDI, pp. 265–283.

Atanov A, Ashukha A, Molchanov D, Neklyudov K, Vetrov D, 2019 Uncertainty estimation via stochastic batch normalization, in: International Symposium on Neural Networks, Springer. pp. 261–269.

Bhalgat Y, Shah M, Awate S, 2018 Annotation-cost minimization for medical image segmentation using suggestive mixed supervision fully convolutional networks. arXiv preprint arXiv:1812.11302.

Bhatia R, 2009 Positive definite matrices. volume 24 Princeton university press.

Bousquet O, Elisseeff A, 2002 Stability and generalization. Journal of machine learning research 2, 499–526.

Buus S, Lauemøller S, Worning P, Kesmir C, Frimurer T, Corbet S, Fomsgaard A, Hilden J, Holm A, Brunak S, 2003 Sensitive quantitative predictions of peptide-mhc binding by a query by committeeartificial neural network approach. Tissue antigens 62, 378–384. [PubMed: 14617044]

Chollet F, et al., 2015 Keras.

Das S, Wong WK, Dietterich T, Fern A, Emmott A, 2016 Incorporating expert feedback into active anomaly discovery, in: 2016 IEEE 16th International Conference on Data Mining (ICDM), IEEE. pp. 853–858.

Deng Y, Sun Y, Zhu Y, Zhu M, Han W, Yuan K, 2018 A strategy of mr brain tissue images' suggestive annotation based on modified u-net. arXiv preprint arXiv:1807.07510.

DeVries T, Taylor GW, 2018a Learning confidence for out-of-distribution detection in neural networks. arXiv preprint arXiv:1802.04865.

DeVries T, Taylor GW, 2018b Leveraging uncertainty estimates for predicting segmentation quality. arXiv preprint arXiv:1807.00502.

Dolz J, Desrosiers C, Wang L, Yuan J, Shen D, Ayed IB, 2017 Deep cnn ensembles and suggestive annotations for infant brain mri segmentation. arXiv preprint arXiv:1712.05319.

Freund Y, Seung HS, Shamir E, Tishby N, 1997 Selective sampling using the query by committee algorithm. Machine learning 28, 133–168.

Gal Y, Ghahramani Z, 2016 Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, pp. 1050–1059.

Gaonkar B, Bui A, Brown M, Macyszyn L, 2018 Extreme augmentation: Can deep learning based medical image segmentation be trained using a single manually delineated scan? arXiv preprint arXiv:1810.01621.

Gaonkar B, Hovda D, Martin N, Macyszyn L, 2016 Deep learning in the small sample size setting: cascaded feed forward neural networks for medical image segmentation, in: Medical Imaging 2016: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 97852I.

Gaonkar B, Villaroman D, Beckett J, Ahn C, Attiah M, Babayan D, Villablanca JP, Salamon N, Bui A, Macyszyn L. Quantitative Analysis of Spinal Canal Areas in the Lumbar Spine: An Imaging Informatics and Machine Learning Study. AJNR Am J Neuroradiol. 2019 9;40(9):1586–1591. doi: 10.3174/ajnr.A6174. [PubMed: 31467240]

Gaonkar B, Beckett J, Villaroman D, Ahn C, Edwards M, Moran S, Attiah M, Babayan D, Ames C, Villablanca JP, Salamon N. Quantitative analysis of neural foramina in the lumbar spine: an imaging informatics and machine learning study. Radiology: Artificial Intelligence. 2019 3 6;1(2):180037.

Gilad-Bachrach R, Navot A, Tishby N, 2006 Query by committee made real, in: Advances in neural information processing systems, pp. 443–450.

Goodfellow IJ, Shlens J, Szegedy C, 2014 Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Greenspan H, Van Ginneken B, Summers RM, 2016 Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Transactions on Medical Imaging 35, 1153–1159.

Jungo A, Reyes M, 2019 Assessing reliability and challenges of uncertainty estimations for medical image segmentation. arXiv preprint arXiv:1907.03338.

Kendall A, Gal Y, 2017 What uncertainties do we need in bayesian deep learning for computer vision?, in: Advances in neural information processing systems, pp. 5574–5584.

Kutin S, Niyogi P, 2012 Almost-everywhere algorithmic stability and generalization error. arXiv preprint arXiv:1301.0579.

Lakshminarayanan B, Pritzel A, Blundell C, 2017 Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in neural information processing systems, pp. 6402–6413.

Larobina M, Murino L, 2014 Medical image file formats. doi:10.1007/s10278-013-9657-9.

Li X, Morgan PS, Ashburner J, Smith J, Rorden C, 2016 The first step for neuroimaging data analysis: DICOM to NIfTI conversion. Journal of Neuroscience Methods 264, 47–56. [PubMed: 26945974]

Louizos C, Welling M, 2017 Multiplicative normalizing flows for variational bayesian neural networks, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org. pp. 2218–2227.

Lowekamp BC, Chen DT, Ibáñez L, Blezek D, 2013 The design of simpleitk. Frontiers in neuroinformatics 7, 45. [PubMed: 24416015]

Nader R, Bretto A, Mourad B, Abbas H, 2019 On the positive semi-definite property of similarity matrices. Theoretical Computer Science 755, 13–28.

Neal RM, 2012 Bayesian learning for neural networks. volume 118 Springer Science & Business Media.

Nguyen A, Yosinski J, Clune J, 2015 Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 427–436.

Poggio T, Rifkin R, Mukherjee S, Niyogi P, 2004 General conditions for predictivity in learning theory. Nature 428, 419. [PubMed: 15042089]

Ronneberger O, Fischer P, Brox T, 2015 U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention {\textendash} MICCAI 2015. Springer, Cham, Cham, pp. 234–241.

Rubens N, Elahi M, Sugiyama M, Kaplan D, 2015 Active learning in recommender systems, in: Recommender systems handbook. Springer, pp. 809–846.

di Scandalea ML, Perone CS, Boudreau M, Cohen-Adad J, 2019 Deep active learning for axon-myelin segmentation on histology data. arXiv preprint arXiv:1907.05143.

Schreiber J, Bilmes J, Noble WS, 2019 apricot: Submodular selection for data summarization in python. arXiv preprint arXiv:1906.03543.

Settles B, 2009 Active learning literature survey Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

Seung HS, Opper M, Sompolinsky H, 1992 Query by committee, in: Proceedings of the fifth annual workshop on Computational learning theory, ACM. pp. 287–294.

Sourati J, Gholipour A, Dy JG, Kurugol S, Warfield SK, 2018 Active deep learning with fisher information for patch-wise semantic segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 83–91.

Terminology CP, 1970 Current procedural terminology (CPT). JAMA : the journal of the American Medical Association 212, 873–874. doi:10.1001/jama.212.5.873b.

Wei K, Iyer R, Bilmes J, 2015 Submodularity in data subset selection and active learning, in: International Conference on Machine Learning, pp. 1954–1963.

Yang L, Zhang Y, Chen J, Zhang S, Chen DZ, 2017 Suggestive annotation: A deep active learning framework for biomedical image segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer pp. 399–407.

Yu B, Kumbier K, 2019 Three principles of data science: predictability, computability, and stability (pcs). arXiv preprint arXiv:1901.08152.

Yu B, et al., 2013 Stability. Bernoulli 19, 1484–1500.

Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig G, 2006 User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. Neuroimage 31, 1116–1128. [PubMed: 16545965]

Zhang Z, Liu Q, Wang Y, 2018 Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters 15, 749–753.

Zhao Z, Yang L, Zheng H, Guldner IH, Zhang S, Chen DZ, 2018 Deep learning based instance segmentation in 3d biomedical images using weak annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer pp. 352–360.

Zhou ZH, 2017 A brief introduction to weakly supervised learning. National Science Review 5, 44–53.

Highlights

- Eigenrank by Committee reduces the need of training data for deep learning-based segmentation in medical imaging.
- Eigenrank by Committee can alert physicians to a likely segmentation failure when using a deep learning methods.
- Eigenrank by Commitee presents a Von Neumann information based theoretical criterion for quantifying deep model disagreement in image segmentation.

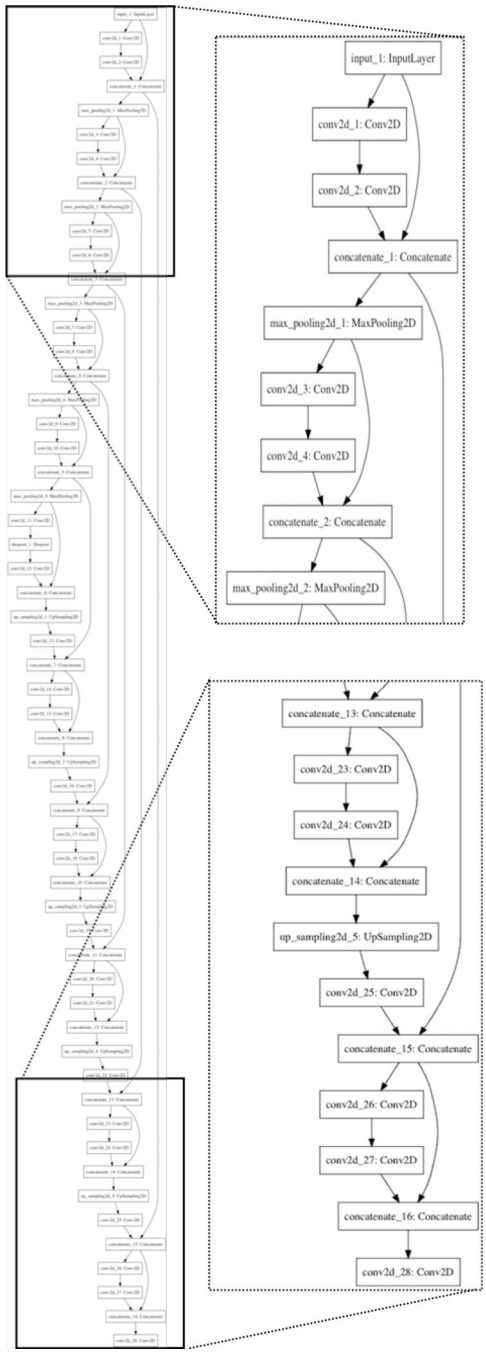


Figure 1:
Residual U-Net model used in our experiments

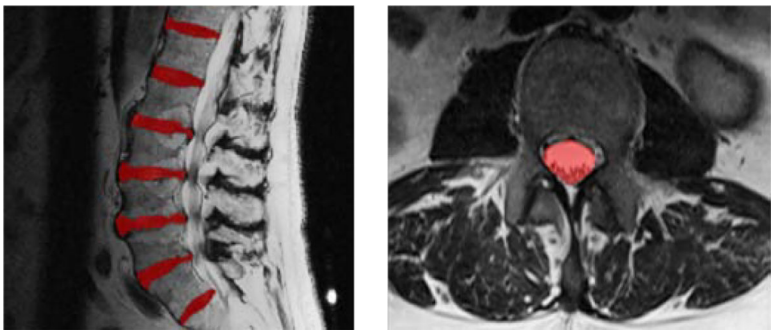


Figure 2:
Illustration of intervertebral disk and spinal canal segmentation

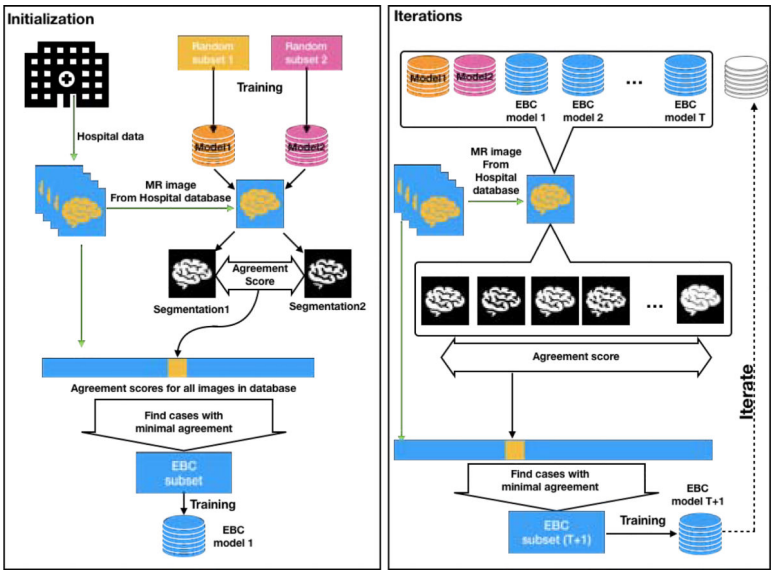


Figure 3:
Illustration of the EBC framework

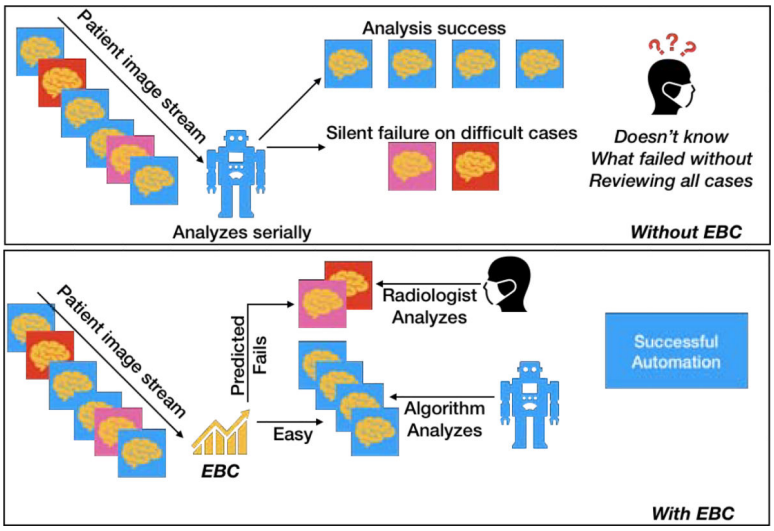


Figure 4:

(Top) Standard machine learning algorithms fail silently. Consequently it is impossible to prioritize physician attention onto cases which are difficult cases. (Bottom) With EBC, there is potential to prioritize cases which are difficult and refer them to human experts and channel the easier cases to an allgoorithm for segmentation.

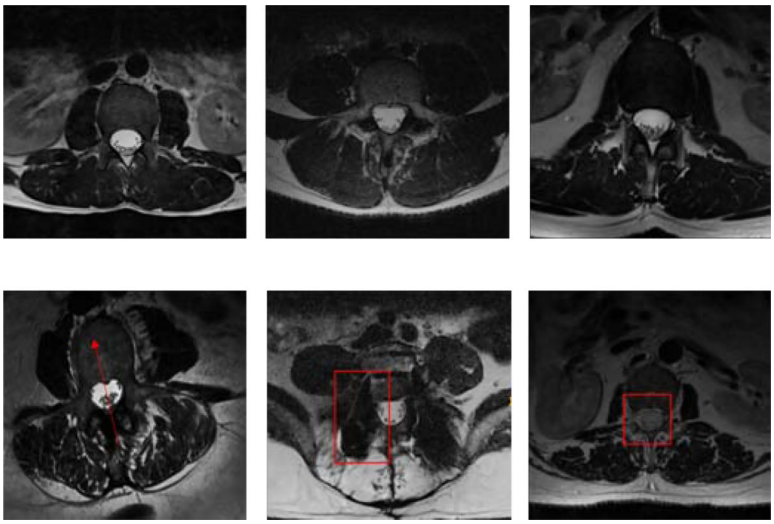


Figure 5:

Top Row Randomly selected subjects **Bottom Row** Subjects selected by the first iteration of the proposed algorithm

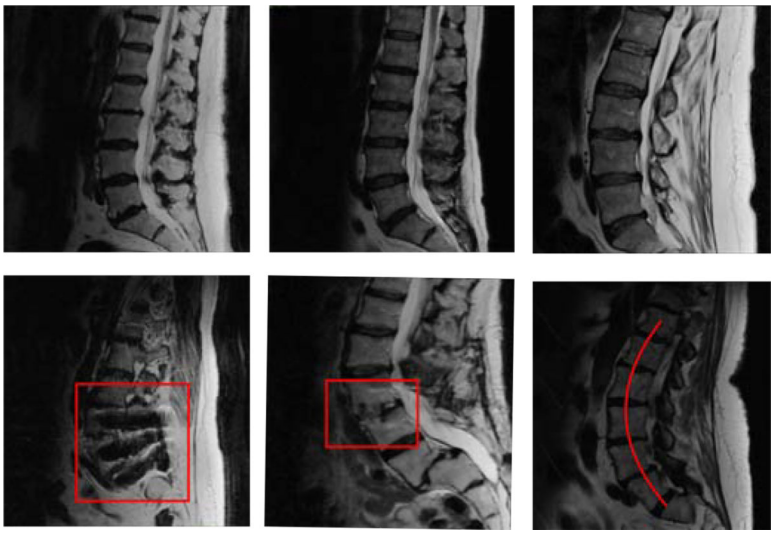


Figure 6:

Top Row Randomly selected subjects **Bottom Row** Subjects selected by the first iteration of the EBC

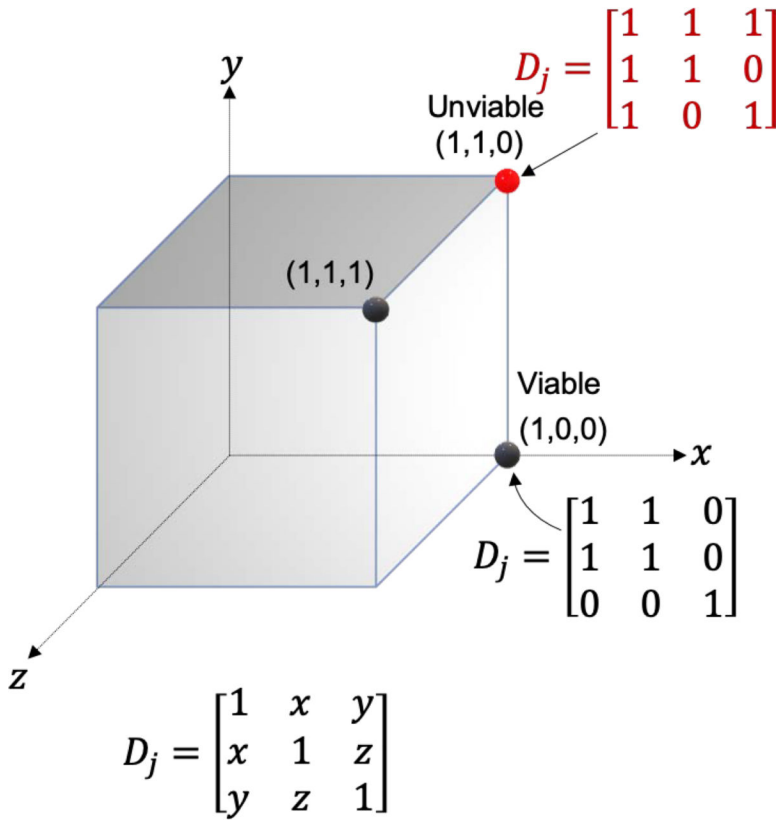


Figure 7:
Viability and unviability of $\mathbf{D}_j$ in comparing three models to each other.

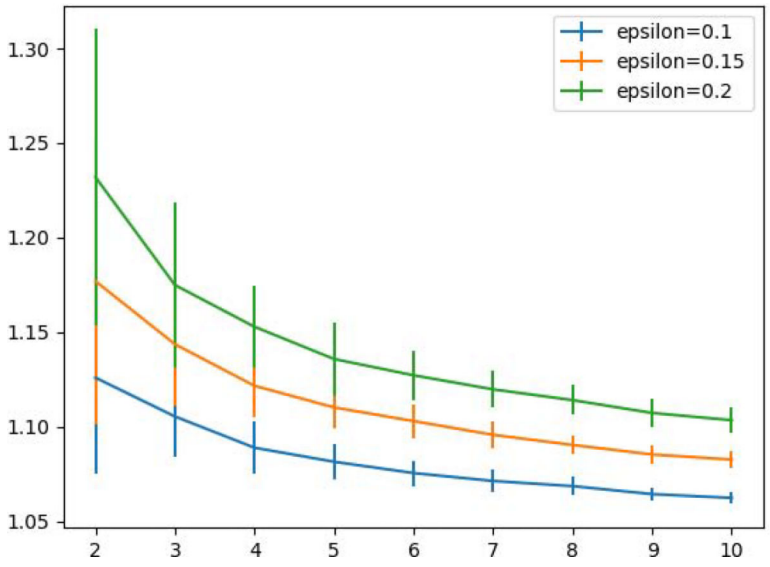


Figure 8:
Why the largest eigenvalue of $\mathbf{D}_j$ suffices as a measure of disagreement for EBC

Table 1:

EBC selection outperforms random selection on spinal canal segmentation. Results comparing EBC and random subset selection on a curated set of lumbar axial MRIs with spinal canals manually segmented for $|\mathcal{T}| = 200$ scans

Iteration	Metric	EBC	Random
7	Mean Dice	0.8062	0.7073
	Stdv Dice	0.0830	0.1831
t-test of means (t= 6.94, p=7.23E-11) Bartletts test of variances (T=99.47, p = 1.99E -23)			
6	Mean Dice	0.8014	0.7049
	Stdv Dice	0.1103	0.1676
t-test of means (t=6.54, p=6.15E-10) Bartletts test of variances (T=30.21, p =3.87 E -8)			
5	Mean Dice	0.8070	0.7050
	Stdv Dice	0.0991	0.1703
t-test of means (t=6.98, p=5.30E-11) Bartletts test of variance (t=50.47, p=1.21E-12)			

Table 2:

Comparing Dice scores for spinal canal generated using training subsets of 21 scans picked using EBC and using random sampling. These subsets were picked from datasets containing $|\mathcal{T}| = 200$, 150 *and* 100 axial MRI scans. The segmentation task used was automated delineation of spinal canals.

	EBC	Random	Statistical testing (t-test for means Bartlett's for variances)
$\|\mathcal{T}\| = 200$			
Mean Dice	0.8062	0.7073	t = 6.94 (p <0.05)
Stdev. Dice	0.0830	0.1831	T = 99.74 (p <0.05)
$\|\mathcal{T}\| = 150$			
Mean Dice	0.7973	0.6981	t = 5.48 (p <0.05)
Stdev. Dice	0.0825	0.1901	T = 80.03 (p <0.05)
$\|\mathcal{T}\| = 100$			
Mean Dice	0.8814	0.8461	t = 2.68 (p <0.05)
Stdev. Dice	0.0372	0.1136	T = 78.13 (p <0.05)

Table 3:

Comparing Dice scores for spinal canal segmentation generated using various values of 'k' in EBC and using the same number of scans picked using random sampling from a data set containing $|\mathcal{T}| = 150$ axial MR scans. The segmentation task was automated delineation of spinal canals.

$\|\mathcal{T}\| = 150$	**EBC**	**Random**	**Statistical testing (t-test for means Bartlett's for variances)**
k=2			
Mean Dice	0.7457	0.6352	t = 5.89 (p <0.05)
Stdev. Dice	0.1128	0.2109	T = 49.89 (p <0.05)
k=3			
Mean Dice	0.7973	0.6981	t = 5.48 (p <0.05)
Stdev. Dice	0.0825	0.1901	T = 80.03 (p <0.05)
k=4			
Mean Dice	0.8301	0.8031	t = 2.15 (p <0.05)
Stdev. Dice	0.0839	0.1201	T = 5.45 (p <0.05)

Table 4:

EBC selection outperforms random selection on intervertebral disk segmentation. Results comparing EBC and random subset selection on a curated set of lumbar sagittal MRIs with intervertebrall disks manually segmented for $|\mathcal{T}| = 103$ scans

Iteration	Metric	EBC	Random
7	Mean Dice	0.8582	0.8425
	Stdv Dice	0.0333	0.0655
t-test of means (t= 2.05, p=0.043) Bartletts test of variances (T=33.06, p = 8.90E-9)			
6	Mean Dice	0.8521	0.8379
	Stdv Dice	0.0329	0.0665
t-test of means (t=2.02, p=0.046) Bartletts test of variances (T=37.11, p =1.11 E-9)			
5	Mean Dice	0.8538	0.8318
	Stdv Dice	0.0371	0.0666
t-test of means (t=3.11, p=0.002) Bartletts test of variance (T=27.11, p=1.91E-7)			

Table 5:

This table shows how Eigenrank run using different metrics for data selection performs as evaluated using each of those metrics. The datasets used contain $|\mathcal{T}| = 100$ axial MRI scans with spinal canal segmentations and $|\mathcal{T}| = 103$ sagittal MRI scans with intervertebral disk segmentations. EBC was run with $k = 3$ for 7 iterations with three metrics and the residual U-Net was trained using the corresponding data subset selections. Evaluations performed using alll three metrics are tabulated.

	Eigenrank (selection metric)			**Random**
	Dice	**Hauss**	**Surf.**	
Validation Metric				
Canal				
Dice	0.88±0.036	0.87±0.053	0.88±0.058	0.84±0.11
Hauss.	69.4±31.7	67.5±29.8	61.8±36.1	75.2±27.8
Surf.	0.92±1.02	0.90±1.05	0.63±0.71	1.45±1.90
Disk				
Dice	0.86±0.033	0.84±0.062	0.86±0.026	0.84±0.065
Hauss.	37.1±25.0	39.2±15.2	33.8±15.8	38.9±26.2
Surf.	0.32±0.22	0.49±0.57	0.33±0.19	0.52±0.71

Table 6:

Using EBC purely as a failure analysis method for spinal canal segmentation, a model is created by training on a left-out set of 15 cases. Out of the remaining 85 validation cases, EBC was used to iteratively remove 'difficult' cases. After each iteration, we compute the mean Dice score of the cases eliminated and of the remaining validation cases

Iteration	For cases eliminated by EBC		For remaining cases	
	Mean Dice	Stdev Dice	Mean Dice	Stdev Dice
1	0.611	0.317	0.835	0.076
2	0.650	0.265	0.839	0.075
3	0.710	0.253	0.840	0.075
4	0.742	0.238	0.840	0.076
5	0.750	0.222	0.842	0.074
6	0.769	0.212	0.840	0.074
7	0.749	0.208	0.851	0.052

Table 7:

Using EBC as a failure analysis method for disk segmentation on sagittal MRI. A model to segment disks in sagittal MRI is created by training on a left-out set of 15 cases. Out of the remaining 87 validation cases, EBC was used to iteratively remove ‘difficult’ cases. After each iteration, we compute the mean Dice score of the cases eliminated and of the remaining validation cases.

Iteration	For cases eliminated by EBC		For remaining cases	
	Mean Dice	Stdev Dice	Mean Dice	Stdev Dice
1	0.681	0.154	0.835	0.040
2	0.739	0.152	0.839	0.040
3	0.773	0.145	0.840	0.040
4	0.789	0.135	0.840	0.040
5	0.804	0.132	0.842	0.040
6	0.805	0.125	0.840	0.038
7	0.809	0.119	0.851	0.037