Assessing Uncertainty and Robustness in a Deep Neural Network Model for the Determination of Gene Mutation Status in Gliomas

Winter 2019 Mechanical Engineering Undergraduate Symposium

Dingkun Guo
Instructor: Xun Huan

May 2019

Abstract

Deep learning has emerged as a powerful and prevalent technique for building data-driven models in medical diagnosis. However, most models only report single-value predictions, and are not capable of providing prediction uncertainty resulting from, among other sources, noisy and limited training data. We thus seek to develop computational capability for quantifying uncertainty in neural network (NN) models in a systematic manner. We focus on a particular residual convolutional NN model developed to predict isocitrate dehydrogenase (IDH) mutation status in gliomas from preoperative brain magnetic resonance imaging (MRI) [2]. Enabling uncertainty quantification for this model is crucial for supporting subsequent treatment decision making. As a first step, we analyze the sensitivity of model prediction to the trained weights. This is achieved through a Monte Carlo sampling approach, where random noise is added to the trained weights. Preliminary results indicate that a 5% perturbation to the weights can alter the prediction probability of IDH mutation up to 10%.

1 Introduction

Modern medical research has witnessed a significant growth in usage of artificial intelligence (AI) technology to assist decisions on patient genetic status, therapies, and diagnosis. One example, the focus of this report, involves utilizing a deep neural network (DNN) model to detect isocitrate dehydrogenase (IDH) gene mutation status from magnetic resonance imaging (MRI) scans of the patient brain tumor. The IDH genetic mutation status carries important information for medical diagnosis and prognosis, where tumors found to be IDH-mutant have observed to significantly increase survival probability of patients compared to their IDH-wild-type (i.e., without the mutation) counterparts [2]. Therefore, early determination of IDH mutation status can change surgical treatment planning and choice of therapy management plans, for example to favor early intervention as opposed to observation under certain situations [2].

However, noninvasive prediction of the IDH mutation status—that is, without extracting glioma tissue and blood samples—remains a challenge One recent development of noninvasive techniques uses a DNN, a common data-driven machine learning (ML) model, to make prediction of the IDH mutation status from MRI scans of the patient brain [2]. While this represents the current state-of-the-art progress in medical AI technology, it is also crucial to understand the uncertainty and quality of such model predictions before using this information for making medical decisions [1], which may be affected by, for example, noisy and limited number of training data. Unfortunately, most current ML models produce only single-value predictions, and the ability to report uncertainty in its predictions is largely missing. We thus seek to enable these crucial capabilities through a formalized research field known as uncertainty quantification (UQ), that quantitatively describes and tracks the effects of noise and uncertainty using rigorous statistical principles. In this report, we conduct initial investigations in assessing the robustness of the publicly available DNN model for predicting IDH mutation status [2], to test against noise in the model weight parameters and input images.

This report is organized as follows. Section 2 introduces the the details of the publicly available DNN model. We note that the DNN model available to us is already trained by its authors, and we do not make alterations to it and only perform robustness assessments by running it. Section 3 describes the methodology behind the robustness test cases. Numerical results can be found in section 4, and the report ends with conclusions in section 5.

2 Model Setup

When we are using this public model as an example to assess robustness, we encountered some technical difficulties. This section records the modification we made when we set up this model.

As shown in Figure 1, the model originally has 5 steps, which are:

1. registration and isotropic resampling,
2. n4 bias correction and skull stripping,
3. image intensity normalization,
4. compile patient samples, and
5. prediction.

These computations are performed on a set of MRI scans accessible from The Cancer Genome Atlas (TCGA) database [4]. The first step involves producing nii-format image files from raw data, which is skipped in our operations because the files from TCGA is already nii form. For step four, because TCGA database does not provide a whole 3D mask of tumor, we create a mask by extruding a 2D cropped slice mask of tumor. In the final prediction, each of the four networks in the model will read one type of MRI scans (in total four types of MRI scans: Flair, T2, T1, and T1post) and make three predictions. Those 12 predictions will be

combined into one in a logistic regression. To compare four models, we use the predictions before logistic regression.

TCGA contains MRI scans of 63 patients, and we use them as test data for conducting the robustness assessment for the DNN model. In this case, every network will produce three predictions for each patient, for a total of 189 predictions (results presented in section 4).

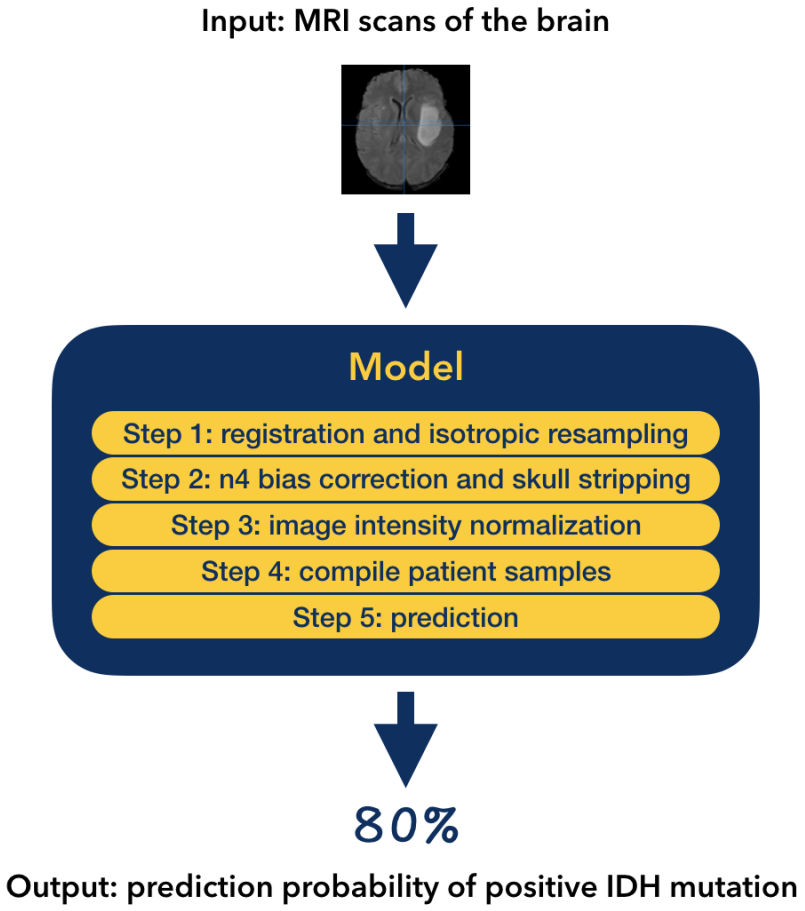

Figure 1: Five Steps to Run the Model

3 Methodology

To easily illustrate and analyze this model, we consider this model as y = DNN(x, w), where input x is the MRI scans, parameter w is weights of model, and output y is the prediction probability of positive IDH mutation status. We target to analyze the sensitivity and robustness of the model predictions with respect to noise through two aspects: (1) noise in the trained model weights, and (2) noise in input data.

3.1 Noise in Weights

A Monte Carlo sampling approach [5] is used, where different levels of random noise (1%, 5%, 10%) is added to trained DNN model weights. Then the output results is compared with no-noise predictions.

Before deploying this technique on the DNN model, we first illustrate this concept through a simple function as an example. Consider the following cubic polynomial model

$$y = 3x^3 + 2x^2 + 1, \tag{1}$$

where the monomial coefficients play the same role as the DNN weights. On the one hand, we first add different levels of noise in all coefficients. The "noisy" weights shift and distort the output curve, and thus 10% noise cause a large output band while the 1%-noise-output still stays around original output. One the other hand, we add same amount of noise in different coefficients. This also results different changes in output as shown in Figure 2.

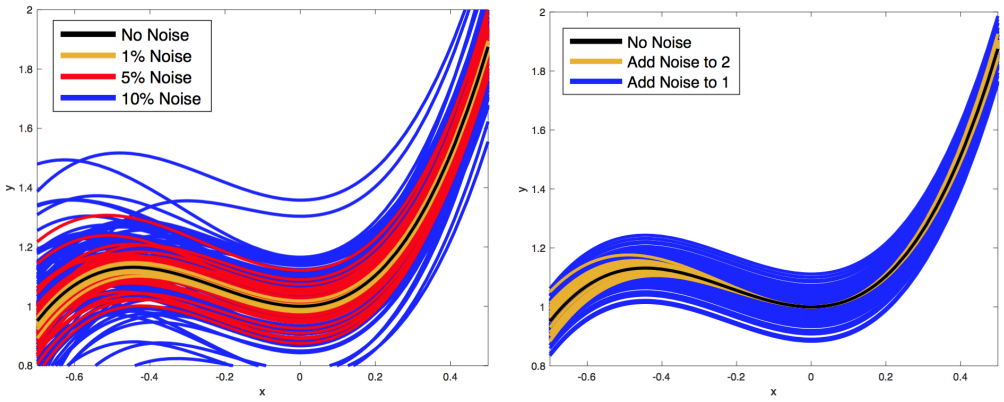

Figure 2: Left: Adding different level of Gaussian noise in all the coefficients in a simple function, Right: Adding 5% Gaussion noise in different coefficients

3.2 Noise in Data

To analyze the uncertainty caused by input data, noise is added into each pixel of MR scans. Rician noise is chosen instead of Gaussian noise because Rician noise is more commonly observed in MR scans [3]. Equation (2) shows the original Rician distribution. To simplify the algorithm, we set θ to zero and get (3). We define σ as v times noise percentage, where the variable v is pixel data. Then we can get the R that is pixel data with noise.

$$R = \sqrt{X^2 + Y^2} \text{ where } X \sim N(v\cos\theta, \sigma^2) \text{ and } Y \sim N(\sin\theta,\ \sigma^2) \tag{2}$$

$$R = \sqrt{X^2 + Y^2} \text{ where } X \sim N(v, \sigma^2) \text{ and } Y \sim N(0,\ \sigma^2) \tag{3}$$

The modified images are used as input to run the model, and the output predictions are shown in subsection 4.3.

4 Results

We perform preliminary investigations on the robustness of the model with respect to noise. Noise is added in weights and input data, and the results of output is shown below.

4.1 Noise in All Model Weights

Different levels of Gaussian noise is added to all the model weights. Figure 3 is obtained after adding noise to weights and making predictions for 10 times. As a result, small amount of noise can change the output a lot. For example, a 5% perturbation to the weights in T1post network can alter predictions up to 20%. Also, T1post network is most sensitive to the noise, and larger uncertainty observed when prediction probability is relatively small. The plot also shows T1 Network is less sensitive to large noise, as large noise only change its output about 10%, while other networks have over 20% difference in output. In addition, most output shifts up, which means overestimated probability. For patient decision making, as having IDH increases probability of survival, overestimated results could lead to less invasive treatment, which can be dangerous.

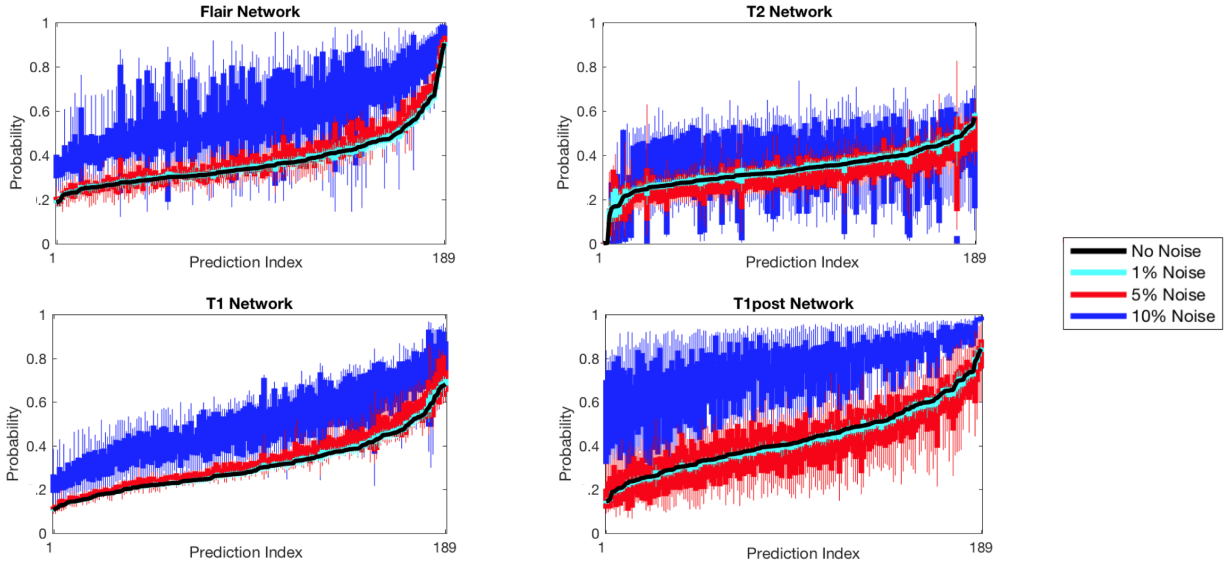

Figure 3: Adding Different Levels of Gaussian Noise to All Model Weights (10 Runs)

4.2 Noise in Different Layer

Five percent Gaussian noise is added to different layers of T1post Network. The results are shown in Figure 4. Although the 211th layer has many more weights than the first layer, its noisy predictions have lower uncertainty. In this case, we can conclude that the first layer contributes more to overall predictive uncertainty than the 211th layer, which suggests some weights may be much more important than others. If knowing which layers are more sensitive, developers can concentrate on improving those important layers.

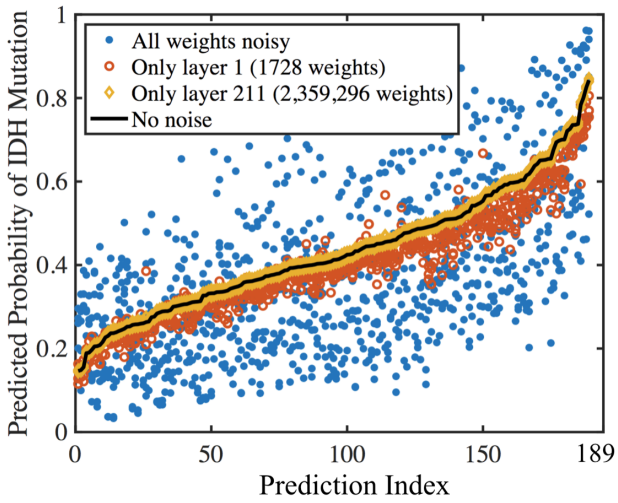

Figure 4: Adding 5% Gaussian Noise to Weights in Different Layers

4.3 Noise in Test Data

Rician noise is added to MR scans in TCGA database. Those modified images are used as input and the results are shown in Figure 5. The output band become condense compared to Figure 3. This illustrates that Rician noise in data cause less uncertainty than noise in weigts. Although the noise in data still shifts predictions, less uncertainty means the model appears more robust against image noise. It is worth noting that T2 network perform quiet well, without much changes even when 10% noise is added.

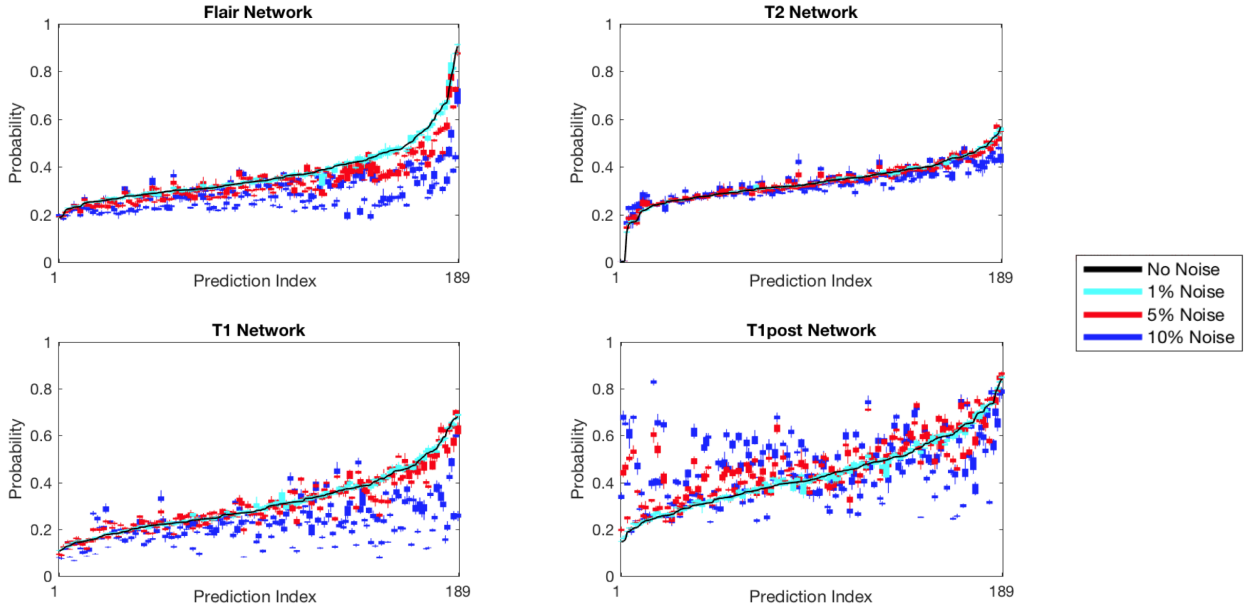

Figure 5: Adding Different Levels of Rician Noise to Images (10 Runs)

5 Conclusions

In this paper, we performed some local sensitive analysis of the DNN model for predicting IDH mutation status in brain tumors. After overcoming some technical difficulties in setting up model, we implemented random sample methods to add noise in trained model weights and test data. Different output caused by different level of noise, different layers of weights, and noise in data were provided and discussed.

In conclusion, predictions from deep neural network models studied here can be quite sensitive to noise in model weights, which are affected by the quality of training data and structure of the deep neural network. Uncertainty should be quantified and subsequently reduced from different aspects of models and data, to enable high-confidence predictions imperative for decision-making for patient treatments.

As preliminary methods and results of accessing robustness and sensitivity of medical AI models, we will repeat this kind of model evaluations to produce additional data for statistical analysis. In addition, these methods and results lead to some interesting thoughts of continuing this research. For the next steps, we will analyze the relationship between uncertainty and other features of tumor such as size and volume and conduct sensitivity analysis layer by layer to identify uncertainty contributions. These will allow us to develop a systematical tool to analyze DNN as a black box and provide a list of criteria for assessing model robustness and generalizability.

References

[1] E. Begoli, T. Bhattacharya, and D. Kusnezov, *The need for uncertainty quantification in machine-assisted medical decision making*, Nature Machine Intelligence, 1 (2019), pp. 20–23, https://doi.org/10.1038/s42256-018-0004-1, http://www.nature.com/articles/s42256-018-0004-1 (accessed 2019-05-03).

[2] K. Chang, H. X. Bai, H. Zhou, C. Su, W. L. Bi, E. Agbodza, V. K. Kavouridis, J. T. Senders, A. Boaro, A. Beers, B. Zhang, A. Capellini, W. Liao, Q. Shen, X. Li, B. Xiao, J. Cryan, S. Ramkissoon, L. Ramkissoon, K. Ligon, P. Y. Wen, R. S. Bindra, J. Woo, O. Arnaout, E. R. Gerstner, P. J. Zhang, B. R. Rosen, L. Yang, R. Y. Huang, and J. Kalpathy-Cramer, *Residual Convolutional Neural Network for the Determination of IDH Status in Low- and High-Grade Gliomas from MR Imaging*, Clinical Cancer Research, 24 (2018), pp. 1073–1081, https://doi.org/10.1158/1078-0432.CCR-17-2236, http://clincancerres.aacrjournals.org/lookup/doi/10.1158/1078-0432.CCR-17-2236 (accessed 2019-03-18).

[3] H. Gudbjartsson and S. Patz, *The rician distribution of noisy mri data*, Magnetic Resonance in Medicine, 34 (1995), pp. 910–914, https://doi.org/10.1002/mrm.1910340618, http://doi.wiley.com/10.1002/mrm.1910340618 (accessed 2019-03-18).

[4] N. Pedano, A. E. Flanders, L. Scarpace, T. Mikkelsen, J. M. Eschbacher, B. Hermes, and Q. Ostrom, *Radiology Data from The Cancer Genome Atlas Low*

Grade Glioma [TCGA-LGG] collection, (2016), http://doi.org/10.7937/K9/TCIA.2016.L4LTD3TK. The Cancer Imaging Archive.

[5] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer New York, New York, NY, 2004, https://doi.org/10.1007/978-1-4757-4145-2, http://link.springer.com/10.1007/978-1-4757-4145-2.