Université de Rennes

ISTIC

M2 MIAGE

EIT Digital

Data Science - Case Study

Author: **Michael Vitali - Smith Trivedi**

Rennes, 01/2024

Contents

List of Figures

1 Introduction

Patients getting MRI scans for multiple sclerosis have images of several sequences in 3D space that highlight different aspects of their brain and spinal cord. In this case, T2 and STIR images are used (which are two varieties of MRI scans) to help in detecting lesions according to which a patient's treatment is decided.

Each additional scan meant to capture a sequence of images is costly both in terms of time and costs, as it increases the treatment time for the patient and the resources used. As a result, the majority of information should be collected in the lowest possible number of scans. Deep learning-based solutions can detect and segment lesions with minimal scans, saving time, energy, and money.

Ricky Walsh, researcher at the INRIA Institute, has previously created a base solution that uses the UNet architecture and additional models to detect and segment lesions. The goal is to increase the performance of existing systems using a variety of approaches, including hyperparameter tuning and ensemble methods.

1.1 Problem statement

Given the T2 and STIR images and the base model, the objective is to develop and design approaches that increase sensitivity, precision, and performance for identifying lesion voxels. As the base model was implemented to work only with T2, we want to understand if the STIR images can enrich the feature and improve the performance of the model.

The proposed research methods to improve lesion detection and segmentation include:

- Ensemble methods.
- Fine-tuning the base model.

1.2 Dataset

The dataset comprises of 255 T2 Scans and 214 STIR MRI scans with the corresponding lesion segmentation. Each file is a 3 dimension scan of variable shape (which is further standardized in the pre-processing stage).
The data comprises of scans focusing on the spine and the brain.

- **T2 Scan**

  T2 images provide a good contrast between various soft tissues based on their water content. Fluids appear brighter. Tissues with high water content, such as cerebrospinal fluid and tumors, appear bright on T2-weighted images

Figure 1: A T2 scan indicating lesion from a doctor's perspective

- **STIR images**

  STIR provides high contrast for tissues with high water content, particularly in areas where fat might obscure the visualization of abnormalities.

Figure 2: A STIR scan of spinal cord

- **P-maps**

  P-Maps or Probability maps are the output of the base model that represents the likelihood of a voxel belonging to a lesion or not. This information gives some control in boosting the sensitivity and improving lesion detection.

  They could be understood as the intensity or the brightness with which a lesion voxel is spotted. A high likelihood of a voxel would indicate the presence of a bright lesion spot whereas a low likelihood indicates nothing to be found.

2 Data pre-processing

To prepare the images for model training, it's necessary to implement a pre-processing step. The reason for this is that the collected images encompass the entire back of the patient, but for training only the region containing the spinal cord is relevant. Thus, this pre-processing step will involve isolating and focusing on the spinal cord segment of the images. The pre-processing is done as following:

- **Step 1:** First of all, the image needs to be cropped around the spinal cord. See Figure 3.
- **Step 2:** The center of the spinal cord is found. See Figure 4.
- **Step 3:** The spinal cord is shifted to have it vertically instead of as an 'S' shape. See Figure 4.

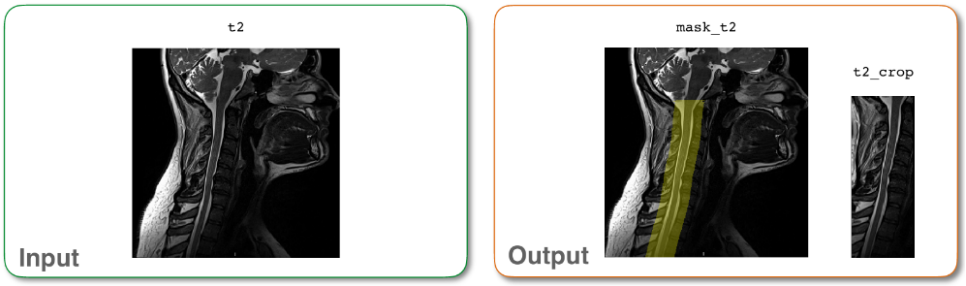

Figure 3: Crop spinal cord

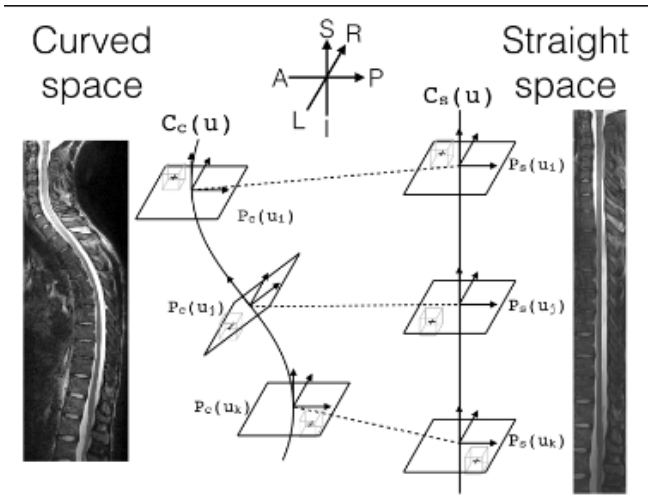

Figure 4: Centering spinal cord

3 Base Model

The base model is done using a UNet network. There are different base models trained on a different number of images and types. All of them were trained using a learning rate of $\mathbf{10}^{-1}$ and for **50000** steps. The image types that could be used for training are T2 and STIR, even though the most relevant ones are the T2 images.
There are two typologies of models: the ones trained only with T2 images and the models trained with both types. There is no model trained only with STIR images. The models are

trained over a different number of images. The number of images used are *30, 60, 120, 240*. However, there are not 240 STIR images in the dataset. Therefore, in the case of the T2-STIR model, the maximum number of images used is 120. All the models are trained 5 times using a cross-validation technique.

As a consequence, the best base model is the one trained on T2 scans using 240 images as input. The average performance on the test set is as follows:

Table 1: Results T2 model with 240 images

	les. sensit.	**les. precis.**	**les. f1**	**voxel sensit.**	**voxel precis.**	**dice coeff.**
Model 1	0.644123	0.631944	0.558350	0.424595	0.533417	0.407078
Model 2	0.699447	0.681017	0.658325	0.517000	0.523243	0.479075
Model 3	0.842337	0.680556	0.696371	0.667517	0.598310	0.583723
Model 4	0.775463	0.676117	0.679422	0.555556	0.528222	0.489749
Model 5	0.755221	0.819475	0.716675	0.541179	0.653462	0.520295
Mean	**0.743318**	**0.697821**	**0.661828**	**0.541169**	**0.567330**	**0.495983**

4 Ensemble

In the ensemble method, the goal is to boost the segmentation accuracy at image and lesion level using the P-Maps from the best base model. Given the P-Maps of the images, different machine learning algorithms are used, such as Decision Trees, Random Forest, Catboost, and Neural Networks.

4.1 Implementations

Initially, standard models were tested and trained using random search cross-validation, trying to find the best hyper-parameters.
All the models are trained using 203 images (folders 1,2,3 for training and folder 4 for validation) and tested on 51 images (folder 5).
On observation, Neural Network outperformed the other models, as it is possible to see from Table 2. Further, once NN was selected as the best model, 5 different NNs were trained using the same cross-validation folder division as the base model.

1. **Load Data:** Load the Pmaps for T2 and T2 + Stir images, and the segmentations.
2. **Data Representation:** The Pmaps are considered as inputs, and the segmentations are treated as labels, both of which are flattened.
3. **Data Cleaning:** Drop data points with P-Map values greater than or equal to 10^{-5} as a threshold.
4. **Model Training:** Train the models using the pre-processed data.
5. **Validation (Neural Network):** Apply the same threshold during validation in the case of neural networks.

6. **Model Testing:** Test the model on complete unfiltered data P-Map values.

Neural Network architecture

Layer (type)	Output Shape	Param #
dense	(None, 64)	192
dense	(None, 32)	2080
dense	(None, 16)	528
dense	(None, 1)	17

Total params: 2817 **Trainable params:** 2817 **Non-trainable params:** 0

4.2 Results and analyses

Table 2: Model selection

	Mean voxel sensit.	Mean voxel precis.	Mean dice coeff.
Decision Tree	0.396	0.717	0.281
Random Forest	0.359	0.652	0.28
Catboost	0.364	0.692	0.283
Neural Network	0.551	0.76	0.638

From the above table, it is observed that the Neural Network model outperforms the other methods.

Table 3: Neural Network models

	Mean voxel sensit.	Mean voxel precis.	Mean f1 score
Model 1	0.314	0.69	0.352
Model 2	0.663	0.767	0.696
Model 3	0.647	0.777	0.684
Model 4	0.707	0.801	0.73
Model 5	0.41	0.73	0.452
Mean	**0.5682**	**0.753**	**0.5828**

4.3 Conclusion

- **Variability Across Models:**

  There is variability in performance across the models indicating that some models are less sensitive to lesions than others.

  Therefore, a voting method could be incorporated into the final output where all the models make a vote with some confidence and a choice is made based on majority voting. This certainly would boost the sensitivity to lesion identification.

In general, it is possible to see that the ensemble method is better on average than the base model (Table 1). Therefore, it is possible to conclude that the STIR images are bringing new relevant information into the model.

5 Fine-tuning

The second idea, in the attempt to improve the base model, is to use the fine-tuning technique. Since the best base model is based only on T2 images, the idea could be to fine-tune it while using also the STIR images and see if they could bring some new information into the model and improve the predictions.

5.1 Implementation

The new model is implemented as follows:

1. **Load the base model**: Firstly, the base model and its associated weights have to be loaded.
2. **Change first layer**: Then, the first convolutional layer needs to be changed with another convolutional layer having a depth of 2, due to the use of both types of image as input.
3. **Load the weights**: Now, since the new layer has randomly initialized weights, the original weights, halved, must be loaded in the new layer.

Now that the model is ready, it has to be fine-tuned. It is done in the following way:

1. **Freeze the entire network besides the new layer**: First of all, the model has to be frozen entirely, leaving only the new convolutional layer trainable.
2. **Fine-tuning first layer**: Then, the model is fine-tuned for 5000 steps using 120 images (both T2 and STIR). In this case, it is used a learning rate of 10^{-3}.
3. **Unfreeze the entire model**: Successively, the entire model has to be unfrozen.
4. **Fine-tuning entire model**: Finally, the entire model is fine-tuned using the 120 images. Also in this case for 5000 steps using a learning rate of 10^{-4}.

The fine-tuning has to be performed for all the 5 best base models.

5.2 Results

The results of the fine-tuned models are the following:

Table 4: Fine-tuning results

	les. sensit.	les. precis.	les. f1	voxel sensit.	voxel precis.	dice coeff.
Model 1	0.779186	0.651270	0.606118	0.522710	0.520833	0.460001
Model 2	0.796164	0.594214	0.646638	0.612667	0.472633	0.495208
Model 3	0.881197	0.757459	0.758760	0.683577	0.597538	0.586815
Model 4	0.789683	0.589446	0.644367	0.610179	0.487143	0.499117
Model 5	0.788998	0.691428	0.685526	0.601016	0.563295	0.526972
Mean	**0.807045**	**0.656763**	**0.668282**	**0.606030**	**0.528289**	**0.513623**

If we compare the above results with the base models in Table 1, we can observe that the sensitivity is improved both at the lesion and voxel level, while the precision is not. In general, the f1-score is lower or very close to the base model.
In conclusion, the STIR images don't provide any additional information to the model.

5.3 What can be changed?

All the previous models are trained using the same learning rates and parameters. Hence, it may be useful to investigate and see if there is a difference when using different learning rates and number of steps.

6 Base Model Result Analyses

Given the results of the base models, it could make sense to perform some analytics to understand better if there is some pattern that is better predicted than others. Here only the most important results are showcased. For all the following analyses the model results related to each image are split into **positive** (f1-score > 0.5) or **negative**, and the images considered are the ones with at least one lesion. The distribution of the results can be seen in Figure 5.

6.1 Section Analysis

One of the possible correlations that can be found is related to the position of the lesions. The lesions can be in two different parts categorized as **Cervical** and **Thoracic**. Given the distribution above, the results obtained are:

Table 5: Section Results

	Cerv	**Thor**
Positive	78	36
Negative	33	32

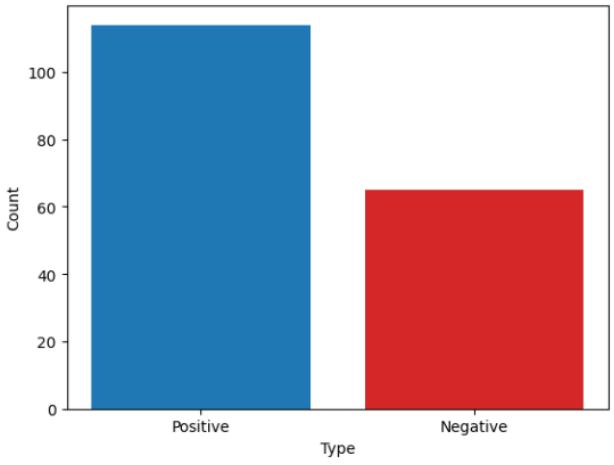

Figure 5: Results distribution

The images in the cervical part are detected positive **70%** of the time, while the images on the thoracic part only **50%** of the time.
With these results, it is possible to say that the lesions on the cervical part are easier to detect, as expected.

To understand if there is any statistical difference between the cervical and thoracic zone the **chi²-test** is performed. The result is a p-value of **0.029** $< 5\%$. Therefore, the null hypothesis, claiming that there is no difference between the two zones, is rejected and the test affirms that a difference exists.

6.2 Device Analysis

In this section, the idea is to see if there is any correlation between the detection of the lesions and the device used to take the images. The results are the following:

	Positive	**Negative**
Prisma	0.67	0.33
Discovery MR750w	0.63	0.37
Aera	0.62	0.38
Avanto	0.52	0.48
Skyra	0.68	0.32
Optima MR450w	0.5	0.5

	Positive	**Negative**
Ingenia	0.77	0.23
Avanto fit	0	1
Spectra	0.33	0.67
Signa HDxt	0.83	0.17
Signa Exp.	0.5	0.5
Verio	0	1
Signa Artist	1	0

In the previous results, we can see that the images taken with Prisma, Discovery, Aera, Skyra, Ingenia, and Signa are most of the time well segmented while the other ones are not. The results with 0 positives or negatives can be caused by the fact that there are not many images taken with that particular device.

The same analyses can be performed from the machine brand point of view. However, there are no interesting results to report.

6.3 Correlation between ground truth volume and detection rate

This section looks into the possible correlation between the volume of the lesions (in the ground truth) and their detection rate. To perform this analysis, the lesions are split into intervals of volume over which the ratio is calculated.
The extremes of the intervals considered are *[0, 35, 55, 80, 115, 200, max volume]*. With this division, all the intervals contain approximately the same amount of lesions.
The results are the following:

- Ratio of found lesions in the interval 0-35: **0.47**
- Ratio of found lesions in the interval 35-55: **0.62**
- Ratio of found lesions in the interval 55-80: **0.74**
- Ratio of found lesions in the interval 80-115: **0.81**

- Ratio of found lesions in the interval 115-200: **0.80**
- Ratio of found lesions in the interval 200-max: **0.87**

As expected, it can be observed that the lower the volume of the lesion to predict, the smaller the predicted ratio. On the contrary, the higher the volume of the lesion to predict, the easier it is to predict the lesion, leading to an increase in the ratio.

6.4 Correlation predicted volume and detection rate

The idea here is to see if there is any correlation between the predicted volume and the prediction rate. To perform this analysis, the lesions are split into volume intervals over which the ratio is calculated.
The extremes of the intervals considered are *[15, 35, 60, 100, 170, max volume]*. With this division, all the intervals have approximately the same number of lesions inside.
The results are the following:

Predicted Volume:

- Ratio of found lesions in the interval 0-15: **0.26**
- Ratio of found lesions in the interval 15-35: **0.37**
- Ratio of found lesions in the interval 35-60: **0.57**
- Ratio of found lesions in the interval 60-100: **0.60**
- Ratio of found lesions in the interval 100-170: **0.73**
- Ratio of found lesions in the interval 170-max: **0.75**

From these results, it can be concluded that the higher the predicted volume of the lesion, the better the lesions are detected. This outcome was unexpected, as one might expect to predict with high frequency a very large volume over a small lesion. However, this situation seems not to occur very frequently.

Reference Volume:

- Ratio of found lesions in the interval 0-15: **0.01**
- Ratio of found lesions in the interval 15-35: **0.78**
- Ratio of found lesions in the interval 35-60: **0.96**
- Ratio of found lesions in the interval 60-100: **0.93**
- Ratio of found lesions in the interval 100-170: **0.97**
- Ratio of found lesions in the interval 170-max: **0.80**

The second results show that the higher the ground truth volume, the higher the accuracy in recognizing the lesions. However, the biggest lesions show a declining detection ratio. This could be explained by the cases in which the lesions are very long and situated in the thoracic part of the spinal cord. Therefore, it is more difficult to detect the entire lesions with respect to the smaller ones.