*Healthcare of the Future*
*T. Bürkle et al. (Eds.)*


*doi:10.3233/978-1-61499-961-4-105*

# Automated Rating of Multiple Sclerosis Test Results Using a Convolutional Neural Network

Martin Eduard Birchmeier [a,1], Tobias Studer [a]
[a] *Bern University of Applied Sciences*

**Abstract.** This work concerns methods for automated rating of the progression of Multiple Sclerosis (MS). Often, MS patients develop cognitive deficits. The Brief Visuospatial Memory Test-Revised (BVMT-R) is a recognized method to measure optical recognition deficits and their progression. Typically, the test is carried out on paper using geometric figures which the patient should recognize and trace. The results are rated manually by a physician. The goal of this work was to digitize the BVMT-R and to support the interpretation of the test results using a machine learning (ML) algorithm. A convolutional neural network (CNN) was used to rate the drawings of a patient. As a result, the correct point value of the BVMT-R could be determined with an accuracy between 57 % and 76% based on a training set of 624 patient drawings obtained from 135 patients. These drawings had been previously physician rated to serve as a gold standard. In our experiment, we obtained reasonable accuracy above 80% when more than 40 drawings were available, but our training sample was too small for more detailed analysis. Conclusion: At the currently achieved classification accuracy, results analysis will remain a physician task, potentially supported with ML based preclassification, but there is hope that ML accuracy can be further improved to enable automated follow-ups.

**Keywords.** Multiple Sclerosis, BICAMS, BVMT-R, Machine Learning, Convolutional neural network, digitalize

## 1. Introduction

Multiple Sclerosis (MS) is a demyelinating disease in which the insulating covers of nerve cells in the brain and spinal cord are damaged. MS causes inflammations in the brain as well as scattered occurrences in the spinal cord resulting in a range of progressively appearing signs and symptoms such as double vision, muscle weakness or coordination problems. It is the most common immune-mediated disorder of the central nervous system and can result in severe neurologic disabilities even in young adults [1]. The progressive cognitive deficits can be divided into domains such as information processing speed, attention function, learning/memory functions as well as executive functions such as planning and execution of complex tasks or problems [2].

In order to investigate these cognitive impairments, an international initiative was formed to recommend and support a fast and universal cognitive assessment named “Brief International Cognitive Assessment for MS” (BICAMS) [3]. The recommended test battery comprises three different tests, including the “Brief Visuospatial Memory

[1] Corresponding Author Martin Birchmeier, Bern University of Applied Sciences, Quellgasse 21, CH-2501 Biel / Bienne, E-mail: m.birchmeier@hotmail.com

Test Revised" (BVMT-R) [4].The BVMT-R test requires the patient to inspect a 2 × 3 stimulus array of abstract geometric figures. There are three learning trials of 10s time. The array is removed and the patient is asked to draw the array from memory, with the correct shapes in the correct position [3]. The test is carried out on paper and rated manually by a physician. Every correct draw of a figure in the correct place receives a rating of 2 points. If the drawing is not correct but similar to the original or correct but in the wrong position, the rating is 1. If the drawing is wrong or in the wrong place, the rating is 0 points.

The long-term goal of this project is the transfer of BVMT-R to a tablet based interface using an app and to automatize the results analysis using a machine learning (ML) algorithm. In this part we demonstrate the results of the automated analysis.

## 2. Method

We chose the "convolutional neural network" (CNN) technology for pattern recognition because this algorithm has been developed for visual object classifications [5,6]. The CNN analyses the images through a row of filters. The output of the CNN is a rating of the image with a probability-value for the reliability of the rating [7]. In our case the CNN was available on Microsoft Azure with the "Custom Vision" algorithm [8].

A total of 779 physician rated drawings from 135 MS patients was obtained from COGITO GmbH Germany. For each of the 6 BVMT-R figures between 127 and 134 drawings were available. All drawings were scanned and digitized with an app to adjust resolution, color and line width. The dataset was then random split in 624 figures (=80%) training and 155 (=20%) test drawings (see table 1).

For each of the six figures a separate CNN was trained.

**Table 1.** Accuracy of the rating of 6 ML algorithms (one for each figure) compared to the physician rating as a gold standard. n = number of test drawings, m is the number of drawings used for training.

| Number | Figure | Rating 0 | Rating 1 | Rating 2 |
|---|---|---|---|---|
| 1 |  (n=26, m=101) | 0% *(n=2, m=5)* | 67% *(n=6, m=23)* | 83% *(n=18, m=73)* |
| 2 |  (n=25, m=102) | 67% *(n 6, m=24)* | 91% *(n=11, m=45)* | 63% *(n=8, m=33)* |
| 3 |  (n=26, m=104) | 67% *(n=4, m=26)* | 63% *(n=10, m=31)* | 67% *(n=12, m=47)* |
| 4 |  (n=26, m=102) | 100% *(n=10, m=38)* | 50% *(n=4, m=18)* | 67% *(n=12, m=46)* |
| 5 |  (n=26, m=107) | 88% *(n=8, m=34)* | 50% *(n=6, m=23)* | 92% *(n=12, m=50)* |
| 6 |  (n=26, m=108) | 93% *(n=14, m=56)* | 20% *(n=5, m=22)* | 86% *(n=7, m=30)* |
| average | n=155, m=624 | 69% *(n=44, m=183)* | 57% *(n=42, m=162)* | 76% *(n=69, m=279)* |

## 3. Results

Figure 1 maps the rating of the physician against the rating of the ML algorithm for all 6 figures. Dot size represents percent values. Diagonal green dots represent matching results of physician and ML rating. The green, top right point (2, 2), e.g. signifies that overall 76% of all drawings rated with 2 points were correctly classified by the ML algorithm. Thus, we measured an overall recognition accuracy of 69% for drawings rated with zero, 57% for drawings rated with one and 76% for those rated with two points (fig 1). Fig 1 also demonstrates that the likelihood of a gross misinterpretation (e.g. the ML

algorithm classifying a 0 for a drawing rated 1 or 2 by the physician) is small. The algorithm tends to rate drawings higher than they are.

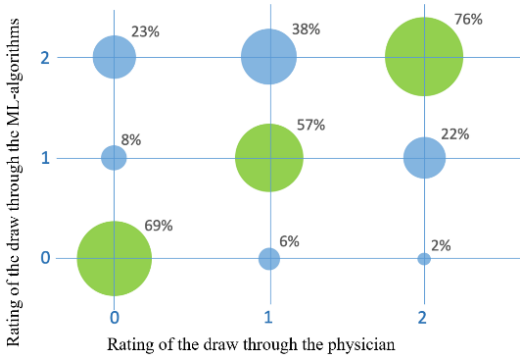


**Figure 1.** Classification by ML algorithms (y-axis) compared with the physician rating (x-axis)

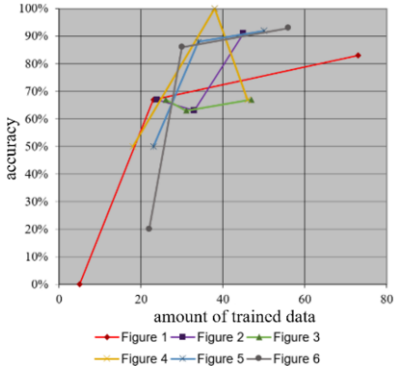


**Figure 2.** Number of training drawings per fig and the accuracy of the algorithm

Based on these results we were interested to determine the required size of the training data set to obtain reasonable accuracy of the ML algorithm classification. Fig 2 plots the number of training drawings per figure against obtained recognition accuracy for the 6 figures of the BVMT-R. For BVMT-R figures 1, 5 and 6 we note a strictly monotonic increasing plot. Figures 2, 3 and 4 are not fully monotonic. Achieved classification accuracy varies between 67 percent for figures 3 and 4 and 93 percent for figure 6. Good classification results start at 30 test drawings for figure 6 resulting 86% accuracy, closely followed by figure 5 (34 drawings resulting in 88% accuracy).

## 4. Discussion

We operated with a comparatively small dataset of between 101 and 108 drawings in the training set for each of the 6 CNN used in this experiment. This difficulty is common in medicine where it is not easy to obtain validated gold standard data for a certain problem, disease or finding.

Considering this fact, our classification results for the automated classification of the BVMT-R, although not brilliant, are encouraging. If a classification accuracy of around 80% can be achieved, it is conceivable that automated classification may be used as a first step in an IT based application to support the physician in his classification task. This is in accordance with Beam [9] who confirms that deep learning approaches, depending on the task, can be used even for small training data sets. It is an advantage that the BVMT-R figures are black-white only and comparatively simple.

BVMT-R figures 5 and 6 delivered better recognition accuracy, achieving more than 80% already with training data sets of 34 and 30 drawings, respectively.

We note that the ML algorithm has a problem to differentiate a semi-correct drawing (score 1) from a fully correct drawing (score 2) (fig 1). On visual inspection we can confirm that these kinds of drawings can often have small differences only, e.g. one extra line starting in the wrong corner of the rectangle.

Our future work, apart from the attempt to obtain additional physician rated training data will focus on the digitizing of the BVMT-R itself. It should be possible to represent the full BVMT-R workflow either on tablet or on another smart device. Obviously, we

will then need a patient study to compare paper based BVMT-R results with those measured with the digital device. We accept the possibility that there may be distinct differences in absolute values. The digitized test, however, offers the opportunity for repeated observations (using all 36 available BVMT-R figures) and thus to follow up the improvement or deterioration of a patient over time. At the currently achieved classification accuracy, results analysis will remain a physician task, potentially supported with ML based preclassification, but there is hope that ML accuracy can be further improved to enable automated follow-ups.

## 5. References

[1] A. Compston, A. Coles. Multiple sclerosis. *Lancet*. **372(9648)** (2008), 1502–1517.
[2] N.D. Chiaravalloti, J. DeLuca. Cognitive impairment in multiple sclerosis. *The Lancet Neurology* **7(12)** (2008), 1139–1151.
[3] D.W. Langdon, M.P. Amato, J. Boringa, B. Brochet, F. Foley, S. Fredrikson et al. Recommendations for a Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS). *Mult Scler.* **18(6)** (2012), 891–898.
[4] R.H.B. Benedict, D. Schretlen, L. Groninger, M. Dobraski, B. Shpritz. Revision of the Brief Visuospatial Memory Test: Studies of normal performance, reliability, and validity. *Psychological Assessment* **8(2)** (1996), 145–153.
[5] D.C. Ciresan, U. Meier, J. Masci, L.M. Gambardella, J. Schmidhuber. *High-Performance Neural Networks for Visual Object Classification*. :12.
[6] A. Krizhevsky, I. Sutskever, G.E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **60(6)** (2017), 84–90.
[7] H.H. Aghdam, E.J. Heravi. *Guide to Convolutional Neural Networks: A Practical Application to Traffic-Sign Detection and Classification* [Internet]. Springer International Publishing; 2017 [cited 2019 Jan 4]. Available from: //www.springer.com/de/book/9783319575490
[8] Custom Vision Service | Microsoft Azure [Internet]. [cited 2019 Jan 4]. Available from: https://azure.microsoft.com/en-us/services/cognitive-services/custom-vision-service/
[9] You can probably use deep learning even if your data isn't that big [Internet]. [cited 2019 Jan 18]. Available from: https://beamandrew.github.io/deeplearning/2017/06/04/deep_learning_works.html