

# Uncertainty quantification using an energy-based method applied to kidney CT images

Andreea Elena Vântu  
Transilvania University of Braşov  
Siemens SRL  
Braşov, România  
andreea.vantu@siemens.com

Alexandru Mihnea Ion  
Transilvania University of Braşov  
Siemens SRL  
Braşov, România  
mihnea.ion@siemens.com

Daniel Bunescu  
Transilvania University of Braşov  
Siemens SRL  
Braşov, România  
daniel.bunescu@siemens.com

Lucian Mihai Itu  
Transilvania University of Braşov  
Siemens SRL  
Braşov, România  
lucian.itu@siemens.com

## Abstract

Uncertainty quantification is an essential topic in medical image analysis. The purpose of this study is to determine if an energy-based function could represent a proxy for identifying misclassified segmentations. We used a kidney CT image dataset, and we conducted experiments on different types of kidney contours.

## CCS Concepts

• **Software and its engineering**; • **Computing methodologies** → **Image segmentation**; • **Applied computing** → **Health care information systems**;

## Keywords

deep learning, energy-based method, medical images, kidneys, contours, uncertainty

## 1 Introduction

Healthcare represents a domain where Artificial Intelligence (AI) technology has many advantages over traditional methods. In this particular case, since AI based systems generate predictions with a specific uncertainty, the reliability of the AI models or systems needs to be as high as possible. Researchers have studied this aspect and have developed different methods to quantify uncertainties and gain the end user's trust. Several studies show the relevance of post-training techniques to quantify uncertainties. Bayesian models are known for their ability to output uncertainties, and the authors of the paper [5] proposed a Bayesian meta-model that was applied post-hoc. Their method consisted of a Dirichlet distribution that would determine the true posterior distribution. This study [6] demonstrated an application of uncertainty quantification methods on cortical lesions. They developed an equation to determine the uncertainty score of each predicted lesion (1). The identified uncertainties described different features and characteristics of those lesions.

$$LSU = 1 - \frac{1}{M} \sum_{m=1}^{M-1} IoU(L, L^m), \quad (1)$$

where  $L$  represents the lesion and  $L^m$  the  $m^{th}$  sampled prediction,  $m = 0, 1, 2, \dots, M - 1$ . There are still some limitations in uncertainty

quantification for medical images. One limitation is that many studies focus on 2D images, whereas those that imply an application to 3D images are fewer. Another limitation is that fewer studies analyse the relationship between the segmentation quality and the uncertainty of a model. Thus, in this research, we considered an energy-based method as a proxy for identifying poor segmentations, by analysing the correlation between energy scores and Dice loss values.

## 2 Methods

For our approach, we used an energy function to determine uncertainties in a medical image segmentation. This method can be used as an Out-Of-Distribution (OOD) method as stated in paper [4], but according to this paper [1], we can shift the perspective and use OOD methods to identify erroneous segmentations. The origin of the energy function is the concept of Helmholtz free energy in thermodynamics. After extrapolating this concept to deep learning, we obtain:

$$E(x) = -T \cdot \log \sum_i^K e^{f_i(x)/T}, \quad (2)$$

where  $x$  is the input,  $T$  is the temperature parameter,  $K$  represents the number of classes, and  $e^{f_i(x)/T}$  is called the partition function.

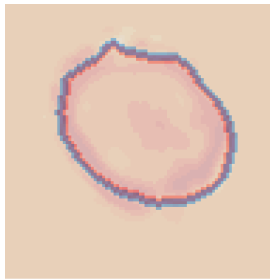
The energy function reflects the model's confidence. High values represent uncertainties, while low values describe strong confidence in the model's reasoning and predictions. For this study, we used two publicly available datasets containing kidney CT scans, KiTS21 [2] and KiTS23 [3]. We used 300 cases from the first dataset to train and validate our model, and retained 200 newly acquired cases from the second dataset as a separate test set. We extracted the regions of interest (ROIs) and used them as inputs for our neural network. We trained a model to segment kidneys. For this task, we used an extension of the U-Net architecture [7], called 3D U-Net. We used the k-fold cross-validation technique to train and evaluate the models and the Dice loss measurement (3) for the evaluation.

$$L = 1 - Dice \quad (3)$$

$$Dice = 2 * IoU \quad (4)$$

We trained the deep learning model on a single NVIDIA GeForce GTX 1080 Ti GPU with 12 GB of memory. For each fold, the model was trained for 50 epochs.

The region most susceptible to uncertainty is the boundary of the kidney. Thus, we focused on the kidney contours and analysed the correlation between the average uncertainty score and the Dice loss value evaluated on the contour. We applied different morphological operations to extract the contours, such as dilation and erosion, and we also increased the thickness of the contour in various ways. We analysed three types of contours (see Fig. 1): outer contour, inner contour, and contextual contour. The last type represents the merged outer and inner contours, to capture more details around the kidney boundary. The width of the contour was changed as follows: increased by one, two, or three voxels.

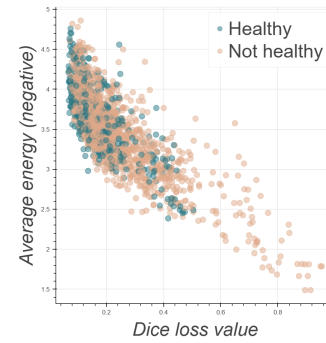


**Figure 1: The overlap of an outer and inner contour placed on an energy mask. The blue voxels represent the outer part of the contour, and the red ones illustrate the inner part of the contour. The contours were placed over the generated energy mask.**

To measure the correlation between the average energy score and the Dice loss value, we calculated the Pearson correlation coefficient and the Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) over the distribution of the results. For the ROC-AUC assessment, we selected the 25<sup>th</sup>, 50<sup>th</sup>, and the 75<sup>th</sup> quartiles as thresholds from the Dice loss value set.

### 3 Results

For our k-fold cross-validation method, k was set to five folds. Upon completion of the cross-validation procedure, we had five models. We chose a suitable epoch for all five models for the experiments session by comparing the mean and standard deviation. The models reached an average Dice loss of 0.047. Fig.2 shows a negative correlation between the absolute value of the average energy score on a kidney contour and the Dice loss score, also evaluated on the contour. The energy scores fall within a negative range. Hence, the actual correlation is positive. An increase in energy values correlates with a simultaneous increase in Dice loss. According to our results, the contextual contour is the best option to distinguish poor segmentations by averaging the energy values on the kidney margins. Moreover, a contextual kidney contour increased by one voxel had the best performance, yielding a Pearson correlation of 0.801 and an ROC-AUC score of 0.888. Our experiments were statistically validated by a *p-value* lower than  $10^{-3}$ .



**Figure 2: The correlation between the average energy and the Dice loss values evaluated on a kidney contextual contour increased by one voxel.**

### 4 Future Work

While this study demonstrates the usage of energy score as a proxy for identifying misleading segmentations, future studies could experiment with other uncertainty methods, such as Mahalanobis distance or Conformal Prediction framework. Additionally, further research studies could be conducted on kidney images to gain valuable technical information, given their unique anatomy and high variability in shape and size across patients.

### 5 Conclusion

We aimed to determine if the energy score could be correlated with a high Dice loss value. After conducting our experiments on different types of kidney contours, we discovered that a contextual kidney contour increased by one voxel performed best in distinguishing good segmentations from poor ones. Our results show that the correlation between the average energy score and the Dice loss value, calculated on the kidney contour, is positive.

### References

- [1] Joris Guérin, Kevin Delmas, Raul Sena Ferreira, and Jérémie Guiochet. 2023. Out-Of-Distribution Detection Is Not All You Need. doi:10.48550/arXiv.2211.16158 arXiv:2211.16158 [cs].
- [2] Nicholas Heller, Fabian Isensee, Klaus H. Maier-Hein, et al. 2021. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. *Medical Image Analysis* 67 (Jan. 2021), 101821. doi:10.1016/j.media.2020.101821
- [3] Nicholas Heller, Fabian Isensee, Trofimova, et al. 2023. The KiTS21 Challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT. doi:10.48550/arXiv.2307.01984 arXiv:2307.01984 [cs].
- [4] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. 2021. Energy-based Out-of-distribution Detection. doi:10.48550/arXiv.2010.03759 arXiv:2010.03759 [cs].
- [5] Maohao Shen, Yuheng Bu, Prasanna Sattigeri, Soumya K. Ghosh, Subhro Das, and Gregory W. Wornell. 2022. Post-hoc Uncertainty Learning using a Dirichlet Meta-Model. *AAAI Conference on Artificial Intelligence* (Dec. 2022). doi:10.48550/arxiv.2212.07359 ARXIV\_ID: 2212.07359 MAG ID: 4311638040 S2ID: 0647c14721528fc1950f883036e0806d7b3be6f9.
- [6] Nataliia Molchanova, Alessandro Cogol, Pedro M. Gordaliza, Mario Ocampo-Pineda, Po-Jui Lu, Matthias Weigel, Xinjie Chen, Adrien Depeursinge, Cristina Granziera, Henning Müller, and Meritxell Bach Cuadra. 2024. Interpretability of Uncertainty: Exploring Cortical Lesion Segmentation in Multiple Sclerosis. *Human Brain Mapping* 45, 9 (June 2024), e26721. doi:10.1002/hbm.26721 arXiv:2407.05761 [eess].
- [7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. doi:10.48550/arXiv.1505.04597 arXiv:1505.04597 [cs].