Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200144

Digitalisation of the Brief Visuospatial Memory Test-Revised and Evaluation with a Machine Learning Algorithm

Martin Eduard BIRCHMEIER [*, a, 1], Tobias STUDER [*, a], Prof. Dr. med. Andreas LUTTEROTTI [b], Prof. Dr. Dipl.-Psych. Iris-Katharina PENNER [c], Prof. Serge BIGNENS [a]

[*] *Those authors contributed equally to the work*
[a] *Bern University of Applied Sciences, Biel, Switzerland*
[b] *Department of Neurology, University Hospital Zurich, Zurich, Switzerland*
[c] *COGITO Center, Neurocognition and Neuropsychology, Düsseldorf, Germany*

Abstract. The disease multiple sclerosis (MS) is characterized by various neurological symptoms. This paper deals with a novel tool to assess cognitive dysfunction. 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. We present a way to make this process more efficient, without losing quality by having the patients using a tablet App and having the drawings rated with the use of a machine learning (ML) algorithm. A dataset of 1'525 drawings were digitalized and then randomly split in a training dataset and in a test dataset. In addition to the training dataset the already trained drawings from a preliminary paper were added to the training dataset. The ratings done by two neuropsychologists matched for 81% of the test dataset. The ratings done automatically with the ML algorithm matched 72% with the ones of the first neuropsychologist and 79% of the ones of the second neuropsychologist. For a semi-automated rating we defined a threshold value for the reliability of the rating of 78.8%, under which the drawing is routed for manual rating. With this threshold value the ML algorithm matched 80.3% and 86.6% of the ratings of the first and second neuropsychologists. The neuropsychologists have in that case to manually check 17.4% of the drawings. With our results is it possible to execute the BVMT-R Test in a digital way. We found out, that our ML algorithms have with the semi-automated method the similar matching as the two professional raters.

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

1. Introduction

The disease multiple sclerosis (MS) is characterized by various neurological symptoms, including motor weakness and spasticity, coordination problems, fatigue, sensory, bladder and cognitive dysfunction [1]. This paper deals with a novel tool to assess cognitive dysfunction. The Brief Visuospatial Memory Test-Revised (BVMT-R) [2], which is part of the Brief Internation Cognitive Assessment for MS Test Battery

[1] Bern University of Applied Sciences Engineering and Information Technology, Höheweg 80, CH-2502 Biel / Bienne, E-mail: m.birchmeier@hotmail.com

(BICAMS) is a validated instrument to assess cognition in MS patients [3]. Currently, the BVMT-R is carried out on paper in the presence of a neuropsychologist. Briefly, the patient is presented a page with six geometric figures for ten seconds to memorize them. Then he tries to redraw them from memory considering the correctness and position of each figure. These drawings are then evaluated by a neuropsychologist. A figure, which is both correct in details and position on the paper 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 the rating is 0 points. This is a time-consuming procedure for both players.

In this paper, we present a way to make this process more efficient, without losing quality by having the patients using a tablet App to look at, memorize and draw the figures and having the drawings rated automatically or semi-automatically with the use of a machine learning (ML) algorithm. We have therefore defined the following research question: "Can the drawings of the BVMT-R be rated by a Convolutional Neural Network (CNN) with the same accuracy as a neuropsychologist?".

2. Method

First of all, a mobile app was developed to digitalize the 294 (page of six drawing of geometric figures) paper-based patient data, which were provided by the COGITO GmbH in Düsseldorf and evaluated by their neuropsychologist (N1). Those patient data contain only the form A (see Table 1 below) of the BVMT-R. After eliminating 239 empty drawings, from the total of 294 patient's data at 6 drawings per page, a total of 1'525 drawings remained and were included in this patient dataset.

This dataset was then randomly split in 1'220 drawings (=80%) for the training dataset and 305 (=20%) for the test dataset. In addition to the training dataset the already trained drawings from the preliminary paper [4] were added, which resulted in a total training dataset of 1'790 drawings (see table 1).

Table 1. Number of training data (m) and test data (n) per figure.

Number	Figure	Data
1		m = 352, n = 57
2		m = 309, n = 52
3		m = 320, n = 54
4		m = 267, n = 47
5		m = 281, n = 48
6		m = 261, n = 47
Total		m = 1'790, n = 305

For the ML algorithm a CNN was chosen, because this algorithm has been developed for visual object classifications [5]. The CNN gives as an output a rating of the image with a probability value of the reliability of the rating [6].

Furthermore, the test dataset was revaluated by a second neuropsychologist (N2) from the University Hospital Zurich, to compare the evaluation between two neuropsychologists from different centers. The evaluation of the neuropsychologist from the University Hospital Zurich is also used as external validation unit.

Since the manual rating procedure of a single neuropsychologist is not representing a clear gold standard for the ML algorithm, its rating results cannot rely on the ratings of a single neuropsychologist. Therefore, in order to answer the research question, we use the result of the comparisons between the ML algorithm and the two individual neuropsychologists and consider these results with the comparison between both neuropsychologists among themselves.

Various statistical methods were used to evaluate the ML algorithm: Sensitivity, specificity, the positive and negative prediction value [7], the significance test [8] and the Cohen's kappa test to evaluate the interrater reliability of assessments between the neuropsychologists themselves and the ML algorithm [9].

In a second step, a semi-automated rating methodology was used, keeping the rating of the ML algorithm when the rating reliability was above a defined threshold value and otherwise redirecting the drawing to the neuropsychologist for manual rating.

3. Results

The two neuropsychologists (N1 and N2) rated the test data and they were compared with each other. Their rating was identical for 81% of the drawings. Cohen's Kappa (K) was 0.62.

We can see that our neuropsychologists have an equal match with a Kappa of 0.62 like the scientists of Brazilian study [10].

The agreement between N1 and ML is by 72% (K = 0.45) and between N2 and ML 79% (K = 0.56).

With the second method (semi-automated rating) in order to set the threshold value, a ROC curve was generated per label and comparison from the sensitivity and the specificity.

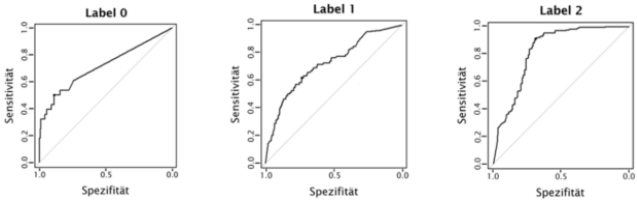

Figure 1. ROC curve per label from the comparison between N1 and ML.

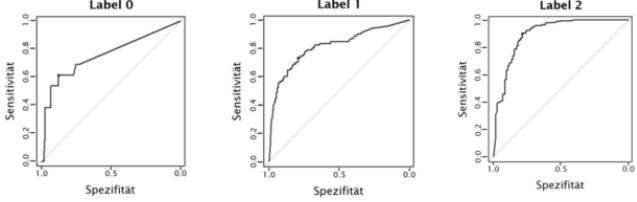

Figure 2. ROC curve per label from the comparison between N2 and ML.

The area below the ROC curve is a measure of the quality of the evaluation per label [11]. This means that if the ML algorithm cannot detect labels, the ROC curve forms a diagonal (50% area below the ROC curve). The better the ML algorithm, the larger the area below the ROC curve. The area below the ROC curve can also be seen as the probability that a label will actually be evaluated from the ML as the label from the neuropsychologist.

Table 2. Area below the ROC curve.

	N1 <-> ML	N2 <-> ML
Label 0	71.5%	75.6%
Label 1	72.0%	81.3%
Label 2	82.6%	89.7%
Average	75.4%	82.2%
Average total	78.8%	

This threshold value was used again to calculate the agreement in percent and in kappa. In addition, the number of drawings that must be checked manually for each threshold value was specified. This shows how the agreement changes and how many drawings have to be checked manually by a neuropsychologist.

With the threshold, the N1 and ML has an agreement of 80.3% (Kappa = 0.62) and the agreement between N2 and ML was 86.6% (Kappa = 0.73). N1 and N2 have to manually check 53 drawings (of the 305).

The significance test is intended to show how many ratings of the ML algorithm do not match with the ratings of N1 and N2 and are not submitted to them for a manually check ($\beta = 20\%$, error type 2) and how many ratings matched, but are submitted to the N1 and N2 for a manually check ($\alpha = 5\%$, error type 1). Follow hypothesis was defined for the significance test: H_0: the ML rated the drawing equal to the neuropsychologist. H_1: the ML rated the drawing not equal to the neuropsychologist.

Table 3. Significance test of N1 <-> ML (left) and N2 <-> ML (right).

	H_0	H_1
H_0	63.0%	19.7%
H_1	9.5%	7.9%

	H_0	H_1
H_0	68.9%	13.4%
H_1	9.8%	7.9%

The error type 1 (cell H_1 / H_0) is on both significance tests over the 5% and not significant. The ML rated 9.5 / 9.8% of the drawings equal to N1 / N2, but with a reliability below the threshold. So those drawings have to be rated manually from a neuropsychologist.

The error type 2 (cell H_0 / H_1) should be as low as possible, because these drawings are not submitted to the neuropsychologist for a manually rating, although they are wrongly assessed. But as we can see, the error type 2 is in both cases under 20%, so β is significant.

Table 4. Overview comparison N1 with ML algorithm and N2 with ML algorithm with threshold of 0% and 78.8%.

	N1 <-> ML		N2 <-> ML	
Threshold (%)	0	78.8	0	78.8
Agreement (%)	72	80.3	79	86.6
Agreement (K)	0.45	0.62	0.56	0.73

4. Discussion

With our results is it possible to execute the BVMT-R Test in a digital way. The rating of the drawings is given by a trained ML algorithm either in a full automated way, or in

a semi-automated way with the possibility to set a threshold value for the reliability of the rating under which the drawing is manually rated by a neuropsychologist.

The biggest challenge to answer the research question was to determine the quality of the ratings received in the test data because there are no rated reference data sets (gold standard rating). With our solution, to measure the matches of the ratings of the test data performed by the COGITO GmbH with the ratings of the University Hospital Zurich, and find out it was 81%, we were then able to compare the results obtained in automated or semi-automated way with our ML algorithm with the results of either one of two neuropsychologists (N1 and N2), results with N1 showing the internal and N2 the external validation.. Consequently, to answer the research question positively the rating of the ML algorithm had to reach at least the same matching as the two professional raters.

With this result it is conceivable to create further projects in this area of science. A more detailed consideration could be a completion of the entire digital BICAMS test set using ML.

In summary, a semi-automated rating with the use of ML algorithms of patient-drawn drawing is possible. The deviation of this solution (matching 80.3% and 86.6% with the two neuropsychologists) is in the same range as the deviation between those two independent professionals (matching of 81%).

5. References

[1] Broicher, Sarah Dinah. Neuropsychologie bei Multipler Sklerose [Internet]. Neuropsychologie bei Multipler Sklerose. 2014 [cited 2019 Feb 27]. Available from: https://www.rosenfluh.ch/media/psychiatrie-neurologie/2014/05/Neuropsychologie.pdf

[2] Benedict, Ralph H. B. Brief Visuospatial Memory Test-Revised | BVMT-R [Internet]. [cited 2018 Dec 28]. Available from: https://www.parinc.com/Products/Pkey/30

[3] Benedict RHB, Amato MP, Boringa J, Brochet B, Foley F, Fredrikson S, et al. Brief International Cognitive Assessment for MS (BICAMS): international standards for validation. BMC Neurol. 2012 Jul 16;12:55.

[4] Birchmeier ME, Studer T. Automated Rating of Multiple Sclerosis Test Results Using a Convolutional Neural Network. Studies in Health Technology and Informatics. 2019;105–108.

[5] Zhang W, Itoh K, Tanida J, Ichioka Y. Parallel distributed processing model with local space-invariant interconnections and its optical architecture. Applied Optics. 1990 Nov 10;29(32):4790.

[6] Aghdam HH, Heravi EJ. Guide to convolutional neural networks: a practical application to traffic-sign detection and classification [Internet]. 2017 [cited 2018 Oct 30]. Available from: http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=1520793

[7] Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol. 2008;56(1):45–50.

[8] Cohen J. Statistical power analysis for the behavioral sciences. [Internet]. 1988 [cited 2019 May 29]. Available from: http://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf

[9] Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960 Apr 1;20(1):37–46.

[10] Caneda MAG de, Cuervo DLM, Marinho NE, Vecino MCA de, Caneda MAG de, Cuervo DLM, et al. The Reliability of the Brief Visuospatial Memory Test - Revised in Brazilian multiple sclerosis patients. Dementia & Neuropsychologia. 2018 Jun;12(2):205–11.

[11] Businger W, Bigler V. Medizinische Statistik [Internet]. Berne Fachhochschule - Technik und Informatik; 2019 [cited 2019 May 26]. Available from: http://moodle.bfh.ch