

# SCREENING OF ALZHEIMER'S DISEASE AND MILD COGNITIVE IMPAIRMENT THROUGH INTEGRATED ON-LINE AND OFF-LINE HOUSE DRAWING TESTS

Nina Hosseini-Kivanani<sup>1</sup>, Elena Salobar-García<sup>2,3</sup>, Lorena Elvira-Hurtado<sup>3,4</sup>, Mario Salas<sup>5</sup>, Christoph Schommer<sup>1</sup> and Luis A. Leiva<sup>1</sup>

<sup>1</sup> University of Luxembourg, Luxembourg

<sup>2</sup> Ramon Castroviejo Institute of Ophthalmologic Research, Spain

<sup>3</sup> Universidad Complutense of Madrid, Spain

<sup>4</sup> Memory Unit, Geriatrics Service, Hospital Clínico San Carlos, Spain

## ABSTRACT

Objective: Evaluate the effectiveness of machine learning (ML) algorithms in classifying mild cognitive impairment (MCI) and Alzheimer's disease (AD) using features derived from the House Drawing Test (HDT). Methods: The HDT was administered to 58 participants, categorized into AD (n = 22), MCI (n= 25), and Healthy Controls (HC, n = 11). Drawings were simultaneously captured using an electronic pen (on-line format) and scanned (off-line format). Results: The models achieved high classification accuracy across all groups: HC vs. MCI (67%), MCI vs. AD (70%), HC vs. AD (76%). Our results showcase the potential of ML for early screening of neurodegenerative diseases.

## KEYWORDS

On-Line, Off-Line, Deep Learning, Alzheimer's Disease, Mild Cognitive Impairment, House Drawing

## 1. INTRODUCTION

Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD) present significant challenges in an aging society, necessitating early diagnosis for effective disease management. MCI, often a precursor to AD, is characterized by a cognitive decline that, while noticeable, does not yet significantly impair daily activities. Identifying MCI as early as possible is essential, as it allows for interventions that may delay the onset of AD. There is thus an urgent demand for diagnostic tools and strategies to facilitate early detection and intervention [Liss et al., 2021; Wimo et al., 2023].

Cognitive assessments have become a focus for early detection of MCI and AD, particularly relying on drawing tasks that assess constructional abilities [Knecht and Lehrner, 2023; Tsatali et al., 2022], such as the Rey-Osterrieth Complex Figure [Cheah et al., 2019], the Clock Drawing Test (CDT) [Cilia et al., 2022], and the House Drawing Test (HDT) [Youn et al., 2021]. However, traditional cognitive assessments, based on pen and paper, are often time-consuming, prompting the development of quicker, semi-quantitative alternatives [Chan et al., 2021; Kobayashi et al., 2022]. Drawing tests based on electronic pens provide more quantifiable metrics (e.g., drawing latency or visual quality) to differentiate between individuals with and without neurodegenerative diseases [Xu et al., 2020; Öhman et al., 2021]. Despite these advancements, little research has compared how traditional (off-line, scanned images) and digital (on-line, time series) drawings perform in practice. While previous work noted that on-line representations offer richer features than off-line data [Bensalah et al., 2023; Cilia et al., 2021], a systematic comparison between these input types for diagnostic accuracy is currently lacking. To bridge this gap, we investigate the impact of data augmentation (DA) on both off-line and on-line representations for MCI and AD screening.

We focus on HDT drawings given the test's complexity and its demand for visuospatial and cognitive planning abilities, as previous research has shown that more complex tasks are more sensitive to early cognitive impairments [Trojano and Gainotti, 2016]. Further, the HDT's requirement for participants to draw from memory, as opposed to copying, places a higher cognitive load, making it a robust tool for early detection of conditions like MCI and AD [Rouleau et al., 1996].

## 2. RELATED WORK

Handwriting analysis<sup>1</sup> has emerged as a cost-effective and reliable method for early detection of AD and MCI. Various studies (e.g., [Ghaderyan et al., 2018]) have used handwriting-based features to differentiate between AD, MCI, and Healthy Controls (HC). However, task effectiveness can vary significantly; for example, symbols like the spiral may not fully capture the fine-grained details of spatial awareness, planning, and memory, which are particularly affected in MCI and AD patients.

Garre-Olmo et al. [2017] analyzed kinematic and pressure features of handwriting in 52 participants. The tasks included drawing of crossed pentagons, spirals, 3D houses, and the CDT. Their study highlighted the potential of on-line features in distinguishing between healthy subjects and those with cognitive impairments. Supporting these findings, [Werner et al., 2006] reported significant differences in temporal measures and pressure among AD, MCI, and HC groups.

Traditional off-line cognitive assessments have primarily focused on identifying outlines and details using scoring systems, often overlooking the sequence of drawing actions. A limited number of studies have employed digital tools, such as pens or tablets, to record the drawing process (see, e.g., [Cilia et al., 2022]). Poreh et al. [2020] used a digital pen to analyze continuity and symmetry variables, offering insights into cognitive functions beyond traditional methods. Similarly, Kim et al. [2020] used a tablet to automatically extract stroke parameters and spatial information. They found that AD patients produced more fragmented drawings, took longer pauses, and demonstrated lower accuracy than individuals with normal cognition.

Our study contributes to the research literature by evaluating the effectiveness of various computational models for detecting AD and MCI (e.g., [Chen et al., 2020; Hosseini-Kivanani et al., 2024]). For example, digital parameters of the CDT have effectively demonstrated cognitive processes and distinguished between patients with amnesic MCI, mild AD, and those with normal cognition [Zhang et al., 2021]. However, specific drawing behaviors in MCI patients remain underexplored.

## 3. METHODOLOGY

We sought to explore cognitive and motor functions through a drawing task designed to assess creativity and precision across different cognitive stages (HC vs. MCI, MCI vs. AD, HC vs. AD). We recruited 58 participants from [redacted], including 11 HC, 25 with MCI, and 22 AD. Cognitive status was assessed using the Mini-Mental State Examination (MMSE). A Chi-square test showed no significant association between sex distribution and diagnosis group,  $\chi^2(1, N = 47) = 2.09, p = 0.148$ , and a  $t$ -test indicated no significant age differences between HC and MCI groups ( $t(36) = 0.65, p > .05$ ) (Table 1). However, a significant difference in MMSE scores was found, indicating lower cognitive function in individuals with MCI compared to healthy controls ( $t(36) = 3.38, p < .05$ ) (Table 1).

---

<sup>1</sup> In this paper, we consider *handwriting* and *hand-drawing* interchangeably, as both involve similar neurophysiological and peripheral processes involved in motor control.

Table 1. Summary of user demographics (Mean  $\pm$ SD) and Age-MMSE correlations

Characteristic	HC (n=11)	MCI (n=25)	AD (n=22)	Total (n=58)
Age (years)	82.6 $\pm$ 2.5	81.4 $\pm$ 5.9	79.4 $\pm$ 4.1	80.9 $\pm$ 5.0
MMSE Score	29.9 $\pm$ 0.8	26.0 $\pm$ 2.1	23.5 $\pm$ 3.6	26.6 $\pm$ 3.3
Gender (F/M)	8/3	15/10	17/5	40/18
Age-MMSE Corr.	-0.16	0.28	-0.23	0.15
<i>p</i> -value	.170	.030	.310	.200

### 3.1 Data Collection and Preprocessing

Participants were instructed to draw a house symbol on a Repaper tablet (dimensions: 10.9 inches)<sup>2</sup> with a blank sheet of paper affixed and using a standard pen equipped with an accelerometer. This setup replicated a typical pen-and-paper drawing experience while capturing digital data via Bluetooth to the Repaper app. A total of 58 drawings were collected. The **on-line** data, representing discrete point sequences, were initially saved as SVG files and then converted to JSON format, containing multivariate sequences of (x,y,t) points. The **off-line** data, captured as high-resolution images using an HP Color LaserJet Pro scanner, were stored as PDFs, converted to PNG format, and resized to 224x224 pixels to standardize inputs for deep learning (DL) models. This resizing aligns with common computer vision practices for compatibility with pretrained DL models. To further enhance image quality, the Canny edge detector was applied to highlight edges in the scanned images.

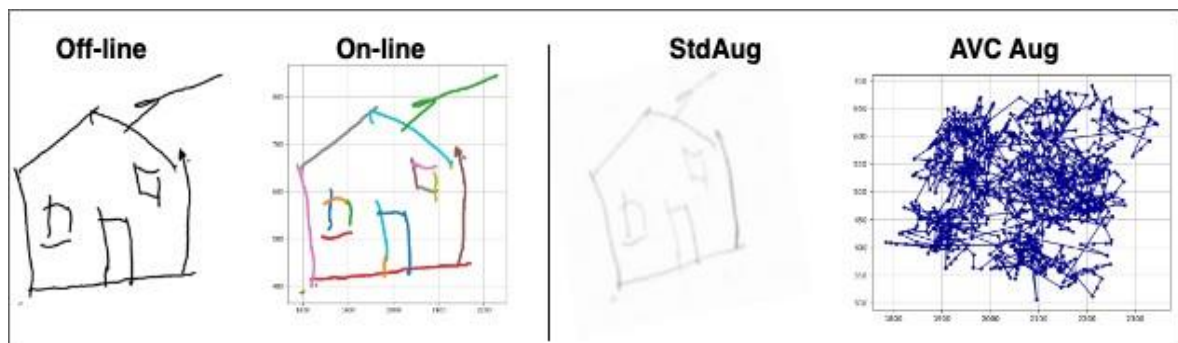


Figure 1. Sample of off-line and on-line drawing with standard and AVC augmentation

**Data Augmentation:** To improve model robustness and generalizability, we applied DA techniques to both on-line and off-line versions. For off-line version, we used geometric transformations such as rotation, translation, scaling, and flipping to increase variability and reduce overfitting. For on-line version, we employed standard techniques such as jittering and, based on recent findings by the AVC technique proposed by [Maslych et al., 2023]. The AVC included Gaussian noise addition, frame-skipping, spatial modifications, perspective adjustments, and scaling. After DA, the dataset included 300 images (off-line version) and 300-point sequences (on-line version), evenly distributed across 100 observations per group.

### 3.2 Experimental Setup

Our study employs Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to analyze handwriting for cognitive impairment assessment. CNNs are used to analyze pixel-based images, detecting spatial patterns and textures that are indicative of subtle cognitive changes. We use three CNN architectures: ResNet50 [He et al., 2016], which employs skip connections to maintain information across deeper layers; DenseNet121 [Huang et al., 2017], known for efficient feature propagation through densely

<sup>2</sup> <https://www.iskn.co/eu>

connected layers; and EfficientNet [Tan and Le, 2019], which optimizes the network’s architecture to handle diverse handwriting styles effectively.

In addition to image-based analysis, RNNs are applied to interpret stroke sequences. This approach captures temporal dynamics and sequential nuances in handwriting, providing insights into the cognitive processes underlying stroke patterns. We implement three types of RNNs: the Bidirectional Vanilla RNN (BiRNN) for straightforward sequential tasks; Bidirectional Long Short-Term Memory (BiLSTM) [Hochreiter and Schmidhuber, 1997] for retaining information across longer sequences; and Bidirectional Gated Recurrent Unit (BiGRU) [Cho et al., 2014], which balances computational efficiency with performance.

**Training Details:** All models were trained using the Adam optimizer with a learning rate of  $\eta = 0.001$  and decay rates of  $\beta_1 = 0.99$  and  $\beta_2 = 0.999$ . We used binary cross-entropy as the loss function for all binary classification tasks (HC vs. MCI, MCI vs. AD, and HC vs. AD). We used a batch size of 32 and up to 100 training epochs, with early stopping (patience of 40 epochs) to avoid overfitting. The augmented dataset was split into 80% for training and 20% for testing, ensuring the test set represented unseen data. Stratified 5-fold cross-validation was conducted on the training set to maintain class proportions across folds. Model performance was evaluated using classification accuracy (Acc.) and the Area Under the ROC Curve (AUC).

## 4. RESULTS AND DISCUSSION

Our experiments are crucial for understanding the progression of cognitive decline and distinguishing between HC, individuals with MCI, and those with AD. As highlighted in previous research [Ding et al., 2022; Werner et al., 2006], distinguishing MCI from HC and AD can be particularly challenging due to overlapping characteristics.

Table 2 summarizes the Accuracy and AUC results for both on-line and off-line datasets, comparing models with and without DA. The data reveal that applying DA, particularly standard DA (StdAug), consistently improves performance across all models and settings.

**Off-line Data:** EfficientNet demonstrated significant performance gains across all binary classification tasks when standard DA was applied. Specifically, for “HC vs. MCI,” accuracy increased from 50%—52% to 65%—66%, for “MCI vs. AD” from 49%—49% to 69%—70%, and for “HC vs. AD” from 53%—55% to 76%—77%. This indicates that off-line data setups benefit substantially from standard DA.

**On-line Data:** BiGRU was the top performer for on-line data, showing marked improvements post-DA. In the “HC vs. MCI” task, performance increased from 51%—54% to 67%—69%, in “MCI vs. AD” from 47%—45% to 70%—72%, and in “HC vs. AD” from 45%—47% to 75%—76%. While on-line data also benefitted from DA, results varied more between models.

**Comparison of DA Techniques:** The comparison between standard DA and AVC DA shows that standard DA generally yields higher performance gains. For example, in the “HC vs. AD” group, GRU achieved similar results with standard DA (75%—76%) and AVC DA (68%—70%), but overall, standard DA consistently outperformed AVC across different settings.

The improvements observed in our study align with existing literature that suggests data augmentation can enhance ML model performance by providing more diverse training data, thereby improving generalization [Shorten and Khoshgoftaar, 2019]. Specifically, our findings underscore that standard DA outperforms more complex techniques like AVC DA, particularly in tasks that require distinguishing subtle cognitive differences, such as between HC and MCI. This suggests that simpler, well-tuned DA methods might be more beneficial for certain medical datasets, where the quality and interpretability of data are paramount [Perez and Wang, 2017].

Table 2. Binary classification results achieved before and after DA (Standard & AVC) for different DL Models

	HC vs MCI					MCI vs AD					HC vs AD				
	Off-line		On-line			Off-line		On-line			Off-line		On-line		
	Before	StdAug	Before	StdAug	AVCaug	Before	StdAug	Before	StdAug	AVCaug	Before	StdAug	Before	StdAug	AVCaug
	(Acc. AUC)	(Acc. AUC)	(Acc. AUC)	(Acc. AUC)	(Acc. AUC)	(Acc. AUC)	(Acc. AUC)	(Acc. AUC)	(Acc. AUC)	(Acc. AUC)	(Acc. AUC)	(Acc. AUC)	(Acc. AUC)	(Acc. AUC)	(Acc. AUC)
Res	42   48	57   49	N/A	N/A	N/A	46   47	63   65	N/A	N/A	N/A	51   50	65   64	N/A	N/A	N/A
Den	49   51	61   63	N/A	N/A	N/A	49   50	60   62	N/A	N/A	N/A	51   50	73   77	N/A	N/A	N/A
Eff	50   52	<b>65   66</b>	N/A	N/A	N/A	49   49	<b>69   70</b>	N/A	N/A	N/A	53   55	<b>76   77</b>	N/A	N/A	N/A
RNN	N/A	N/A	50   51	61   60	56   59	N/A	N/A	50   53	61   65	55   59	N/A	N/A	48   52	72   75	60   58
LSTM	N/A	N/A	50   55	65   66	59   59	N/A	N/A	47   49	65   66	60   57	N/A	N/A	46   52	75   75	59   59
GRU	N/A	N/A	51   54	<b>67   69</b>	59   60	N/A	N/A	47   45	<b>70   72</b>	61   64	N/A	N/A	45   47	<b>75   76</b>	68   70

## 4.1 Limitations and Future Work

Our study has some limitations worth of mentioning. Mainly, the small sample size, which is a pervasive problem in medical studies [Chen et al., 2016; Hosseini-Kivanani et al., 2024; Impedovo and Pirlo, 2018], and the focus on a single type of drawing task may limit the generalizability of our findings. Additionally, the study's reliance on a specific neuropsychological test (the HDT) may not fully capture the diversity of cognitive impairments across different populations and tasks. Future research should explore other cognitive assessment tasks to validate further our findings. Despite these limitations, our results hold promise and could pave the way for future clinical applications using a simple handwriting test as a non-invasive, low-cost method.

## ACKNOWLEDGEMENT

Work supported by the UCM research group (Grupo de Investigación básica en Ciencias de la Visión del IIORC, UCM-GR17-920105), the Horizon 2020 FET program of the European Union (grant CHIST-ERA-20-BCI-001), and the European Innovation Council Pathfinder program (grant 101071147).

## REFERENCES

- Bensalah, A., Parziale, A., De Gregorio, G., Marcelli, A., Fornés, A., & Lladós, J. (2023). I can't believe it's not better: In-air movement for Alzheimer handwriting synthetic generation. *In Proceedings of IGS*.
- Chan, J. Y., Bat, B. K., Wong, A., Chan, T. K., Huo, Z., Yip, B. H., & Tsoi, K. K. (2021). Evaluation of digital drawing tests and paper-and-pencil drawing tests for the screening of mild cognitive impairment and dementia: A systematic review and meta-analysis of diagnostic studies. *Neuropsychology Review, 31*(4).
- Cheah, W.-T., Chang, W.-D., Hwang, J.-J., Hong, S.-Y., Fu, L.-C., & Chang, Y.-L. (2019). A screening system for mild cognitive impairment based on neuropsychological drawing test and neural network. *In Proceedings of IEEE SMC*.
- Chen, H.-Y., Li, D.-C., & Lin, L.-S. (2016). Extending sample information for small data set prediction. *In Proceedings of IIAI-AAI*.
- Chen, S., Stromer, D., Alnasser Alabdallah, H., Schwab, S., Weih, M., & Maier, A. (2020). Automatic dementia screening and scoring by applying deep learning on clock-drawing tests. *Scientific Reports, 10*(1).
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *In Proceedings of EMNLP*.
- Cilia, N., D'Alessandro, T., De Stefano, C., & Fontanella, F. (2022). Deep transfer learning algorithms applied to synthetic drawing images as a tool for supporting Alzheimer's disease prediction. *Machine Vision and Applications, 33*.
- Cilia, N. D., D'Alessandro, T., De Stefano, C., Fontanella, F., & Molinara, M. (2021). From online handwriting to synthetic images for Alzheimer's disease detection using a deep transfer learning approach. *IEEE Journal of Biomedical and Health Informatics, 25*(12).
- Ding, Z., Lee, T., & Chan, A. (2022). Digital cognitive biomarker for mild cognitive impairments and dementia: A systematic review. *Journal of Clinical Medicine, 11*.
- Garre-Olmo, J., Faúndez-Zanuy, M., López-de Ipiña, K., Calvó-Perxas, L., & Turró-Garriga, O. (2017). Kinematic and pressure features of handwriting and drawing: Preliminary results between patients with mild cognitive impairment, Alzheimer's disease, and healthy controls. *Current Alzheimer Research, 14*(9).
- Ghaderyan, P., Abbasi, A., & Saber, S. (2018). A new algorithm for kinematic analysis of handwriting data; Towards a reliable handwriting-based tool for early detection of Alzheimer's disease. *Expert Systems with Applications, 114*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of CVPR*.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8).
- Hosseini-Kivanani, N., Salobrar-García, E., Elvira-Hurtado, L., López-Cuenca, I., de Hoz, R., Ramírez, J. M., Gil, P., Salas, M., Schommer, C., & Leiva, L. A. (2024). Ink of insight: Data augmentation for dementia screening through deep learning. *In Proceedings of ICMHI, Japan, Yokohama*.
- Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *In Proceedings of CVPR*.
- Impedovo, D., & Pirlo, G. (2018). Dynamic handwriting analysis for the assessment of neurodegenerative diseases: A pattern recognition perspective. *IEEE Reviews in Biomedical Engineering, 12*.
- Kim, K. W., Lee, S. Y., Choi, J., Chin, J., Lee, B. H., Na, D. L., & Choi, J. H. (2020). A comprehensive evaluation of the process of copying a complex figure in early-and late-onset Alzheimer's disease: A quantitative analysis of digital pen data. *Journal of Medical Internet Research, 22*(8), e18136.
- Knechtl, P., & Lehrner, J. (2023). Visuoconstructional abilities of patients with subjective cognitive decline, mild cognitive impairment, and Alzheimer's disease. *Journal of Geriatric Psychiatry and Neurology*.
- Kobayashi, M., Yamada, Y., Shinkawa, K., Nemoto, M., Nemoto, K., & Arai, T. (2022). Automated early detection of Alzheimer's disease by capturing impairments in multiple cognitive domains with multiple drawing tasks. *Journal of Alzheimer's Disease, 88*(3).
- Liss, J. L., Seleri Assunção, S., Cummings, J., Atri, A., Geldmacher, D. S., Candela, S. F., & Sabbagh, M. N. (2021). Practical recommendations for timely, accurate diagnosis of symptomatic Alzheimer's disease (MCI and dementia) in primary care: A review and synthesis. *Journal of Internal Medicine, 290*(2).
- Maslych, M., Taranta, E. M., Aldilati, M., & Laviola, J. J. (2023). Effective 2D stroke-based gesture augmentation for RNNs. *In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*.
- Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv:1712.04621*.
- Poreh, A., Levin, J. B., & Teaford, M. (2020). Geriatric complex figure test: A test for the assessment of planning, visual spatial ability, and memory in older adults. *Applied Neuropsychology: Adult, 27*(2), 101–107.
- Rouleau, I., Salmon, D. P., & Butters, N. (1996). Longitudinal analysis of clock drawing in Alzheimer's disease patients. *Brain and Cognition, 31*(1).



- Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1).
- Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *In Proceedings of ICML*.
- Trojano, L., & Gainotti, G. (2016). Drawing disorders in Alzheimer's disease and other forms of dementia. *Journal of Alzheimer's Disease*, 53(1).
- Tsatali, M., Avdikou, K., Gialaouzidis, M., Minopoulou, D., Emmanouel, A., Kouroundi, E., & Tsolaki, M. (2022). The discriminant validity of Rey complex figure test (RCFT) in subjective cognitive decline, mild cognitive impairment, and Alzheimer's disease dementia in Greek older adults. *Applied Neuropsychology: Adult*.
- Werner, P., Rosenblum, S., Baron, G., Heinik, J., & Korczyn, A. (2006). Handwriting process variables discriminating mild Alzheimer's disease and mild cognitive impairment. *Journal of Gerontology Series B*, 61(4).
- Wimo, A., Seeher, K., Cataldi, R., Cyhlarova, E., Dielemann, J. L., Frisell, O., & Dua, T. (2023). The worldwide costs of dementia in 2019. *Alzheimer's & Dementia*, 19(7).
- Xu, F., Ding, Y., Ling, Z., Li, X., Li, Y., & Wang, S. (2020). DCDT: A digital clock drawing test system for cognitive impairment screening. *In Proceedings of ICDE*.
- Young, Y. C., Pyun, J. M., Ryu, N., Baek, M. J., Jang, J. W., Park, Y. H., Ahn, S. W., Shin, H.-W., Park, K.-Y., & Kim, S. Y. (2021). Use of the clock drawing test and the Rey-Osterrieth complex figure test-copy with convolutional neural networks to predict cognitive impairment. *Alzheimer's Research & Therapy*.
- Zhang, X., Zhao, Y., Lv, L., Min, G., Wang, Q., & Li, Y. (2021). A study on the performance characteristics and diagnostic efficacy of digital clock drawing test in patients with amnesic mild cognitive impairment. *Chinese Journal of Behavioral Medicine and Brain Science*, 30.
- Öhman, F., Hassenstab, J., Berron, D., Schöll, M., & Papp, K. V. (2021). Current advances in digital cognitive assessment for preclinical Alzheimer's disease. *DADM*, 13(1).