Developing a Machine Learning Model for Automated Scoring on the Cube in Cognitive Assessments: A Pilot Study

L. Shpeller[1], C. Cadonic[2], S. Phrakonkham[2], P. Nasiopoulos[3] and Z. Moussavi[2]

[1] University of British Columbia, School of Biomedical Engineering, Vancouver, Canada
[2] University of Manitoba, Department of Electrical and Computer Engineering, Winnipeg, Canada
[3] University of British Columbia, Department of Electrical and Computer Engineering, Vancouver, Canada

Abstract— **Psychological assessments are often used to help assess cognitive impairments. Inconsistencies in marking these assessments in general, and in cube drawing tests in particular, can lead to misdiagnoses and irregularity in accurate monitoring of the cognitive status; that can be crucial especially in multi-site studies. As a pilot study, a machine learning model using a convolutional neural network was developed to classify drawn cube shapes as ”correct” or ”incorrect” automatically. Techniques such as K-fold cross validation, image augmentation, and early stopping were used to optimize the model using training data. A model with a final validation accuracy of 85.7% was developed as a proof of concept; suggestions for further improvement are presented in this paper. This model will eventually help to ensure similar scoring across different sites when patients are assessed by different assessors.**

Keywords— **cognitive assessment, convolutional neural network, image processing, machine learning**

I. Introduction

Standardized psychological assessments for cognition, such as the Montreal Cognitive Assessment (MoCA) [1], the Wechsler Memory Scale (WMS) [2], the Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) [3], etc., often include visuospatial tasks in which participants are instructed to reproduce geometrical shapes such as a cube. Based on the accuracy of how participants re-draw the shape, they receive a score from their human assessor. The scoring of such tests can vary slightly among assessors due to the subjective nature of the assessment. Although tests such as ADAS-Cog have proven to be quite accurate for severe cases [4], studies have shown that such assessments may not be able to consistently detect mild cognitive impairments [5]. This may be contributed by the fact that even one score higher or lower in these assessments can be crucial [6], as it significantly affects the evaluation and change in cognition of a patient before and after a treatment. As a result, the variability of scoring among assessors, particularly in multi-site studies, may lead to a systematic bias in the assessments. In this pilot study, our aim is to develop an image processing algorithm to score such drawings automatically avoiding variability in scoring due to its subjective nature.

II. Materials and Methods

A. Data

A data set of 185 cube drawings was collected from a current clinical trial [7] from both MoCA and ADAS-Cog cube drawing part of the assessments. For consistency, all data used for training and measuring the algorithm's accuracy was additionally examined by one trained assessor.

B. Machine Learning Model

Convolutional Neural Networks (CNNs) are a common machine learning model to use when classifying images [8],[9]. The CNN works by applying filters, known as kernels, to the images to extract relevant features needed to classify it [8],[10]. A binary classification system was used to classify correct cubes as 1s and incorrect cubes as 0s. Out of the 185 collected data in this study, the trained assessor scored 118 as incorrect (0) and 67 as correct (1), thus, presenting a lack of balance between the two class sizes. If not handled correctly, an unbalanced data set may affect proper training as the model will have more exposure to one class, thus damaging generalizability of the model.

C. Pre-processing

Pre-processing of the data consisted of cropping each image to a perfect square, scaling to 240 x 240 pixels, converting to grey scale, then using a threshold value ($> 225/255$) to represent the image pixels in a binary matter (0 or 1). Finally, these images were up sampled by randomly selecting correct cubes from the data set using a uniform probability distribution to have an equal number of correct and incorrect cubes to train the model with,thus, reducing bias related to the initially unbalaned data set. A second data set was created by augmenting 20% of the data; horizontally stretching

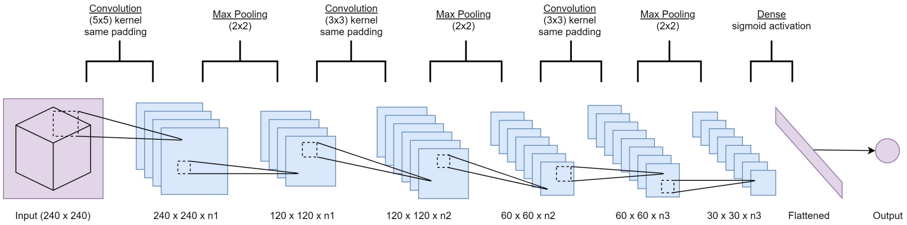


Fig. 1: CNN Diagram

each incorrect cube by a random value and adding a border of 30 pixels around each correct cube, creating a new data set of 222 cubes.

D. Training

Training was done using K-fold cross validation [11] with the keras package and with k = 5 folds using the entire data set. The parameters with the best performance were then trained and tested with an 80-20 ratio and validation accuracy was measured. Early stopping was used to discourage over-fitting of the model by monitoring validation accuracy with patience = 15 epochs.

Parameters were determined by storing trial results in a pandas data frame and selecting the highest performing parameter combination. The parameters chosen were batch size = 90, epochs = 150, learning rate = 0.001, l2 weight regularization = 0.01, dropout rate = 0.1, and sigmoid activation applied to the final layer. The RMSProp optimizer was used along with keras binary cross entropy loss function [12].

Three final models with altered parameter values were created to best suit different data sets. The three data sets were Scaled Data (non-binary) – SD, Augmented Data – AD, and Non-augmented Data – ND. These sets were chosen to represent the importance of having binary data and to understand the effects of the augmentation.

CNN depth was determined from experimentation of smaller networks, proving to have a difficult time learning from the inherent slight bias of the upsampled data set. Adding batch normalization and creating other variations of the layers provided little improvement to our validation accuracy. The greatest improvement was found by increasing the depth of hidden layers, thus arriving at the current architecture.

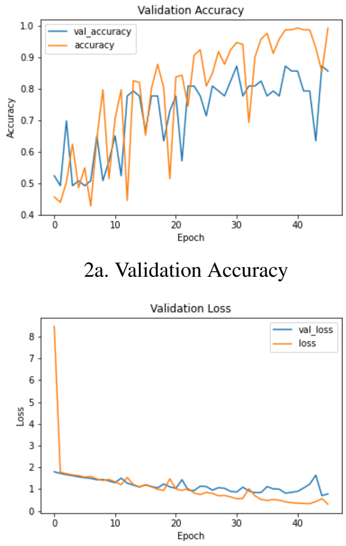
Each model was trained on a Microsoft Surface Book 2 with 1.9 GHz CPU, 16.0 GB RAM, and did not utilize a discrete GPU.

Table 1: Results of final 3 models

Model	ROC AUC	Validation Accuracy Score
SD	85.5	74.6%
AD	91.9	82.4%
ND	91.5	85.7%

III. Results

Various parameters were methodically tested to determine the best model for each data set, and the results can be seen in Table 1. ND was determined to have the highest validation accuracy of 85.7%, showing a high confidence in clearly correct or incorrect cubes as seen in Fig. 3 (99.8% and 0.027%).



2a. Validation Accuracy

2b. Validation Loss

Fig. 2: Graphs displaying Validation Accuracy and Validation Loss

The Receiver Operating Characteristic (ROC) graph has an Area Under Curve (AUC) of 0.92, demonstrating a strong ability to distinguish between the two classes [13]. Figure 2 showcases the validation accuracy and loss as the model is trained. The use of early stopping helped eliminate over-fitting, and a steady increase in accuracy and decrease in loss reveals the simplicity of the image and the model's ability to learn quickly and effectively.

Table 2: Confusion matrices of final 3 models

Model	TN	FP	FN	TP
SD	0.69	0.31	0.19	0.81
AD	0.85	0.15	0.22	0.78
ND	0.78	0.22	0.06	0.94

The ND model displayed a recall of 0.94, precision of 0.81, and misclassification rate of 0.14 for the test data as presented in Table 2. The ND model had a higher performance classifying the positive class (correct cubes). In the context of this study, there is no inherent benefit to having unbalanced classification, meaning further tuning of the model would be beneficial to finding more balance between the classes.

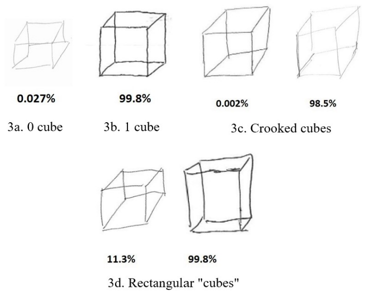
IV. Discussion

For multi-site clinical trials, inter-rater reliability is a major issue of concern if a standardized assessment (i.e. ADAS-Cog, WMS, etc.) is used as the main outcome measure. A score lower or higher can change how the effect of the treatment is marked for a particular participant. This concern is amplified for large multi-site studies as re-scoring and re-evaluation of all assessments by one trained assessor is time-consuming and costly [14]. Additionally, cognitive assessments containing a visuospatial component, such as the Mini-Mental State Examination (MMSE), the most prevalent assessment tool for cognitive impairment in a clinical setting [15][16], retain the aforementioned issues with scoring reliability. Thus, the effect of scoring variation directly impacts the severity of cognitive impairment measured by clinicians and health professionals. For these reasons, we have initial evidence to show that a machine learning model could reliably be used to score subjective visuospatial tasks. In this study, we accomplished this by first training on a simple cube drawing task.

The results of our pilot study show an accuracy of 85.7%. This is encouraging given that inter-rater reliability for a cognitive assessment has been shown to be 88% [17]. However, we do acknowledge that our presented model accuracy is biased since we used the entire data set to derive the parameters of the model while its classification accuracy was determined using k-fold cross-validation with folds = 5. Nevertheless, future studies using larger samples will most likely increase the accuracy and reliability of the automated scoring model.

In designing such an automated scoring model, thresholding, the use of a cutoff value to accentuate pixels assisting with feature extraction in a CNN [18], is of particular importance. In this study, thresholding was used to simplify the cube image and remove the bias related to the shade of the image. The cutoff value of 225/255 was chosen by testing various values and comparing the processed image to the original to determine the highest threshold which properly represented the entire cube. The benefits of thresholding can be seen in Tables 1 and 2 comparing the results of SD with ND.

The architecture of ND model was sufficiently simpler than VGG16 [19] and offered higher performance on this data set. Thus, three layers was determined to be an adequate number of layers to capture the interactions between line segments and patterns between sections of the cube. The choice of sigmoid activation was made to provide an output in the range of 0-1; as well as it presented the highest validation accuracy when compared to other activation methods.



3a. 0 cube

3b. 1 cube

3c. Crooked cubes

3d. Rectangular "cubes"

Fig. 3: 3a. An incorrect cube correctly classified, 3b. A correct cube correctly classified, 3c. Inconsistencies in classifying crooked or tilted cubes, 3d. Inconsistencies in classifying rectangular shaped ”cubes”

Achieving a validation accuracy of 85.7% (Fig. 2a) demonstrates effective model training with the help of early stopping to prevent over-fitting . This consistency can also be seen in decreasing validation loss while maintaining little divergence from training loss in Fig. 2b.

On the other hand, the model was inconsistent in classifying rectangular cubes or those with non-parallel lines, as shown in Fig. 3b and 3c. This inconsistency is a limitation of the training data and the inconsistencies in the initial classifications. The MoCA criteria for scoring the cube drawing are not explicit, requiring an assessor to determine if lines are “parallel enough” or “straight enough”; hence leading to inconsistencies in classifying the cubes. Furthermore, this might relate to why the data augmentation (AD) hurt the performance of the model.

Further experimentation is needed for improving accuracy and reliability. This could include, but is not limited to using He initialization [20], experimenting with optimizers, and using cross-validation before selecting parameters to ensure a high blind test accuracy. To reduce the inconsistencies within classifying the training data, multiple assessors could classify all the data and a voting system could be used to determine the final classifications. This can help bolster the data presented herein, so the model is no longer learning from data from a single trained assessor, but now averages out the assessor bias. Finally, a further experiment could involve having multiple assessors classify a data set and compare these to determine an inter-rater reliability score for human scored assessments. This score could be compared to the reliability of the CNN model to determine its effectiveness and if it is a beneficial supplement to human assessors.

V. Conclusion

The results of this pilot study are encouraging for using machine learning to remove assessor bias and inter-rater variability when scoring visuospatial tasks such as the shape drawing in standard psychological assessments, particularly for drawing a cube. Further investigations could quantitatively determine if the model exceeds the reliability of human scoring in shape drawing, and whether finer tuning of the model's structure could lead to improved validation accuracy and precision. Once improved and tested, such automated scoring for shapes will help with the consistency and reliability of assessments done by different assessors in a large multi-site study; thus, removing the bias and reducing the cost of the studies. Thus, the results herein provide evidence of a first step towards applying machine learning to improve scoring reliability in cognitive assessments containing a visuospatial element.

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgements

The authors would like to thank Cristina Francisco for her assistance with this project. A further thank you to the Natural Sciences and Engineering Research Council of Canada for providing an Undergraduate Student Research Award to finance this project.

References

1. Nasreddine Z, Phillips N, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: abrief screening tool for Mild Cognitive Impairment *J Am Geriatr Soc.* 2005;53(4):695–699.
2. Wechsler D. *Wechsler memory scale*. Psychological Corporation 1945.
3. Rosen W, Mohs R, Davis K. A new rating scale for Alzheimer's disease *The American Journal of Psychiatry.* 1984;141(11):1356–1364.
4. Rutherford G, Lithgow B, Moussavi Z. Short and Long-term Effects of rTMS Treatment on Alzheimer's Disease at Different Stages: A Pilot Study *Journal of Experimental Neuroscience.* 2015;9:33–51.
5. Dautzenberg G, Lijmer J, Beekman A. Diagnostic accuracy of the Montreal Cognitive Assessment (MoCA) for cognitive screening in old age psychiatry: Determining cutoff scores in clinical practice. Avoiding spectrum bias caused by healthy controls *Int J Geriatr Psychiatry.* 2020;35(3):261–269.
6. Ng A, Chew I, Narasimhalu K, Kandiah N. Effectiveness of Montreal Cognitive Assessment for the diagnosis of mild cognitive impairment and mild Alzheimer's disease in Singapore *Singapore Med J.* 2013;54(11):616–9.
7. Moussavi Z, Rutherford G, Lithgow B, et al. Protocol of Repeated Transcranial Magnetic Stimulation for improving cognition in Alzheimer's Disease: A Randomized Double-Blind Placebo Controlled Trial (Preprint) *JMIR Research Protocols.* 2020;10.
8. Krizhevsky A, Sutskever I, Hinton G.E. ImageNet Classification with Deep Convolutional Neural Networks in *Advances in Neural Information Processing Systems* (Pereira F., Burges C. J. C., Bottou L., Weinberger K. Q.. , eds.);25:1097–1105Curran Associates, Inc. 2012.
9. Albawi S., Mohammed T. A., Al-Zawi S.. Understanding of a convolutional neural network in *2017 International Conference on Engineering and Technology (ICET)*:1-6 2017.
10. Zeiler M.D, Fergus R. Visualizing and Understanding Convolutional Networks 2013.
11. Efron B. Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation *Journal of the American Statistical Association.* 1983;78:316-331.
12. Boer PT, Kroese D, Mannor S, Rubenstein R. A Tutorial on the Cross-Entropy Method *Annals of Operations Research.* 2005;134:19–67.
13. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve *Radiology.* 1982;143(1):29–36.
14. Ansell J. Inter-assessor variability: scant data proves the point *Developmental Medicine & Child Neurology.* 2016;58:111-111.
15. Mitchell Alex J. The Mini-Mental State Examination (MMSE): update on its diagnostic accuracy and clinical utility for cognitive disorders in *Cognitive screening instruments*:37–48Springer 2017.
16. Judge Davneet, Roberts Jenna, Khandker Rezaul Karim, Ambegaonkar Baishali, Black Christopher M. Physician practice patterns associated with diagnostic evaluation of patients with suspected mild cognitive impairment and Alzheimer's disease *International Journal of Alzheimer's Disease.* 2019;2019.
17. Peters L, Maathuis K, Kouw E, Hamming M, Hadders-Algra M. Test–retest, inter-assessor and intra-assessor reliability of the modified Touwen examination *European Journal of Paediatric Neurology.* 2008;12:328 - 333.
18. Ker J, Singh S.P, Bai Y, Rao J, Lim T, Wang L. Image Thresholding Improves 3-Dimensional Convolutional Neural Network Diagnosis of Different Acute Brain Hemorrhages on Computed Tomography Scans *Sensors.* 2019;19(9):2167.
19. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition 2015.
20. He J, Lan M, Tan C, Sung S, Low H. Initialization of cluster refinement algorithms: a review and comparative study in *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*;1:297-302 2004.

Author: Laia Shpeller
Institute: University of British Columbia
Street: 2222 Health Sciences Mall
City: Vancouver
Country: Canada
Email: laiashpeller@gmail.com