The Magic Number: Impact of Sample Size for Dementia Screening Using Transfer Learning and Data Augmentation of Clock Drawing Test Images

Nina Hosseini-Kivanani
Dept. of Computer Science
University of Luxembourg
Esch-Belval, Luxembourg
nina.hosseinikivanani@uni.lu

Christoph Schommer
Dept. of Computer Science
University of Luxembourg
Esch-Belval, Luxembourg
christoph.schommer@uni.lu

Luis A. Leiva
Dept. of Computer Science
University of Luxembourg
Esch-Belval, Luxembourg
name.surname@uni.lu

***Abstract*—Dementia is a disease characterized by memory impairment and a gradual disability in performing daily activities. Automated screening for early detection of dementia can lead to more adequate and timely treatment. Our work focuses on predicting various stages of dementia severity using pre-trained Deep Learning (DL) models and a public Clock Drawing Test (CDT) dataset. However, the relationship between sample size and model performance is not yet well understood. This may lead to an overreliance on a large number of samples for model training, which may eventually deter reliable outcomes. We found that the classification performance of DL models tends to plateau once a certain number of samples is reached, therefore, it is possible to work on a small data regime with DL models in this task. This research not only advances the field of medical image analysis for dementia screening but also offers broader implications for DL applications in healthcare. Ultimately, the understanding of how sample size affects model performance can guide future research and support more intelligent and efficient utilization of DL models in addressing complex health-related challenges.**

***Index Terms*—Sample size estimation; Clock Drawing Test; Deep Learning; Alzheimer's disease**

I. Introduction

Dementia is a progressive neurological disorder that causes memory loss and cognitive decline, critically impairing an individual's ability to perform daily activities. The prevalence of dementia is increasing, with Alzheimer's disease (AD) being identified as the most common form, accounting for the majority of cases [1]. Despite substantial research progress, a cure for dementia continues to remain out of reach, highlighting the need for research and development of innovative interventions [1]. The impact of this disease extends beyond the affected individuals, including their families, caregivers, and healthcare systems. According to current estimates, the global cost of dementia reaches one trillion dollars annually and is expected to increase in the future [2].

Detecting dementia typically requires a variety of cognitive tests for neuropsychological impairment [3], among which the Clock Drawing Test (CDT) is a widely used tool. Specifically for the study of AD, the CDT has demonstrated a high diagnostic efficiency [4], [5], especially among the elderly [6], [7]. It assesses the cognitive health of patients through a simple yet revealing task: drawing a clock set to a specific time (usually ten minutes after eleven). The simplicity, non-invasiveness, and intuitiveness of this test make it an accessible tool for assessing cognitive health across diverse populations, including those with limited literacy or physical disabilities.

The evaluation of CDT images involves analyzing the quality of the drawings, specifically the positioning of the numbers and the clock hands. Clinicians use different scoring systems for their assessments, which help determine the severity or progression of dementia. In this paper, we rely on the well-established Shulman six-point scale [8] to classify CDT images with Deep Learning (DL) models.

Recently, significant advancements have been made with regard to the analysis and interpretation of CDT images. Jimenez-Mesa et al. [9] introduced a computer-aided diagnosis system based on DL for automated diagnosis. Similarly, [10] developed a deep neural network (DNN) model using 40000 CDT drawings. Their model achieved an accuracy of 90% in binary classification (impaired vs. control participants), and up to 77% accuracy in identifying individuals with probable dementia. Nevertheless, their research primarily centered on the binary classification problem. In another study by [11], a DNN-based prediction model was designed to detect cognitive decline, effectively distinguishing between cognitively impaired and control participants. All these previous works aimed at automating the scoring process for CDT images, specifically targeting their use for screening purposes, and mostly focusing on binary classification tasks.

A current limitation of the state of the art, as discussed in the next section, is the lack of systematic and comprehensive understanding of the optimal sample size required for deep learning (DL) models, particularly in the context of dementia diagnosis, and the absence of focused studies on the nuanced effects of varying sample sizes on model performance. This is important because it is often assumed that DL requires a large number of samples for model training. At the same time, collecting and labeling CDT images is time-consuming. Therefore, it would be a quite feat if DL models could provide reliable outcomes with small data.

This work concentrates on optimizing the sample size used for DL model training with CDT images. By exploring various sample sizes and their effects on model performance, we aim to improve the efficiency and effectiveness of diagnostic processes for dementia. Our findings shed light on the important topic of sample size in DL model performance, providing a practical roadmap for researchers, clinicians, and practitioners dealing with limited data availability.

II. Related Work

The growing body of research supports the potential of DL models for early and accurate dementia detection, facilitating more timely and effective treatment for patients [12]. For example, previous work has shown promise in the diagnosis of neurological disorders [13] and the prediction of early-stage dementia [14].

Recent research proposed the use of automatic scoring systems as alternatives to traditional manual evaluation techniques [15], [16]. Notably, Chen et al. [16] aimed to automate the CDT scoring process using DL models, reporting an accuracy of 96.65% for binary classification and 72.2% for multi-class classification based on the six-point Shulman scale [8], which categorizes drawings from perfect clock representations to those that are severely disorganized and unidentifiable as clocks (see Figure 2).

However, the process of collecting a large quantity of high-quality data for DL models is both time-consuming and expensive. This represents a significant challenge, especially in medical research where data collection involves strict ethical regulations and privacy concerns [17]. Nonetheless, the relationship between sample size and DL performance is poorly understood. While larger datasets can potentially lead to improved model performance, the results are not guaranteed. Therefore, determining an appropriate sample size is vital, as it significantly affects the robustness, reliability, and generalizability of a model's predictions.

Knowing the adequate amount of training data is essential [17], but few studies have systematically evaluated the impact of sample size on model accuracy [18]. Althnian et al. [19] showed that smaller sample sizes when combined with careful feature selection, can enhance the performance of machine learning (ML) classifiers. This lack of a clear rule emphasizes the need for further research on how the quantity of data influences the performance of ML models in the medical field. Our study aims to contribute to this ongoing topic, examining the influence of data size on the effectiveness of DL models in dementia diagnosis.

Several studies have investigated the impacts of dataset size on the classification performance in the medical domain [18], [20]. For example, Varma and Simon [21] used a dataset comprising only 40 samples and examined the performance of models using two different Cross-validation (CV) approaches for data selection. Their study primarily focused on the choice of validation method. In contrast, Combrisson et al. [22] varied the sample size and used K-fold CV exclusively. Their study reported that with smaller sample sizes, classification accuracy was above the chance level 62.5% with the p-value $< .05$ (in a 2-class or 4-class classification problem). In another study in the medical domain, Althnian et al. [19] prepared three subsets of different sizes and employed a range of metrics to compare the performance of six classic ML models. They concluded that a set of 10 features and a smaller amount of data could enhance the performance of classifiers. Meanwhile, Han et al. [23] took a slightly different approach and investigated the optimal number of feature sets using a random forest classifier. They suggested that optimal data can vary from one dataset to another if no specific pattern is defined. Hence, to address this, they proposed using an out-of-bag error and 'SearchSize' exploration, leading to an improvement in accuracy.

Finally, various studies have tackled the challenge of small datasets by augmenting training sets; see e.g. [24]. However, most of this research has predominantly focused on increasing the data size, with little attention given to examining the impact of the sample size on performance. Mostly, existing research has concentrated on the extent to which the dataset size can affect the classification performance in different domains (e.g., [25], [26]).

In sum, this paper addresses a gap in the current research on using DL models in the clinical domain, in particular for dementia diagnosis. The paper investigates the optimal sample size required for DL models, exploring its impact on model performance. By clarifying the optimal data requirements, this study paves the way for more efficient dementia diagnostic processes, even in settings with limited data resources.

III. Materials and Methods

Our focus is to investigate whether using an optimal dataset size for DL model training can achieve similar, or even improved, classification performance levels for dementia screening, as compared to using the full dataset. Therefore, we sought to develop DL models capable of predicting different stages of dementia severity using an optimal sample size. To achieve this, our approach builds on and replicates the study conducted by Chen et al. [16], who previously demonstrated the efficacy of training DL model on a public CDT dataset.

Given the characteristics of the CDT dataset (small size and imbalanced classes), we used data augmentation techniques, which have been proven effective in generating new training samples, by applying various transformations to the original images [27] to see how the model performance with and without augmentation in different sample sizes will change. This approach is two-fold. Firstly, it helps balance the dataset by generating a range of inputs from which the model can learn. Secondly, it acts as a robust regularization technique, ensuring that the model generalizes well to new data and preventing overfitting [28], [29].

A. Dataset

The CDT dataset we used in this work has 1375 images. The participants' age varied from 18 to 98 years, with an average age of 69.8 years$\pm$14.7 years. Based on the Shulman scoring system, the images have been classified into

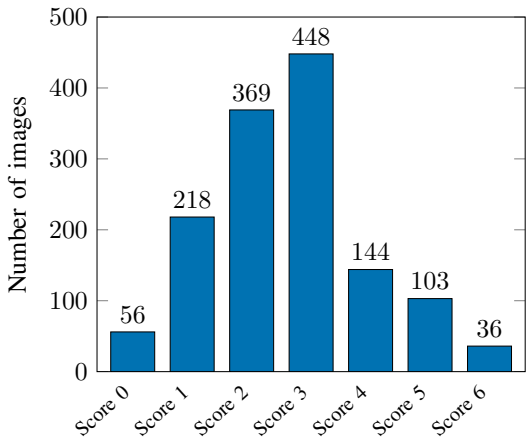


Fig. 1: Total number of drawings for screening and scoring. Score 0 refers to healthy subjects. Score 6 refers to subjects who are unable to draw anything related to a clock.

six categories, each representing different stages of dementia severity. The scores ranged from 1 to 6, each indicative of increasing dementia severity, with 1 indicating low severity and 6 indicating high severity. A score of 0 represents healthy subjects (Figure 1).

During data acquisition, participants were presented with a piece of paper containing a pre-printed circle. They were then instructed to draw the clock numbers from 1 to 12 and to set the clock hands to point to "11:10 o'clock." As illustrated in Figure 2 (with a Shulman Score 0), the clock hands should be positioned on '11' and '2' to accurately represent this time.

1) Data pre-processing: The paper-and-pencil drawings of both patients and healthy participants were scanned in 256-bit grayscale PNG format at 849×1168 px, which were resized to 224×224 px, according to the expected input size of our pre-trained DL models, as explained in the next section. We manually revised all images and removed 20 low-quality images, mostly due to bad scanning and noisy images.

As mentioned before, the dataset is highly imbalanced and not very large for today's standards in DL, so we use data augmentation to address this issue in our study. However, note that not all data augmentations apply in our case. For example, mirroring CDT images would destroy the semantics of the drawings. Similarly, changing the hue or saturation has no effect on those images since they are grayscaled. Therefore, we applied the following operations: scaling, rotation, and translation. These operations produce new, transformed images that help to increase the size and diversity of the training data without compromising its clinical validity [30]. The resulting dataset was perfectly balanced, comprising 448 CDT images per class, or 3136 images overall.

Further, to quantify the quality of the augmented data, we computed the structural similarity index measure (SSIM) [31] of all augmented images against the original images. As shown in Figure 3, the SSIM values are overall between 0.65 to 0.85 with the highest frequency range from 8 to 10, which indicates that the augmented images are not near-duplicates of the original data. Rather, they are new images that, as shown later, eventually helped to improve model performance.

B. Deep learning models

We employed transfer learning to leverage the power of pre-trained DL models in our experiments. Transfer learning is an approach that enables the use of neural network models that have been previously trained on a representative dataset, such as ImageNet [32], to be used as a starting point for solving a related problem by fine-tuning the model on a new dataset [33]. This method helps reduce the need for large computational resources and extensive labeled data, yet still allows us to achieve high performance on the target task.

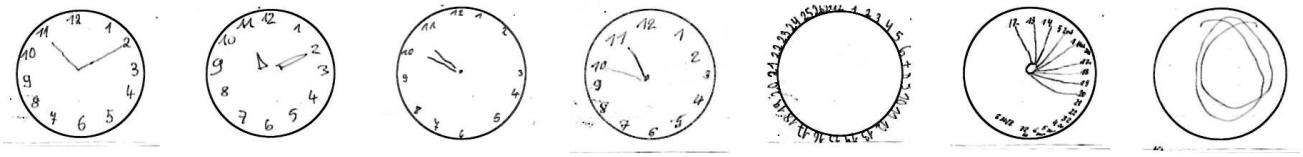
We fine-tuned three Convolutional Neural Networks (CNNs) for binary (healthy vs. non-healthy) and multi-class (six Shulman scores) classification tasks. Binary classification provides a fundamental understanding of the model's capability to differentiate between normal and abnormal cognitive functioning. Conversely, the multi-class classification task aims to distinguish among the six stages of dementia severity according to Shulman's scale, which enables a more detailed understanding of dementia progression, as reflected in the CDT drawings. In the following, we delve into the specifics of the CNN architectures used:

- VGG-16 [34]: It comprises 16 CNN layers with a 3x3 kernel and three subsequent fully connected (FC) layers, was designed by the Visual Geometry Group (VGG) at Oxford University. It has garnered recognition due to its straightforward architecture and efficiency in extracting features.
- ResNet-152 [35]: It comprises 152 CNN layers, providing flexibility and a smaller parameter count compared to models like VGG. It effectively minimizes the error rate to 3.5% and owes its remarkable performance to the use of skip connections.
- DenseNet-121 [36]: It comprises 121 CNN layers, ensuring that every layer has a direct connection to the outputs of all the layers preceding it. It also comprises DenseBlocks interconnected by transition layers.

IV. Sample size analysis

The size of a dataset plays a central role in ML, enabling the effective training and testing of predictive models. A frequent question that often arises is how much data is sufficient or required for model training, and this remains an open challenge [37]. The answer to this question is not straightforward, as it involves finding a balance influenced by various factors. These include the complexity of the task, the diversity present within the data, and the sophistication of the chosen model. Further, small-scale studies carry a higher likelihood of committing either type I or II errors, thereby reducing the probability of identifying true effects [38].

Therefore, it is critical to identify and apply strategies that can reduce data requirements without excessively compromising the model performance. How to effectively use smaller



(a) Score 0 (b) Score 1 (c) Score 2 (d) Score 3 (e) Score 4 (f) Score 5 (g) Score 6

Fig. 2: Sample drawings of the CDT dataset that was used in this work, scored according to Shulman's scale.

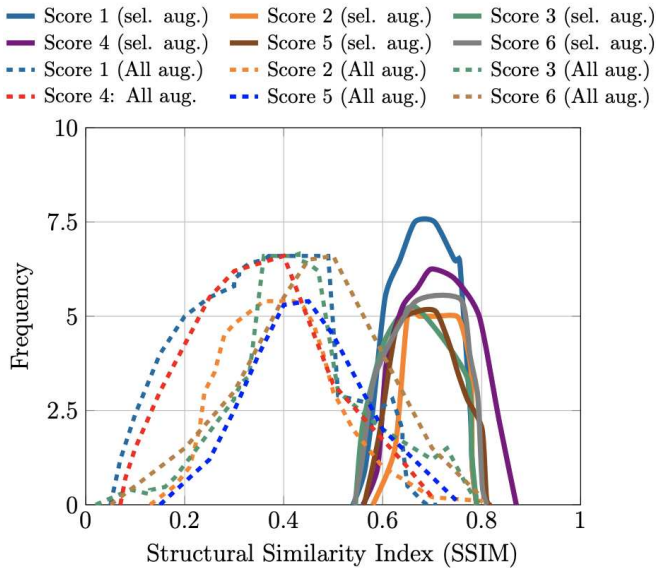


Fig. 3: SSIM distributions. Dashed plots correspond to the results considering all the augmentation (All aug.) techniques collectively. The selected augmentation techniques are Rotation, Scaling, and Translation offset.

datasets is especially beneficial for researchers with restricted access to large datasets. We aim to demonstrate how these strategies can be effectively applied to deal with the limitations imposed by smaller datasets while ensuring robust model performance. We will report the usual model performance metrics to assess the efficacy of our approach: classification accuracy and area under the Receiver Operating Characteristic curve (AUC). These metrics will provide a comprehensive understanding of the model's capacity to correctly classify and distinguish between different stages of dementia severity.

A. Learning Curves

A learning curve (LC) is a graphical representation that illustrates the performance of a model over time [39]. The LC serves beyond merely visualizing the current performance of the model. It can also be used as a predictive tool. Once the LC has been established, we can extrapolate to estimate the accuracy of the model if it is to be trained on the entirety of the available training data. This allows us to predict the potential performance using additional data.

In this work, we generated LCs for accuracy and AUC to visually assess the model's improvement and to identify any potential areas where performance may decline. This can help researchers to use data wisely, e.g. to decide whether to stop or continue model training based on the observed performance over different data splits.

B. Model Training

We selected training subsets corresponding to splits of 10%, 25%, 50%, 75%, and 95% of the entire dataset for our study. A training split of 10% was fixed in each case, except for the 95% training split where the test split was set to 5%. For each of the selected subsample splits, the model was trained from scratch. We repeat this procedure five times, using different initialization seeds for each split to verify the validity of our results and consider any variability that may occur during individual training iterations.

To ensure consistency, we used the experimental setup outlined in the study conducted by Chen et al. [40]. Our CNN models were trained using the Adam optimizer algorithm, with learning rate values ranging from 0.0001 to 0.1, while maintaining a fixed batch size of 16. Additionally, the cross-entropy loss function was used in the training process to measure the performance of a classification model.

Additionally, the recorded performance metrics facilitated the plotting of the LC in the subsequent stages of our experiment. This analysis enables us to effectively analyze, interpret, and optimize the performance of our models. This also helps us understand the trade-off between computational cost and model performance, thereby allowing us to maximize the efficiency of our model given the available data.

V. Results and Discussion

TABLE I: Best and second-best Accuracy and AUC results overall, relative to the total size of the CDT dataset.

	Binary Classification		**Multi-class Classification**	
Classifier	Accuracy @ Sample size		Accuracy @ Sample size	
VGG-16	0.97 @ 100%	0.95 @ 95%	0.68 @ 100%	0.68 @ 95%
ResNet-152	0.97 @ 100%	0.88 @ 95%	0.71 @ 100%	0.71 @ 95%
DenseNet-121	0.98 @ 100%	0.98 @ 95%	0.77 @ 100%	0.77 @ 95%

The results for both binary and multi-class classification can be found in Table I. DenseNet-121 performed the best, with almost 98% accuracy for binary classification, even with the small sample size of 1585 images (50% of the data, after data augmentation). For multi-class classification experiments,

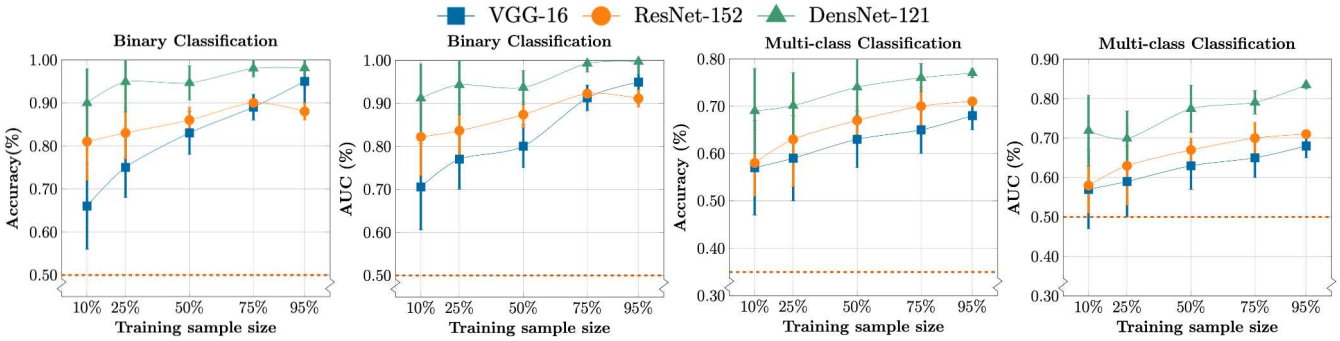


Fig. 4: LC for binary classification (two leftmost plots), and multi-class classification (two rightmost plots), showing model accuracy and AUC results vs. training sample size. The dashed orange lines represent the performance of a random classifier.

ResNet-152 achieved an accuracy of 88% for binary classification using 95% of the data. These results suggest that these models can provide competitive performance when fine-tuned on small data samples.

The LCs in Figure 4 are aligned with common trends observed in previous studies (e.g., [22]), where a swift performance enhancement was noticeable with the expansion of the training set size. For sample sizes larger than 50%, the performance of the DenseNet model stabilized and reached the second-highest accuracy when using the whole dataset (Table I).

In terms of binary classification accuracy, our analysis did not reveal any statistically significant differences when using DenseNet and 25% of the dataset compared to using more data. This was confirmed by a chi-squared test of proportions ($\chi^2(4, N = 417) = 13.13, p > .05$). Similarly, for multi-class classification accuracy, no significant differences were noted when employing 50% of the dataset compared to higher data splits ($\chi^2(4, N = 834) = 2.28, p > .05$). These findings suggest that using all the available data may not always be crucial to obtain the best performance results. For architectures like VGG and ResNet, a minimum of 75% of the dataset seems necessary to attain peak performance.

In contrast to previous studies, that employed ML models using all the available data, our experiments examined the role of sample size and model performance across varying sample sizes. In a nutshell, if the performance of a classifier is good enough with only a subset of the dataset, then such a classifier can be used in conditions where limited data is present, which is quite frequent in clinical domains.

Although the emphasis in previous research has been on how increasing data size improves the ability of CNNs (e.g., [37]), there is a big gap in studies regarding the optimized sample sizes. Our results showed that with 75% of the data we reached to the highest accuracy in all CNN models but DenseNet required only 50% of the data to achieve statistically similar results as when using all the data. Until now, it was expected that adding more labeled data would improve model performance, but we have shown that we do not need as much data.

The advantage of applying DL on a small dataset is quite apparent, given that CNNs are highly data-dependent and usually necessitate more computational costs than traditional ML models. Although our findings are specific to a public CDT dataset and a handful of state-of-the-art CNN models used in a previous study (i.e. [40]), we believe that our results provide valuable insights into the understanding of selecting the optimum sample size for the development of improved DL models.

VI. Conclusion and Future Work

We have investigated the impact of sample size and the performance of DL models (state-of-the-art CNN classifiers). Our findings indicate that classification accuracy and AUC tend to plateau once a certain number of samples are reached. Therefore, we do not really need to use all the available data for training. For multi-class classification, results suggest we may need a larger portion of the dataset, even though the results shown in Tables I report very similar figures. Future research should explore other datasets together with other DL architectures. This way, further improvements could be achieved in terms of predictive capabilities for the early detection of dementia.

References

[1] M. Monica Moore, M. Díaz-Santos, and K. Vossel, "Alzheimer's association 2021 facts and figures report," *Alzheimer's Association*, 2021.

[2] Alzheimer's Association, "2022 Alzheimer's disease facts and figures," *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, vol. 18, no. 4, pp. 700–789, Apr. 2022.

[3] E. Kaplan, "The process approach to neuropsychological assessment of psychiatric patients." *The Journal of Neuropsychiatry and Clinical Neurosciences*, 1990.

[4] K. S. Lee, E. A. Kim, C. H. Hong, D.-W. Lee, B. H. Oh, and H.-K. Cheong, "Clock drawing test in mild cognitive impairment: quantitative analysis of four scoring methods and qualitative analysis," *Dementia and Geriatric Cognitive Disorders*, vol. 26, no. 6, pp. 483–489, 2008.

[5] B. J. Mainland and K. I. Shulman, "Clock drawing test," *Cognitive screening instruments: A practical approach*, pp. 67–108, 2017.

[6] S. Bandyopadhyay, J. Wittmayer, D. J. Libon, P. Tighe, C. Price, and P. Rashidi, "Explainable semi-supervised deep learning shows that dementia is associated with small, avocado-shaped clocks with irregularly placed hands," *Scientific Reports*, vol. 13, no. 1, p. 7384, May 2023.

[7] S. Kim, S. Jahng, K.-H. Yu, B.-C. Lee, and Y. Kang, "Usefulness of the Clock Drawing Test as a Cognitive Screening Instrument for Mild Cognitive Impairment and Mild Dementia: an Evaluation Using Three Scoring Systems," *Dementia and Neurocognitive Disorders*, vol. 17, no. 3, pp. 100–109, Dec. 2018.

[8] K. I. Shulman, D. Pushkar Gold, C. A. Cohen, and C. A. Zucchero, "Clock-drawing and dementia in the community: a longitudinal study," *International journal of geriatric psychiatry*, vol. 8, no. 6, pp. 487–496, 1993.

[9] C. Jimenez-Mesa, J. E. Arco, M. Valentí-Soler, B. Frades-Payo, M. Zea-Sevilla, A. Ortiz, M. Ávila-Villanueva, D. Castillo-Barnes, J. Ramírez, T. del Ser-Quijano *et al.*, "Automatic classification system for diagnosis of cognitive impairment based on the clock-drawing test," in *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 2022, pp. 34–42.

[10] K. Sato, Y. Niimi, T. Mano, A. Iwata, and T. Iwatsubo, "Automated evaluation of conventional clock-drawing test using deep neural network: Potential as a mass screening tool to detect individuals with cognitive decline," *Frontiers in neurology*, vol. 13, p. 896403, 2022.

[11] Y. C. Youn, J.-M. Pyun, N. Ryu, M. J. Baek, J.-W. Jang, Y. H. Park, S.-W. Ahn, H.-W. Shin, K.-Y. Park, and S. Y. Kim, "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*, vol. 13, no. 1, pp. 1–7, 2021.

[12] L. G. Apostolova and P. M. Thompson, "Mapping progressive brain structural changes in early alzheimer's disease and mild cognitive impairment," *Neuropsychologia*, vol. 46, no. 6, pp. 1597–1612, 2008.

[13] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[14] D. Aschwanden, S. Aichele, P. Ghisletta, A. Terracciano, M. Kliegel, A. R. Sutin, J. Brown, and M. Allemand, "Predicting cognitive impairment and dementia: A machine learning approach," *Journal of Alzheimer's Disease*, vol. 75, no. 3, pp. 717–728, 2020.

[15] S. Amini, L. Zhang, B. Hao, A. Gupta, M. Song, C. Karjadi, H. Lin, V. B. Kolachalama, R. Au, and I. C. Paschalidis, "An Artificial Intelligence-Assisted Method for Dementia Detection Using Images from the Clock Drawing Test," *Journal of Alzheimer's Disease*, vol. 83, no. 2, pp. 581–589, 2021, publisher: IOS Press BV.

[16] S. Chen, D. Stromer, H. A. Alabdalrahim, S. Schwab, M. Weih, and A. Maier, "Automatic dementia screening and scoring by applying deep learning on clock-drawing tests," *Scientific Reports 2020 10:1*, vol. 10, no. 1, pp. 1–11, Nov. 2020, publisher: Nature Publishing Group.

[17] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLOS ONE*, vol. 14, no. 11, p. e0224365, Nov. 2019.

[18] I. Balki, A. Amirabadi, J. Levman, A. L. Martel, Z. Emersic, B. Meden, A. Garcia-Pedrero, S. C. Ramirez, D. Kong, A. R. Moody *et al.*, "Sample-size determination methodologies for machine learning in medical imaging research: a systematic review," *Canadian Association of Radiologists Journal*, vol. 70, no. 4, pp. 344–353, 2019.

[19] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. B. Dris, N. Alzakari, A. Abou Elwafa, and H. Kurdi, "Impact of dataset size on classification performance: an empirical evaluation in the medical domain," *Applied Sciences*, vol. 11, no. 2, p. 796, 2021.

[20] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PloS one*, vol. 14, no. 11, p. e0224365, 2019.

[21] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–8, 2006.

[22] E. Combrisson and K. Jerbi, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," *Journal of neuroscience methods*, vol. 250, pp. 126–136, 2015.

[23] S. Han and H. Kim, "On the optimal size of candidate feature set in random forest," *Applied Sciences*, vol. 9, no. 5, p. 898, 2019.

[24] H.-Y. Chen, D.-C. Li, and L.-S. Lin, "Extending sample information for small data set prediction," in *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 2016, pp. 710–714.

[25] J. G. A. Barbedo, "Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification," *Computers and electronics in agriculture*, vol. 153, pp. 46–53, 2018.

[26] X. Zhu, C. Vondrick, C. C. Fowlkes, and D. Ramanan, "Do we need more training data?" *International Journal of Computer Vision*, vol. 119, no. 1, pp. 76–92, 2016.

[27] J. Wang, L. Perez *et al.*, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Networks Vis. Recognit*, vol. 11, no. 2017, pp. 1–8, 2017.

[28] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, 2022.

[29] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[30] N. Hosseini-Kivanani, E. Salobrar-García, L. Elvira-Hurtado, I. López-Cuenca, R. de Hoz, J. M. Ramírez, P. Gil, M. Salas, C. Schommer, and L. A. Leiva, "Better together: Combining different handwriting input sources improves dementia screening," in *Proc. eScience: AI4Health*. IEEE, 2023.

[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[33] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[36] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[37] S. Calderon-Ramirez, D. Murillo-Hernandez, K. Rojas-Salazar, D. Elizondo, S. Yang, A. Moemeni, and M. Molina-Cabello, "A real use case of semi-supervised learning for mammogram classification in a local clinic of costa rica," *Medical & biological engineering & computing*, vol. 60, no. 4, pp. 1159–1175, 2022.

[38] E. C. Carter, F. D. Schönbrodt, W. M. Gervais, and J. Hilgard, "Correcting for bias in psychology: A comparison of meta-analytic methods," *Advances in Methods and Practices in Psychological Science*, vol. 2, no. 2, pp. 115–144, 2019.

[39] C. Perlich, "Learning curves in machine learning," in *Encyclopedia of Machine Learning*, 2010.

[40] S. Chen, D. Stromer, H. A. Alabdalrahim, S. Schwab, M. Weih, and A. Maier, "Automatic dementia screening and scoring by applying deep learning on clock-drawing tests," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.