A six-year Australian and New Zealand experience of artificial intelligence techniques applied to MRI liver data - tricks and traps

Award: Wiley-Blackwell Best Exhibit Award, Radiology
Poster No.: R-0041
Congress: 2018 ASM
Type: Educational Exhibit
Authors: M. Blake¹, T. Cooper², S. Khorshid¹, T. St Pierre¹, ¹WA/AU, ²NZ
Keywords: Computer Applications-Detection, diagnosis, Neural networks, Computer applications, Image verification
DOI: 10.1594/ranzcr2018/R-0041

Any information contained in this pdf file is automatically generated from digital material submitted to EPOS by third parties in the form of scientific presentations. References to any names, marks, products, or services of third parties or hypertext links to third-party sites or information are provided solely as a convenience to you and do not in any way constitute or imply ECR's endorsement, sponsorship or recommendation of the third party, information, product or service. ECR is not responsible for the content of these pages and does not make any representations regarding the content or accuracy of material in this file.

As per copyright regulations, any unauthorised use of the material or parts thereof as well as commercial reproduction or multiple distribution by any traditional or electronically based reproduction/publication method is strictly prohibited.

You agree to defend, indemnify, and hold ECR harmless from and against any and all claims, damages, costs, and expenses, including attorneys' fees, arising from or related to your use of these pages.

Please note: Links to movies, ppt slideshows and any other multimedia files are not available in the pdf version of presentations.

www.myESR.org
Learning objectives

This paper presents a six-year experience in an Australian and New Zealand setting of applying computer algorithms to automate the analysis of MRI images of the liver for the quantification of iron and fibrosis, and the insights and learnings achieved over this period to transform these algorithms into a regulated medical device.
Background

The following terms are used either throughout the paper or are essential background knowledge to an understanding of the methodologies relating to image processing and computer based artificial intelligence:

**Machine Learning** is the use of algorithms to allow computers to learn models from data without being explicitly programmed.

**Support Vector Machines** - use *supervised learning* methods (see below) for classification. The SVM is presented with examples of images that have been classified into one of two categories. The SVM builds a model that predicts whether new images fall into one category or the other. In part one of the Procedure Details section, we trained an SVM to categorize T1 VIBE images of the liver into high or low-grade fibrosis based on training data from biopsies.

**Clustering** - uses *unsupervised learning* methods (see below) to assign a set of observations to subsets (called *clusters*) so that observations within the same cluster are similar according to some pre-designated criterion or criteria, while observations drawn from different clusters are dissimilar. We tried this approach to find features associated with high and low-grade fibrosis in MR liver images.

**Artificial Neural Networks** - are software structures that mimic aspects of biological neural networks. Layers of nodes are connected by "axons" with weightings. Data are input to the "top" layer and output is retrieved from the "bottom" layer. By training the network with multiple examples of input with known outputs, the weightings of the "axons" are adjusted such that a trained network can receive an input dataset and predict the output from the trained weightings. (Further described in parts two and three of the Procedure Details section)

**Deep Learning** - uses artificial neural networks with extra layers of nodes. Extra computing power is required for deep learning. We built a computer specifically for the task (described in part three of the Procedure Details section).
**Transfer Learning** - is a method of training an artificial neural network for a particular task using training data from another task. For example, an artificial neural network could be trained to recognise different makes and models of cars from the thousands of examples on the internet. The trained network can then be further trained to recognise different makes and models of motorcycle with a much smaller number of samples but with a high level of success. This approach can be useful when you do not have a large dataset for training.

**Supervised learning** [1] takes an input variable and an output variable and uses an algorithm to learn the mapping function from the input to the output. The aim is to approximate the mapping function so when new input data is presented the model can predict the output variable for that data. The learning stops when the algorithm achieves an acceptable level of performance. All the data are labelled and the algorithm learns to predict the output from the input data.

**Unsupervised learning** [1] is used for input data where there is no corresponding output variable. The goal of unsupervised learning is to model the underlying structure or distribution in the data to learn more about the data. The algorithm is left to its own devices to discover and present the interesting structure in the data. This can be grouped into a clustering problem where you want to discover the inherent groupings in the data or an association problem where you want to discover rules that describe large portions of the data. All data are unlabelled and the algorithms learn the inherent structure from the input data.

**Semi-supervised** refers to when some data are labelled but most of them are unlabelled and a mixture of supervised and unsupervised techniques can be used.
**Imaging findings OR Procedure details**

**Part One: Support Vector Machine Learning - Liver Fibrosis**

In 2012, we applied support vector machine learning (SVM) to a cohort of 28 patients with hepatitis C to assess textural differences in MRI images [2] to build a predictive model for categorising the patients' Ishak fibrosis scores [3]. Figure 1 shows an example liver image used in the study. The assumptions and limitations of the modelling of small patient numbers, that include cross-validation and overfitting, are presented.

The small subject population was divided into both training sets and validation sets using a "leave one out" cross validation technique. Features in the images associated with liver fibrosis (as assessed by biopsy) were identified using clustering techniques. The parameter optimisation was performed using 27 of the 28 patient image sets as the training data set and then the performance of the resulting model was tested on the 28th patient as the test set. This was repeated 28 times, each time leaving out one of the image sets as the single test case. The SVM model yields a number of "votes" for whether or not the image has features associated with liver fibrosis. The results from the training model for the 3-minute post contrast 3D VIBE images are shown in Figure 2.

The subjects were re-scanned 6 weeks later. Using the leave one out models trained on 27 of the 28 visit 1 images sets; 27, 28, and 28 out of 28 cases were correctly classified in visit 1 data for pre-, 3-mins post, and 10 mins post-contrast images, respectively (Table 1). The model trained from all 28 visit 1 datasets correctly classified 23 out of 27, and 26 out of 28, and 25 out of 28 cases for the pre-, 3-mins post, and 10 mins post-contrast images from visit 2 (Table 2).

**Table 1. Summary of the results for the leave-one-out classification of visit 1 MRI data**

<table>
<thead>
<tr>
<th>Image Type</th>
<th>Classification Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-contrast</td>
<td>27 out of 28 cases (96%)</td>
</tr>
<tr>
<td>3-min post contrast</td>
<td>28 out of 28 cases (100%)</td>
</tr>
<tr>
<td>10-min post contrast</td>
<td>28 out of 28 cases (100%)</td>
</tr>
</tbody>
</table>

**Table 2. Summary of the results for the classification of visit 2 MRI data based on the model trained on visit 1 data.**
<table>
<thead>
<tr>
<th>Image Type</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-contrast</td>
<td>80%</td>
<td>92%</td>
</tr>
<tr>
<td>3-min post contrast</td>
<td>93%</td>
<td>92%</td>
</tr>
<tr>
<td>10-min post contrast</td>
<td>93%</td>
<td>85%</td>
</tr>
</tbody>
</table>

While these results are promising, the validation at visit 2 was on the same patients involved in the training of the SVM model. There is a danger of "overfitting" the model, especially when a small number of cases is involved. Further studies on independent cohorts of patients are now required.

**Cross validation** assesses how the results of a statistical analysis will generalise to an independent data set. It is mainly used where the goal is to test the model's ability to predict new data that were not used in training it, and to give insight on how the model will generalise to an independent dataset.

**Overfitting** is the production of a prediction model that corresponds too closely or exactly to the data upon which it is trained and may therefore fail to accurately predict future observations reliably. See Figure 3.

**Part Two: Transfer Learning**

In 2016, our focus shifted from attempting to predict the pattern of liver fibrosis to predicting liver iron concentration (LIC) from MR images. Our rationale was that we needed to work on problems where we had larger datasets available to us to assess the influence of the number of datasets on the machine learning and to have sufficiently large and diverse populations for validation of the models. We had available to us several hundred sets of MR liver images acquired using the FerriScan® protocol together with associated liver iron concentrations. The FerriScan method involves acquiring 11 abdominal axial slices through the liver. Analysts choose the most appropriate slice and manually segment the liver to carry out the liver iron concentration analysis.

To model the relationship between the images and the liver iron concentrations, transfer learning techniques with a convolutional neural network (CNN) were used [4]. CNNs have been found to be effective at building models from images [5] and this method along with gradient boosting machines (GBMs) was chosen for this project. GBMs are efficient for building models from numerical data [6].

Our initial attempt used 200 MRI datasets to further train a pre-existing artificial neural network that had been trained on 2.6 million human face images of 2600 people [7].
MRI data comprised images where the image slice for analysis was chosen manually to match exactly the image slice that had been used in the FerriScan® analysis. By splitting the 200 datasets into training and validation sets, we were able to determine how the number of datasets in the training set determined the accuracy of the model (as determined by the magnitude of the square root of the mean squared difference between the model prediction and the liver iron concentration measured by FerriScan® - see Figure 4).

It can be seen that the accuracy of the model improves (RMSE decreases) as the number of images in the training set increases. This is to be expected - more data means better models. As the number of patients approaches 199 the spread about the line increases. This is because the number of patients in the test set is decreasing (at 199 there is only one patient in the test set) so the measurement error increases.

From this learning curve, we decided to increase the number of training images to 400 based on the concept of diminishing returns. The resulting model was then tested on 100 subjects not involved with the training of the CNN and using scanners different from those used to train the CNN [8]. The results of the testing are shown in Figure 5.

The advantages of a machine learning approach over a manual approach using a parametric formula (such as FerriScan®) are potentially that:

- the machine can use other organs in the image set besides the liver to do its calculations;
- it can use all slices in the image set instead of just the best slice;
- it can figure out and allow for scanner differences;
- it can figure out and allow for patient differences such as age, gender etc;
- it can learn the patterns of and account for noise in the images;
- it can work out its own formula for LIC instead of being restricted to the bi-exponential formula.

**CNNs are suited to regression and classification problems which produce a prediction model, typically of decision trees, built in a staged fashion to allow optimisation of the mathematical function. For the CNNs the MatConvNet Toolbox [9] for MATLAB was chosen and for GBM the xgboost package [10] for R was used.**

*There are two ways to use transfer learning with CNNs. One way is to leave some of the CNN layers fixed and to retrain other layers to model the MRI images. The other way is to extract “activations” from one of the CNN layers and to feed them into a separate machine learning model such as GBM.*
The former method is problematic because so many parameters are retrained in the CNN that with only a few hundred images it is possible to model the images perfectly and predict LIC exactly. This model would be overfitted which means that it is specific to images used in the training and would not generalise to unseen images. GBMs have fewer parameters and so are less prone to overfitting. They also allow the use of extra information such as age and gender of patients, although this was not used for the initial proof of concept.

In our example, the CNN trained on images other than MRI images. We used the vgg_face CNN descriptor from Oxford University [6]. This network was trained on 2.6 million face images of 2600 people. We tried other pre-trained CNNs and this was found to be the best. The reason may be due to the similarity in shape and features between faces and MRI abdominal images. This descriptor requires colour images, so we colourised the MRI images by creating a gradient of skin tones to map the grey pixels onto. Skin tones were scraped from face images on the internet.

Some trial and error was used to find the best activation layer and we found this to be the fully convolutional layer "fc6." This layer has 4096 numerical values. So in essence the vgg face CNN was used to transform a 256 x 256 pixel MRI image into 4096 numbers in a way that retained the essence of the image.

In addition, we used the 65536 raw pixel values as inputs to a GBM to find which pixels were most valuable in predicting LIC. The most valuable pixels were, as to be expected, clustered around the liver. We also found that pixels near the edges were valuable - possibly these had something to do with calibration of the MRI images.

We chose the 4096 most valuable pixels and the 4096 activations to create a GBM model that predicted $R^2$ values from the 8192 inputs. We used 5-fold cross validation to prevent overfitting for all our models.

Limitations of the approach:

- the number of samples provided for training were small;
- the low sample count forced the choice of transfer learning to attempt to solve the problem;
- manual selection of image slice and manual segmentation of the torso were required.

Part Three: Deep Learning and Convolutional Neural Networks
In 2017, we were able to access a much larger dataset of MR images (several thousand) with associated liver iron concentration values. With the much larger dataset, we decided to approach the problem by

- training CNNs from scratch rather than using transfer learning;
- training a CNN to select the image slice for analysis;
- training a CNN to predict the liver iron concentration without the need for manual image segmentation.

The new approach leveraged the latest open source deep learning libraries (see box below for details). For training image slice selection, the original slice number chosen manually by FerriScan® analysts was used to train a CNN. For training liver iron concentration prediction, the resulting FerriScan® result was used to train a separate CNN.

The machine used contains a single 1080ti NVIDIA graphics card, 128 gigabyte of memory and a 1 terabyte NVMe ssd drive. The open source deep learning libraries used were Tensorflow & Keras.

After training on 15,354 image datasets, the trained CNNs were tested on a further 906 new datasets that had not been used in the training of the CNNs. Furthermore, the unseen 906 datasets used for testing the performance of the CNNs had been acquired from 77 scanners that had not been used for acquiring the training data. The predictive performance of the trained CNNs can be seen in Figure 6.

---

**Saliency Maps**

A question that inevitably arises when using artificial intelligence to analyse images is "what is the CNN doing?" This question is almost impossible to answer but some clues can be given by examining "saliency maps" that indicate the weighting that each pixel in an image has with respect to particular aspects of the neural network that result in the final output. In our example of the measurement of liver iron concentration from MR images, manual analytical analysis involves modelling the exponential decay of signal in a series of spin echo images (see Figure 7). The rate of decay is related to the liver iron concentration by a calibration curve.

For some scanners, signal gain changes significantly between echoes and hence analysts normalise tissue signal intensities to that observed in the normal saline bag that is scanned along side the patient. Analysts also use the signal variation in subcutaneous fat to help model the RF signal intensity throughout the torso. With these observations in mind, it is very noteworthy that saliency maps generated by the CNN
used for liver iron concentration measurement show that significant attention is paid to the liver (as expected), but also to the subcutaneous fat (Figure 8) and also to the normal saline bag (Figure 9) for scanners that change gain between echoes. The startling aspect of these saliency maps is that they show that the CNN has "learned" to examine these relevant parts of the images without having been explicitly trained or instructed to do so!

**Lessons Learned**

- Thousands of datasets were required to train the CNN from scratch;
- Reasonable results can be achieved with large numbers of well labelled data;
- CNNs can completely automate an analysis process.
Fig. 1: Assigning degree of fibrosis to MR images of the liver - contrast enhanced 3D VIBE cross section of the liver.

© Resonance Health Analysis Services Pty Ltd
Fig. 2: Classification of 28 subjects by the SVM model on 3 min post contrast 3D VIBE images. Blue bars indicate patients with biopsy proven liver fibrosis with Ishak stage 2 or less. Red bars indicate patients with biopsy proven liver fibrosis with Ishak stage 3 or greater. The overall model gives 100% correct classification using the "leave one out" cross validation method.

© Resonance Health Analysis Services Pty Ltd
Fig. 3: Schematic diagram illustrating an overfitted model (green line) for distinguishing between red and blue populations and a more robust model (black line). Figure taken from Wikipedia.

© Wikipedia
Fig. 4: Learning curve showing the root mean squared error (RMSE) vs numbers of datasets used in each training set.

© Resonance Health Analysis Services Pty Ltd
**Fig. 5:** Predictions of liver iron concentration from the transfer learning neural network plotted against values obtained by the reference standard, FerriScan®, for 100 iron loaded subjects. The diagonal line is the line of equivalence and not a regression fit.

© Resonance Health Analysis Services Pty Ltd
Fig. 6: Predictions of liver iron concentration from the convolutional neural networks plotted against values obtained by the reference standard, FerriScan®, for 906 subjects. The diagonal line is the line of equivalence and not a regression fit.

© Resonance Health Analysis Services Pty Ltd
Fig. 7: Series of single spin echo images of an iron loaded liver with increasing echo time. Note how the signal intensity in the liver decreases with echo time. A bag of normal saline is included in the image to act as reference standard for scanner gain.

© Resonance Health Analysis Services Pty Ltd
**Fig. 8:** Spin echo image together with saliency map. Note how not only pixels in the liver are used but also pixels in the subcutaneous fat.

© Resonance Health Analysis Services Pty Ltd

**Fig. 9:** Spin echo image together with saliency map. Note how the CNN uses pixels in the normal saline bag in this case. This particular scanner changes gain settings between echoes.

© Resonance Health Analysis Services Pty Ltd
Conclusion

Other things to consider when building a neural network:

This paper reports a six-year experience applying machine learning models to small and large image datasets. The key elements that have resulted in an acceptable model are the growth in computer processing power, access to advanced machine learning algorithm libraries, and access to large volumes of consistently acquired and well labelled data.

Although the majority of the analysis in this paper has focussed on the application of computer algorithms to automate the analysis of MR images of the liver for the quantification of iron and fibrosis and the insights and learnings achieved over this period, there must also be consideration given to the successful deployment of the CNN as a regulated medical device. When seeking regulatory clearance of the CNN (Software as a Medical Device, SaMD), the requirements for security, traceability, reproducibility, and performance all need to be addressed, and it is more efficient to consider these from the outset.

In the development stage, the emphasis is on performance (training the network to improve accuracy and reproducibility). Once a suitable performance is achieved, in order to meet regulatory requirements (FDA, CE Mark, TGA) for reproducibility, the algorithm must no longer be learning. In our example, which has now achieved regulatory clearance in several jurisdictions, the final model had been trained on a sufficient number of datasets and had reached a saturation point, where diagnostic performance did not significantly improve with additional training data. The software was then able to be pushed to production, with a defined diagnostic performance.

Whilst there may be examples where CNNs have not been trained on sufficient datasets to reach maximum accuracy and continued learning may result in increased/changed performance, from a regulatory perspective these models will not meet the current fundamental requirement for reproducibility. In effect, it is a particular output that receives regulatory clearance. If a device attains regulatory authority clearance and then the performance improves due to additional training of a test system, this improvement is not cleared without additional validation work, and potentially a new submission to regulatory authorities.

Regulatory clearance for SaMD also requires absolute security and traceability of data. All data need to be secure at all times (in transit and at rest) for HIPAA compliance, and the ability to trace data access and usage is essential. Regulators assess these factors...
very closely to ensure the entire device meets these requirements. With our CNN, we allowed for several ways for the user to access the algorithm (web-portal and third-party platforms). Each of these access methods has had best endeavours consideration given to meeting the full requirements of regulators.

Conclusion:

The last six years have seen an enormous growth in the use of computer programs, collated under the broad term "Artificial Intelligence" across all industries. This paper tracks those changes in a local scenario culminating in the production of a regulatory cleared Medical Device that uses artificial intelligence to quantify liver iron concentration from MRI.

There are valid fears that have been expressed about the impact Artificial Intelligence will have on jobs as part of the fourth industrial revolution. However, there are many solutions to many diagnostic problems that will need to come together before radiologists are replaced by machines. Current commercial offerings, like the solution described above, solve one problem only.

We prefer to refer to AI as Augmented Intelligence, creating tools that will assist and enhance radiologists in the performance of their duties. Augmented Intelligence will reduce the mundane and the routine, improving product and service by increasing diagnostic accuracy, reducing error, and increasing throughput.

When used as decision making support tools for increased efficiency and accuracy, AI will allow capacity for reproducible and rapid decision making to radiologists globally. The liver iron measurement tool described above is now available in nine developing countries that struggle with access to affordable healthcare, either due to cost or lack of availability of appropriately trained medical staff. AI may be able to bring solutions at low cost to many countries that currently have insufficient expertise and will provide increased efficiencies and support for radiologists in general.

The development of these tools requires large bodies of accurately labelled data on which to learn. Australian and New Zealand radiology is well placed to participate in building these solutions. We encourage you to find partners and get to work!
Personal information

Martin Blake: MBBS, FRANZCR, FAANMS, MBA, FAICD

Martin is a Radiologist and Nuclear Medicine Specialist, with 30 years medical imaging experience and has been a partner of Perth Radiological Clinic since 1997 and Chairman since 2009. He has also been a non-executive director of Resonance Health (ASX:RHT) since 2007 and Chairman since 2012.

Tony Cooper: BSc (Hons), MS (Statistics)

Tony is a Data Scientist and is a founder of Precision AI which provides artificial intelligence solutions to European and Australasian industry. He has a distinguished academic record from Stanford University and 25 years industrial and scientific experience in data science.

Sherief Khorshid: BSc

Sherief is a Data Scientist and Founder of Three Springs Technology, an Artificial Intelligence consulting business. Sherief has 20 years commercial experience primarily in finance technology and AI. He is also an active member of Perth's Machine Learning community.

Tim St Pierre: BSc (Hons), PhD

Tim is a widely published Medical Physicist in the Department of Physics, University of Western Australia. He led the team that developed the non-invasive liver iron measurement technology FerriScan® and has also developed technologies for measurement of liver fat and detection of parasites in human blood, faeces and urine.
References


Acknowledgements: Alison Laws, CEO, Resonance Health Ltd; Ajmal Mian, The University of Western Australia; Sander Bangma, Moodle; Wenjie Pang, Resonance Health Ltd; Mike House, Resonance Health Ltd.