LEX LOCALIS-JOURNAL OF LOCAL SELF-GOVERNMENT
ISSN:1581-5374 E-ISSN:1855-363X
VOL. 24, NO. 4 (2026)

A DEEP SIMILARITY LEARNING FRAMEWORK FOR AUTOMATED VISUOSPATIAL COGNITIVE ASSESSMENT AND EARLY COGNITIVE IMPAIRMENT SCREENING USING SIAMESE NETWORK

S Jeyanthi[1], Dr. Gayathri K S[2]

[1]Research Scholar, Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India.
[2]Associate Professor, Department of Information Technology, Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India. Email ID: gayathriks@ssn.edu.in

Corresponding author Email ID: prabakaran.jeyanthi@gmail.com

Abstract
The brain network which is responsible for visuospatial abilities is mainly controlled by the parietal and occipital lobes. The damages in the occipital lobe represents the individual's vision disability which fails to recognize the objects. The damages in the parietal lobe causes an inability to interact with or manipulate objects in the daily activities. Existing visuospatial conventional approaches suffer from clinician subjectivity, lower reproducibility, time consuming, manual assessment and scoring. This guarantees the development of the automated assessment tools for visuospatial ability assessment. Our proposed Siamese backbone ResNet50 approach evaluates the individuals drawing like infinity loop and 3D cube by comparing with the reference images, then automatically award the marks based on the image similarity scores. This score helps to identify the cognitive impairment at the early stage. Experimental results demonstrate that the proposed approach provides accurate, consistent and fine-grained visuospatial assessment, outperforming traditional subjective scoring methods. The proposed Siamese with ResNet50 network provides a scalable and reliable solution for early cognitive screening, longitudinal observing and remote assessment of visuospatial deficits.

Keywords: Visuospatial Assessment, cognitive, similarity calculation, deep learning, ResNet50

1 Introduction

A disability in multiple cognitive domains which consist of memory, executive motor function, language, attention, perception and visuospatial cognitive ability is called neurodegenerative disorder can be categorized into dementia and Alzheimer diseases. According to global epidemiological scientific reports, the occurrence of dementia cases is increasing quickly due to an aging population, posing significant clinical, social and economic challenges in worldwide. Due to the critical in the early detection remains in delaying disease progression and affecting the patient's quality of life. Though memory impairment has conventionally been highlighted in dementia diagnosis, growing evidences are indicate that visuospatial dysfunction often manifests in the early stages of Alzheimer's disease and related dementias. Consequently, sensitive, adaptive and automated assessment of visuospatial abilities has emerged as an crucial component of cognitive assessment and its severity evaluation.

Visuospatial ability refers to the perceive [1], analyze and perceptually manipulate spatial relationships among objects. It encompasses skills such as spatial orientation, geometric reproduction, distance judgment, depth perception and visual construction [5]. These functions are primarily associated with the parietal and occipital cortical regions of the cerebrum which are vulnerable to neurodegenerative changes in dementia [2]. Impairment in cognitive function can lead to lack of memory and recall [3], lack of language and communication, difficulties in navigation, misjudgement of object placement, facing difficulties in driving and challenges in performing activities of daily living. Therefore, accurate evaluation of visuospatial performance not only supports diagnosis but also provides insights into functional independence and safety risks in affected individuals.

Clock Drawing Test (CDT), Cube Copying Test, Rey–Osterrieth Complex Figure (ROCF) Test and Infinity Loop Test are the conventional visuospatial assessment methods. Among these methods, the Clock Drawing and Cube Copying Test (CCT) [16] are widely agreed assessment tools for the reason

that its simplicity, rapid administration and minimal resource requirements. In the image-based clock drawing assessment, individuals are asked to draw the face of a clock in a circle shape, asked to write the numbers from 1 to 12 in the appropriate places and draw all the clock hands for the mentioned time. An individual's assessment performance is evaluated based on structural organization, their numerical placement, symmetric and hand orientation. Likewise, in the geometric copying tasks, evaluate constructional ability by demanding patients to replicate three-dimensional figures like cubes or intersecting or infinity loop. Though these tests are clinically authenticated and easy to manage, their scoring processes are largely depend on subjective or expert judgment. Inter-rater variability, limited geometric precision and coarse grading scales reduce sensitivity, particularly in detecting subtle impairments during early disease stages.

Rey Osterrieth Complex Figure Test (ROCF) [13] is widely used assessment method for visuo-constructional ability [5][11], visual memory [8], planning and spatial orientation. It is sensitive to functional decline in multiple cognitive domains in dementia, brain diseases and psychiatric disorders. Battery for Visuospatial Abilities (BVA) offers domain-specific assessment, distinguishing between visuo-perceptual [6] and representational skills and is particularly useful in conditions like ALS 9. Corsi Block-Tapping Test [14] and Design Organization Test (DOT) [17] assess visuospatial memory and organization, with performance improving through adolescence and correlating with executive function. New tools like the Vestibular Cognition Assessment System (VCAS) and mobile-based apps [12] for Parkinson's disease [2] provide multidimensional, dynamic and user-friendly assessments, overcoming limitations of traditional paper-based tests. VR-based assessments increase ecological validity, allow for 3D spatial evaluation and can reduce cognitive load, making them promising for both assessment and training. VisMET [15], a Passive Eye-Tracking enables sensitive, efficient and non-invasive detection of visuospatial memory impairment, particularly useful for early Alzheimer's disease screening.

Visuo-cognitive [7] skills comprise abilities related to vision [18] and thinking. These abilities indicate cognitive in dementia, but few research has observed them in older Indian adults. Understanding these skills can help with early detection and treatment. This study used data from 1,271 urban individuals aged 45 and older in Bangalore, India. Participants' visuo-cognition was assessed using Visuospatial span (VS), Visual Attention (VA), Modified Taylor Complex Figure (MTCF) Copy, Immediate Recall(IR), Delayed Recall (DR) and visuospatial domain from Addenbrooke's Cognitive Examination (ACE) III [19]. Results showed that among the group, females outperformed males in some areas and education positively impacted visuo-cognitive performance. Age and education factors have influenced various visuo-cognitive components, though gender also had precise impacts on visual attention and MTCF measures.

Visuospatial processing speed is important for daily tasks like driving and walking and is thought to peak in early adulthood. A new task called the visuospatial processing speed (VIPS) [9] was created, combining two existing tests. This task, tested with 86 young adults, showed significant correlations with attention and memory abilities. A larger group of 4395 participants exposed that performance is very high in the early 20s and declines with age, doubling in time taken by those in their 60s.

Image similarity [18] is about finding similar images based on a reference in the CIFAR10 dataset which is a computer-vision dataset used for object recognition. SimNet, a deep multi-scale Siamese network, is trained with positive and negative image pairs using an online pair mining strategy. A multi-scale CNN combines embeddings from different layers, showing improved performance in capturing fine-grained image similarities compared to traditional CNNs. a few versions of distance metrics Euclidean, cosine and L1 (city-block) distance has been used, Euclidean out perform all other metrics. Traditional visuospatial assessment methods such as spiral, infinity and cube copying [16], clock drawing, figure identification, occluded or partially visible letter identification and dot-counting exercises are primarily rely on paper-pencil based or visually guided tasks. These assessments are very simple and easy to administer, cost-effective and widely accepted in medical practices. They provide an intuitive measure of spatial construction, visual perception and motor coordination, these measures constructing them fit for rapid bedside assessment and early cognitive screening.

Notwithstanding these advantages, conventional visuospatial assessments suffer from several inherent limitations due to paper-based assessment. Manual scoring is largely depend on subjective and typically on clinician expertise or knowledge, it is dominated by inter-rater variability and partial reproducibility.

These conventional assessments were naturally employ coarse, rule-based marking schemes that fail to capture subtle structural distortions, partial inaccuracies and fine-grained spatial abnormalities. Additionally, traditional approaches do not support automated assessment and analysis, large-scale deployment or remote assessment, restricting their applicability in telemedicine and longitudinal monitoring settings.

In this context, the present study proposes an automated deep learning and computer vision based visuospatial assessment framework for early dementia screening. It provides new opportunities to address these challenges by enabling objective, data-driven analysis of visuospatial tasks. In this work, a deep learning framework has been proposed to automate visuospatial cognitive [20] assessment using multiple visually driven tasks. The framework integrates drawing-based figure analysis using a Siamese ResNet network for similarity comparison of infinity and cube. By combining complementary visuospatial tasks and leveraging modern deep learning architectures, the proposed system aims to deliver a scalable, consistent and fine-grained assessment of visuospatial abilities, overcoming the limitations of traditional manual evaluation methods.

The system digitizes these tasks, including the Infinity and Cube Copying Test and furthermore applies advanced image pre-processing, feature embedding [19] and similarity measurement techniques to evaluate structural accuracy. By computing quantitative similarity metrics and geometric consistency measures, the framework generates objective scores and classifies impairment severity levels. The proposed approach aims to enhance early detection sensitivity, reduce clinician bias and provide scalable solutions for large-scale screening and remote cognitive assessment.

The primary contributions of this work are threefold. First, it introduces a digital pipeline for automated visuospatial drawing analysis using deep neural feature representations. Second, it develops a similarity-driven scoring mechanism that maps geometric distortions to clinically interpretable marks. Third, it validates the framework using standard evaluation metrics, thereby establishing its potential utility in real-world dementia screening scenarios. By integrating artificial intelligence with established neuropsychological testing paradigms, this research advances the development of objective, reproducible and clinically meaningful cognitive assessment tools.

2 A Deep Similarity Learning and Evaluation Framework

This research proposes an automated image-based drawing assessment using a deep similarity learning and evaluation framework involves a Siamese network with CNN backbone Resnet50. A ResNet50 is a deep learning architecture that belongs to the family of Convolutional Neural Networks (CNNs). It has been introduced to address the degradation and vanishing gradient problems encountered in very deep CNN models through the use of residual or skip connections. The proposed architecture performs hierarchical feature extraction from images by learning or extracting spatial patterns such as edges, textures, contours and high-level semantic representations. Therefore, ResNet50 itself is fundamentally a CNN-based architecture has been designed for image recognition and feature learning tasks.

A deep learning framework, the Siamese Network is composed of two or more identical subnetworks that share the same architecture and weights to perform their tasks. The main objective of a Siamese architecture is to learn similarity or distance relationships between image pairs. As an alternative, it directly performing a classification task, generate feature embeddings from paired inputs and compares them using similarity measures such as cosine similarity, Euclidean distance, hybrid or contrastive distance metrics.

In computer vision-based applications, including similarity analysis in the images, signature verification, face recognition, medical image matching and cognitive drawing assessment, ResNet model is generally used as the backbone for feature extractor within a Siamese architecture. In this configuration, two identical ResNet branches process two input images independently while sharing parameters. The extracted feature vectors are then compared to determine the degree of similarity between the inputs. The operational framework of a ResNet-based Siamese Network can be represented in the fig. 1.

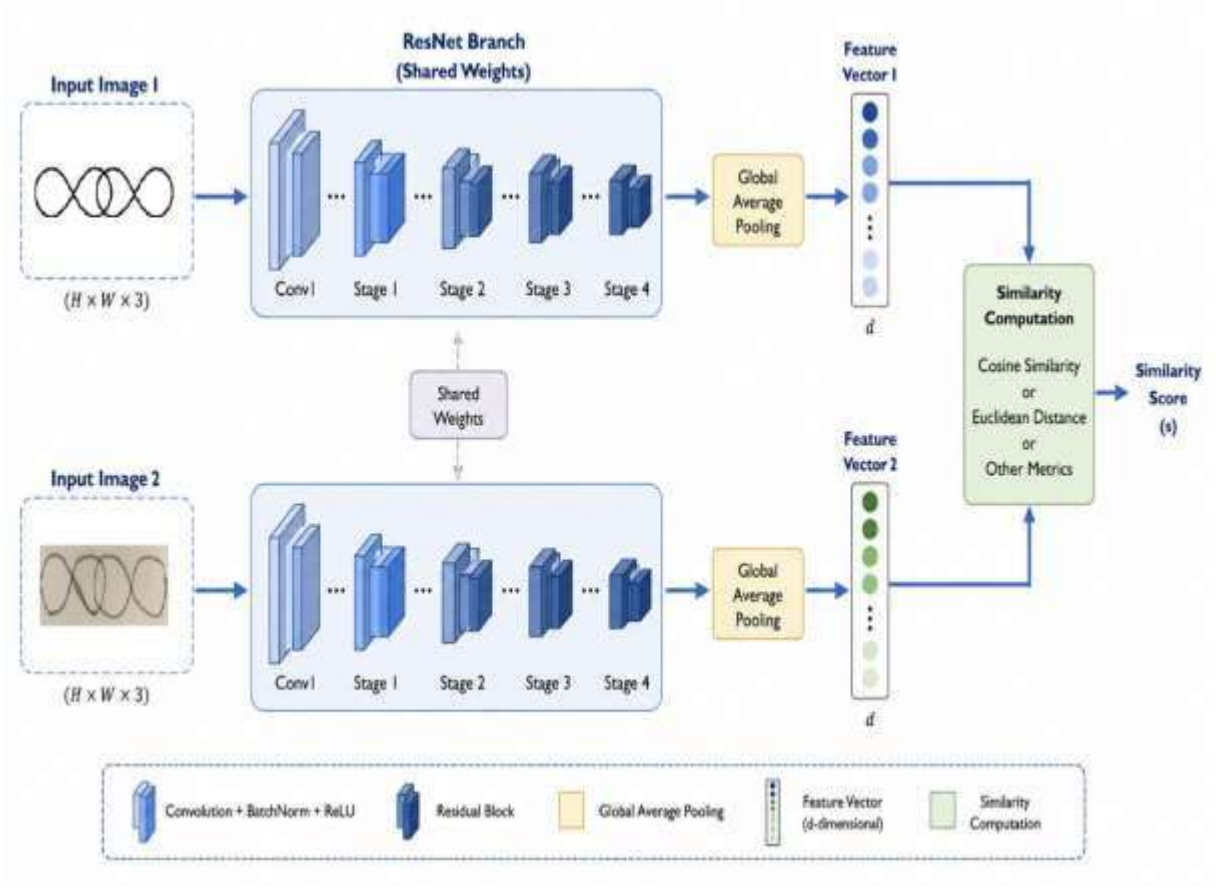


Fig 1. Operational Framework of a ResNet50-based Siamese Network

The shared-weight mechanism in Siamese model ensures that both branches can learn the identical feature representations, thus enabling consistent comparison between image pairs. The similarity score generated by these network indicates that how closely the two images be similar to each other in the learned feature space. Therefore, ResNet50 alone should not be considered as a Siamese network autonomously. Instead, ResNet50 is a CNN architecture which is used for feature extraction, Siamese Network is a framework which is used for similarity-learning between image pairs and these impression represents a Siamese network using ResNet50 as the CNN backbone.

The proposed system relies on a dedicated image dataset containing geometric and abstract figure drawings as shown in fig 2. The dataset consists of two major categories are reference images and user-drawn images. Reference images include professionally constructed drawings of cubes and infinity figures that serve as ground-truth visual templates. User images are collected from participants who reproduce these figures as part of a digital cognitive assessment process. Since user drawings naturally exhibit variations in size, orientation, stroke thickness and drawing precision, a computational mechanism is required to fairly evaluate their quality. This dataset forms the basis for training and validating the similarity-based scoring model implemented in the study.

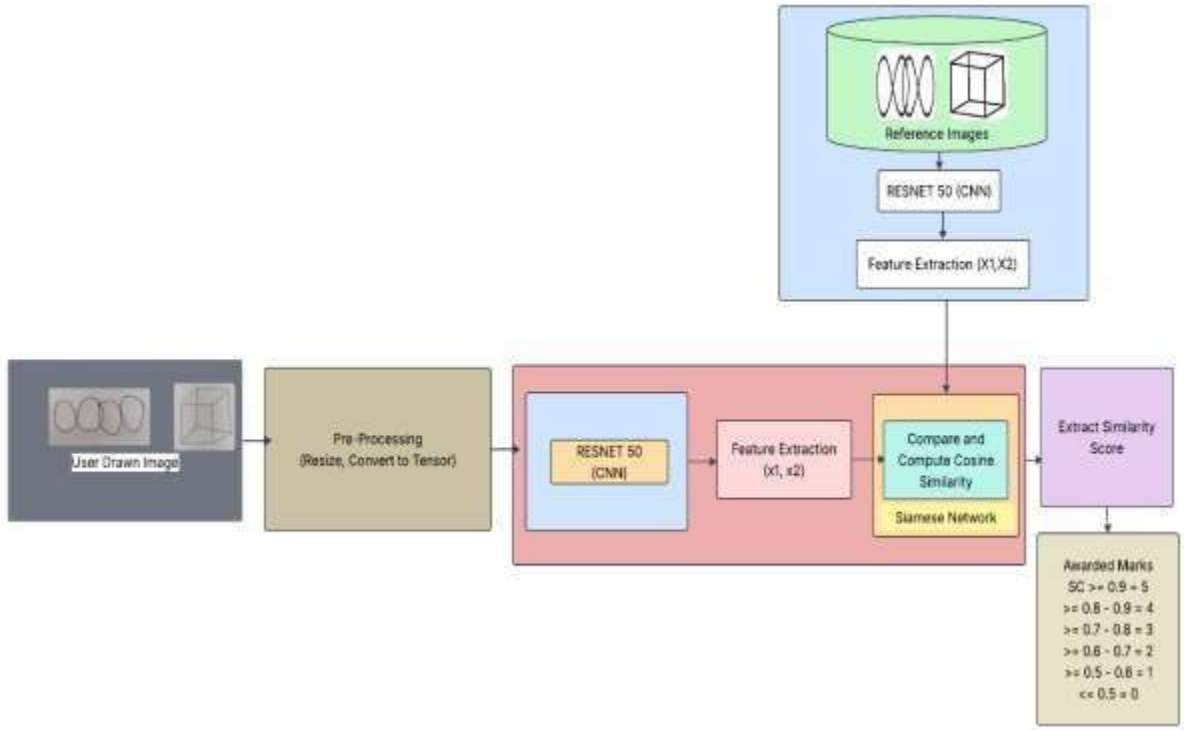


Fig 2. Working Process of Visuospatial Cognitive Assessment using proposed Deep Similarity Learning (DSL) framework

Prior to similarity evaluation, all images undergo a comprehensive preprocessing pipeline. Raw images captured from different users cannot be directly utilized due to inconsistencies in resolution and format. Therefore, uniform preprocessing operations are applied to both reference and user drawings to ensure compatibility with convolutional neural network architectures. Each image is first resized to a fixed dimension, typically 224 × 224 pixels, to satisfy the input requirements of ResNet models. Images are then converted into numerical tensor representations to enable mathematical processing on GPU platforms. Pixel intensities are scaled and normalized using standard mean and deviation values to minimize illumination differences and background noise. When necessary, grayscale drawings are converted into three-channel inputs to maintain uniform structure across the dataset. The output of this stage is a pair of standardized tensors, denoted as X_1 for the reference image and X_2 for the user-drawn image, which are subsequently used as inputs to the Siamese model as shown in fig. 3

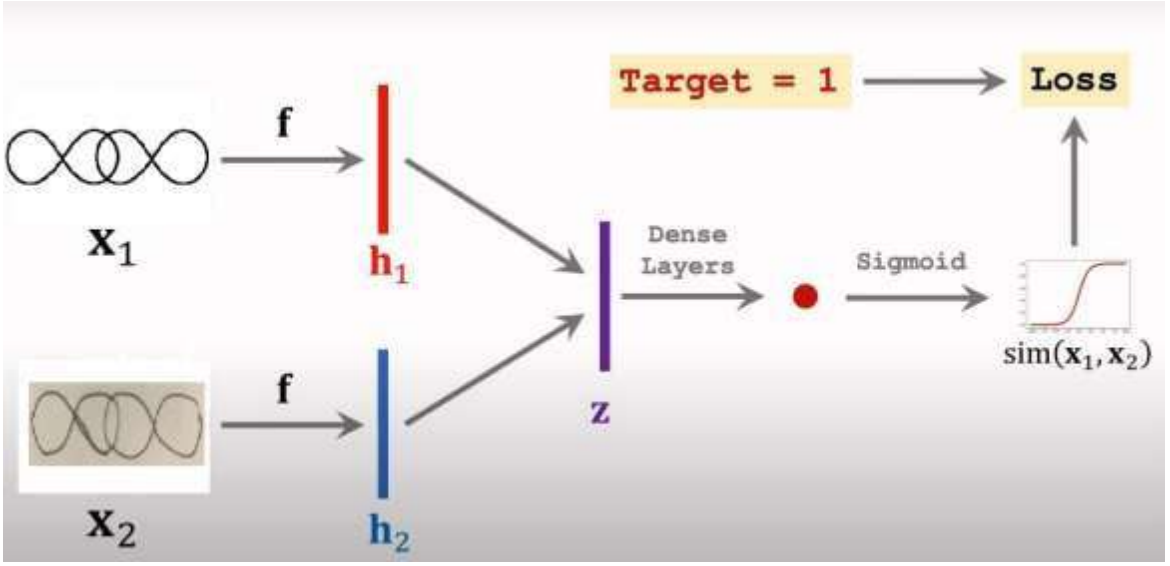


Fig 3. Learn similarity between two input samples

In the proposed framework, the first branch receives the preprocessed reference drawing, whereas the second branch processes the corresponding participant or user-generated drawing. Each branch integrates a ResNet-based convolutional encoder framework that acts as a high-level feature extractor which is used for extracting features from the pairs of images. Unlike traditional pixel-wise assessment techniques, the ResNet model captures discriminative semantic characteristics including geometric structure, spatial organization, shapes, contours, curves, edges and fine-grained visual patterns. The residual learning strategy within ResNet facilitates efficient deep feature learning while mitigating the vanishing gradient problem, thereby improving representation capability for complex visuospatial patterns.

Let the two input, user and reference images can be represented as (X_1) and (X_2). These inputs X1 and X2 are simultaneously propagated through the twin subnetworks that can be represented by the nonlinear mapping function ($f(\cdot)$). Since both subnetworks share identical weights, the extracted embeddings are projected into the same feature domain, enabling meaningful similarity analysis. The embedding generation process can be expressed as follows:

$h_1 = f(X_1)$ (1)

$h_2 = f(X_2)$ (2)

where h_1 and h_2 are the high-dimensional feature embeddings which is parallel to the reference and user drawings respectively. These extracted feature embeddings encode the crucial structural and semantic information of these two images while removing irrelevant variations and noise.

2.1 Similarity Representation

Following feature extraction task, the embedding vectors are combined to construct a relational similarity representation. The similarity modelling stage targets to quantify the degree of correspondence between these two embeddings by learning their structural associations in the latent space. Several transformation strategies can be employed for this purpose, including vector concatenation, element-wise absolute difference, cosine similarity, Euclidean distance and hybrid similarity metrics. The generated similarity representation effectively highlights both the common and discriminative characteristics between the two input drawings.

Afterwards, the combined feature representation goes through fully connected dense layers. These layers help the system to learn the complex nonlinear interactions between the embeddings and encouraging its ability to make the differentiation between these representations. Also, it refining the similarities and improving the robustness of comparative analysis. At last, the network generates a similarity score that reflects the degree of correspondence between the participant drawing and the reference images. Overall, the proposed Siamese-ResNet50 framework enables efficient and accurate similarity assessment by integrating deep residual feature extraction with metric-based relational learning.

The sigmoid activation function is used in the final layer of the proposed system to generate a similarity score between 0 and 1. This similarity score represents the degree of similarity between the two inputs, where the scores close to 1 indicate high similarity and the scores close to 0 indicate dissimilarity as illustrated in fig. The network is trained using the pairs of labeled input, where each pair of labels are associated with a ground-truth label indicating whether the samples are similar or dissimilar. To optimize the model, a binary cross-entropy loss function is used for minimizing the variance between what the model predicts and the true label. Through this supervised learning process, the network learns an embedding space in which similar inputs are positioned closer together, whereas dissimilar inputs are mapped farther apart.

Cosine similarity measures the angular distance between the feature vectors x1 and x2 and produces a continuous value between 0 and 1 to calculate the similarity between the images as illustrated in fig 4. Where vector x1 denotes the reference image and x2 represents the user drawn image, and the similarity scores closer to 1 that indicate the strong structural correspondence with the reference images, whereas the lower values denote the poor similarity. The similarity score which is represent by S is calculated as the dot product of the vectors divided by the product of their magnitudes, guaranteeing that the measure is independent of vector scale. This approach is well-suited for comparing user drawn freehand

drawings where magnitude variations are common but geometric structure is more important for evaluation. The extracted similarity score becomes the quantitative indicator used by the automated marking module of the system.

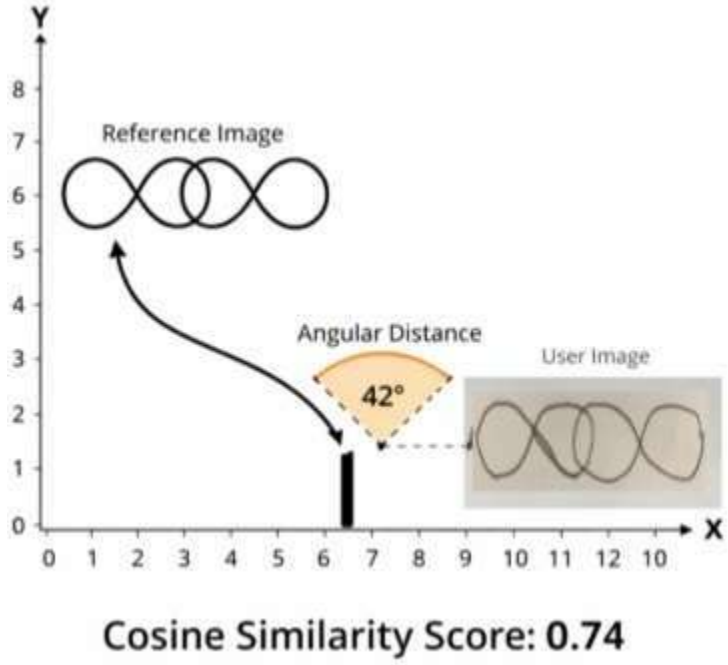


Fig 4. Angular Distance and Cosine Similarity score calculations.

2.2 Scoring Strategy

A transparent threshold-based scoring mechanism converts the continuous score into discrete marks. For instance if the similarity score is greater than or equal to 0.9, the system awards 5 marks, it defines both images are near-perfect drawing. The subsequent lower ranges of scores like ≥ 0.8, ≥ 0.7, ≥ 0.6 and ≥ 0.5 which are assigned the following 4, 3, 2 and 1 marks respectively. This decision structure ensures consistent, transparent and interpretable evaluation for all participants. The mapping logic is very simple yet high effective, enabling instant conversion from deep model output similarity scores to cognitive performance marks.

Finally, our system displays the evaluation result to the user. The awarded marks and similarity score are presented through the graphical interface or console output. The proposed system can also store the results numerically for additional analysis. Showing results in real time enables instant feedback to the individual and supports longitudinal observation of individual performance through multiple assessment. Largely, the system proves an end-to-end automated assessment framework integrating pretrained feature extraction, Siamese-based comparison, cosine similarity computation and objective rule-based mark awarding schemes. Each step in our system contributes to decreasing human-based bias, improving reproducibility and enabling scalable digital visuospatial cognitive assessment.

3 Results and Discussions

During training phase, the Siamese with ResnNet50 network has been optimized using the pairs of images labelled or categorized as either similar or dissimilar. Each pair of input is forwarded simultaneously through the twin branches of the network, which share identical weights to ensure the consistency of feature extraction. The output feature embeddings from both branches are then used to compute the Euclidean distance, which serves as the basis for contrastive loss computation.

The objective of the training phase is to minimize the contrastive loss computation over the entire training process. Backpropagation has been applied to update the parameters of the shared convolutional backbone and enabling the network to progressively learn to extract more discriminative feature representations. The optimization process in the network pushes similar image pairs more closer in the embedding space while lashing dissimilar pairs have farther apart beyond the specified margin.

Stochastic gradient descent or adaptive optimizers called Adam are employed to minimize the computation of loss function. The network is trained for multiple times using multiple epochs until convergence is attained. Through this metric learning strategy, the Siamese network with ResNet50 has the capable of generating robust embeddings that effectively capture visual and spatial similarity between images.

Let our proposed model extract feature embeddings from the reference image and the user drawn image. These extracted feature embeddings are represented as high-dimensional vectors:

$$f_r = [f_{r1}, f_{r2}, \ldots, f_{rd}] \quad (3)$$

$$f_t = [f_{t1}, f_{t2}, \ldots, f_{td}] \quad (4)$$

where
f_r denotes the feature vector of the reference image,
f_t denotes the feature vector of the test or user drawn image, and
d is the dimensionality of the feature space.

3.1 Cosine Similarity

After feature embeddings extraction, similarity has been computed between the two feature vectors using cosine similarity, which is used to measure the angular similarity between the feature embeddings and is defined as:

$$S_{cos} = \frac{[f_r, f_t]}{||f_r||\ ||f_t||} \quad (5)$$

where
f_r and f_t represents the dot product between vectors
(|·|) denotes the Euclidean norm.

The obtained similarity score lies in the range between 0 and 1, where higher similarity scores specify greater visual similarity whereas the lower specify the poorer similarity between the user drawn and reference images. Based on these similarity values, a discrete performance score or marks is awarded using the predefined threshold criteria. This framework enables robust and quantitative evaluation of visual resemblance using deep feature representations and metric learning principles.

In this study, various deep learning architectures were evaluated to determine their effectiveness in feature-based similarity comparison. The selected models include convolutional neural networks such as ResNet, DenseNet, VGG, EfficientNet, AlexNet, ConvNeXt**,** and transformer-based models such as Vision Transformer and Swin Transformer. These architectures are widely adopted in visual feature extraction due to their ability to capture both local and global patterns.

To enhance interpretability and clinical or cognitive relevance, the similarity scores were mapped into a standardized scoring system as defined in table 1. Specifically, similarity values were scaled and converted into marks out of five as shown in the table X. This transformation aligns with conventional scoring approaches used in visuospatial cognitive assessments, where performance is evaluated using graded scoring criteria.

Table 1. Conversion of Similarity Scores into Performance Marks

Models	**Similarity Scores**				**Marks (Out of 5)**			
	Cosine	**Euclidean Distance**	**MSE**	**Hybrid Cosine + Euclidean Similarity**	**Cosine**	**Euclidean Distance marks**	**MSE**	**Hybrid Cosine + Euclidean Similarity marks**
Resnet_18	0.7422	0.4962	0.999	0.5662	3.71	2.48	4.99	2.83
Resnet_50	0.864	0.5756	0.9995	0.6356	4.32	2.88	5	3.18
Resent_101	0.7965	0.5257	0.9992	0.595	3.98	2.63	5	2.97

Efficientnet_B0	0.6827	0.4702	0.9987	0.537	3.41	2.35	4.99	2.68
Efficientnet_B1	0.6752	0.4673	0.9987	0.5333	3.38	2.34	4.99	2.67
Efficientnet_B2	0.7536	0.5018	0.999	0.5721	3.77	2.51	5	2.86
Efficient_v2_L	0.5123	0.4172	0.9981	0.4599	2.56	2.09	4.99	2.3
Densenet_121	0.7445	0.4973	0.999	0.5674	3.72	2.49	4.99	2.84
Densenet_161	0.738	0.4942	0.999	0.5642	3.69	2.47	4.99	2.82
Densenet_169	0.7245	0.4879	0.9989	0.5573	3.62	2.44	4.99	2.79
Densenet_201	0.7682	0.5094	0.9991	0.5798	3.84	2.55	5	2.9
Inception_v3	0.588	0.4379	0.96984	0.4932	2.94	2.19	4.99	2.47
VGG11	0.7563	0.5032	0.999	0.5734	3.78	2.52	5	2.87
VGG16	0.8067	0.5321	0.9992	0.6005	4.03	2.66	5	3
VGG19	0.7643	0.5074	0.9991	0.5775	3.82	2.54	5	2.89
VGG19_bn	0.8196	0.5407	0.9993	0.6079	4.1	2.7	5	3.04
Swin_b	0.7552	0.5026	0.999	0.5737	3.78	2.51	5	2.87
Alexnet	0.8527	0.5657	0.9994	0.6282	4.26	2.83	5	3.14
convnext_base	0.7396	0.4949	0.999	0.5658	3.7	2.47	4.99	2.83
vit_b_16	0.6761	0.4677	0.9983	0.5346	3.38	2.34	4.99	2.67
vit_l_32	0.6868	0.4718	0.9988	0.5388	3.43	2.36	4.99	2.69

The similarity between the reference and user drawn images were quantified using multiple distance and similarity measures. Cosine similarity, Euclidean distance, Mean Squared Error (MSE), and a hybrid similarity metrics have been used for computation process as illustrated in the fig 5. Each metric uses different aspects to calculate similarity. Cosine similarity evaluates the angular relationship between feature vectors, this making it invariant to magnitude variations. Euclidean distance has been used to measure the spatial separation between vectors and reflects absolute feature differences. MSE metric is different from all other metric, which quantifies pixel-level or feature-level reconstruction error, useful in capturing fine-grained deviations. The hybrid similarity score combines complementary properties of Cosine and Euclidean metrics which improves the robustness and also reduce the influence of noise.

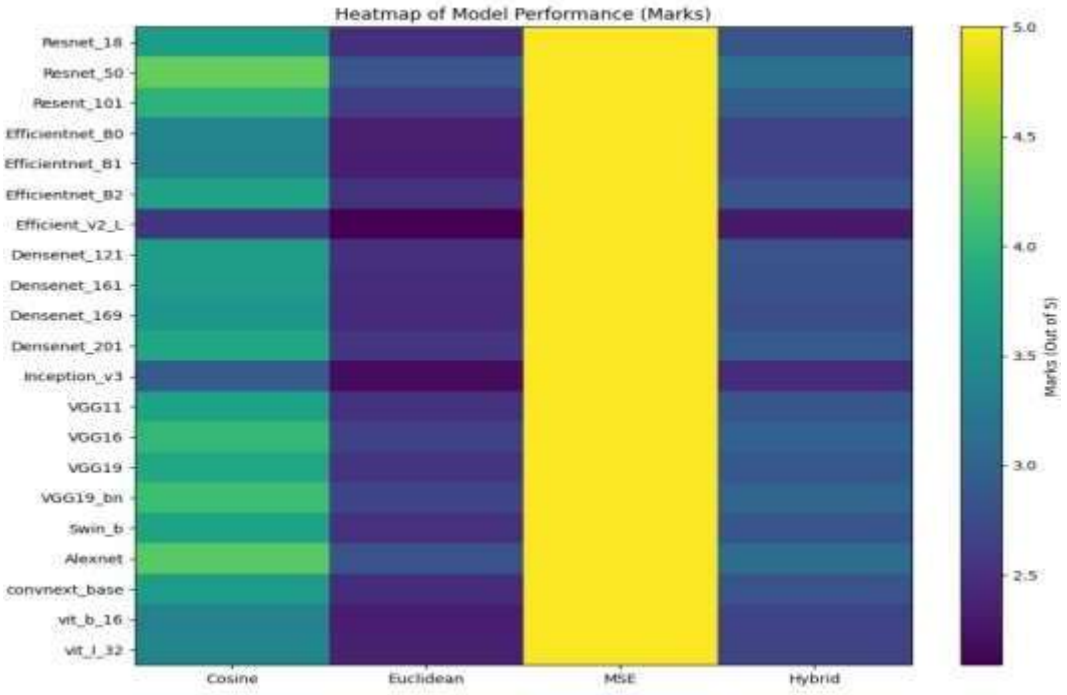


Fig 5. Models performances based on similarity metrics and scores

The conversion procedure guarantees that higher similarity values and its generated scores correspond to better model performance. It also bridges the gap between machine learning outputs and human-interpretable evaluation frameworks. These kinds of scoring systems are predominantly useful in healthcare applications such as automated dementia screening, where clinicians necessitate interpretable and standardized performance indicators.

3.2 Similarity evaluation on other than Siamese models

Table 2 and Figure 6 present the comparative performance analysis of eight deep learning models evaluated using cosine similarity, hybrid similarity, Euclidean similarity, and final scoring metrics. The experimental results demonstrate clear performance variations among the tested architectures, highlighting the effectiveness of recurrent models for the given recognition task.

The experimental results presented in the table demonstrate the comparative performance of multiple deep learning architectures using three similarity measures, namely cosine similarity, hybrid similarity, and Euclidean distance, along with the final evaluation score expressed as marks out of 5. The analysis reveals that high similarity values alone do not necessarily guarantee higher scoring performance, since the final assessment depends on the consistency, structural preservation, semantic understanding, and robustness of the generated or reconstructed outputs.

The obtained results in the above table indicates that recurrent architectures such as LSTM and GRU achieved the highest similarity scores due to their superior capability in preserving sequential dependencies and contextual information. Though their cosine similarity values are comparable to those of GNN values, their hybrid similarity scores are significantly higher, this indicating that enhanced alignment between structural and semantic features. Consequently, these models have provided more accurate and reliable representations, this leading to maximum evaluation scores.

Table 2. Comparative performance of multiple deep learning architectures.

Models	Similarity			Marks (out of 5)
	Cosine Similarity	Final Hybrid similarity	Euclidean	
AutoEncoder	0.8481	0.85	-	4
GAN	0.8462	0.5322	0.0612	1
CapsNetworks	0.8078	0.6325	0.3696	2

GNN	0.9977	0.7828	0.4604	3
LSTM	0.9953	0.8611	0.7413	3
GRU	0.9979	0.8611	0.5742	3
CLIP	0.708	0.6656	0.5668	2
ResNet50	0.864	0.6356	0.5756	4.32

Interestingly, the GNN model achieved an extremely high cosine similarity score is 0.9977, though it achieved only 3 marks. These anomalies bring into perspective an inherent weakness associated with the use of cosine similarity measures as a means of evaluation. The cosine similarity measure is a method employed to gauge directionality between feature vectors, but it fails to consider local distortions and artifacts reconstruction. So that, in spite of exhibiting strong vector-level similarity, the outcomes of the GNN still contain structural deviations that affect the qualitative assessment and final scoring in negatively.

Similarly, the GAN model demonstrated moderate cosine similarity but received the lowest score. This is primarily because GAN-based architectures may generate visually plausible outputs while lacking faithful reconstruction consistency. The significantly lower hybrid similarity score suggests instability in feature preservation, and the very small Euclidean distance alone was insufficient to reflect perceptual quality. This indicates that isolated metric optimization cannot guarantee robust assessment performance.

Capsule Networks and CLIP exhibited moderate performance. Although these models capture spatial and semantic relationships effectively, their relatively lower hybrid similarity values indicate incomplete alignment between extracted features and target representations. Accordingly, their outputs were less reliable in attaining high assessment accuracy.

Among the above evaluated models, our proposed Siamese with ResNet50 demonstrated the most balanced and stable performance across the above all metrics. Though its cosine similarity was not the uppermost, the proposed model consistently maintained strong hybrid similarity and moderate Euclidean distance, it resulting in an overall score of 4.32 out of 5. It's because of the consistent and stable characteristics that are a result of the deep residual learning of ResNet50 architecture. It enables the features to be extracted without being exposed to the challenge of vanishing gradients. The use of residual connections guarantees both low and high level visual representations.

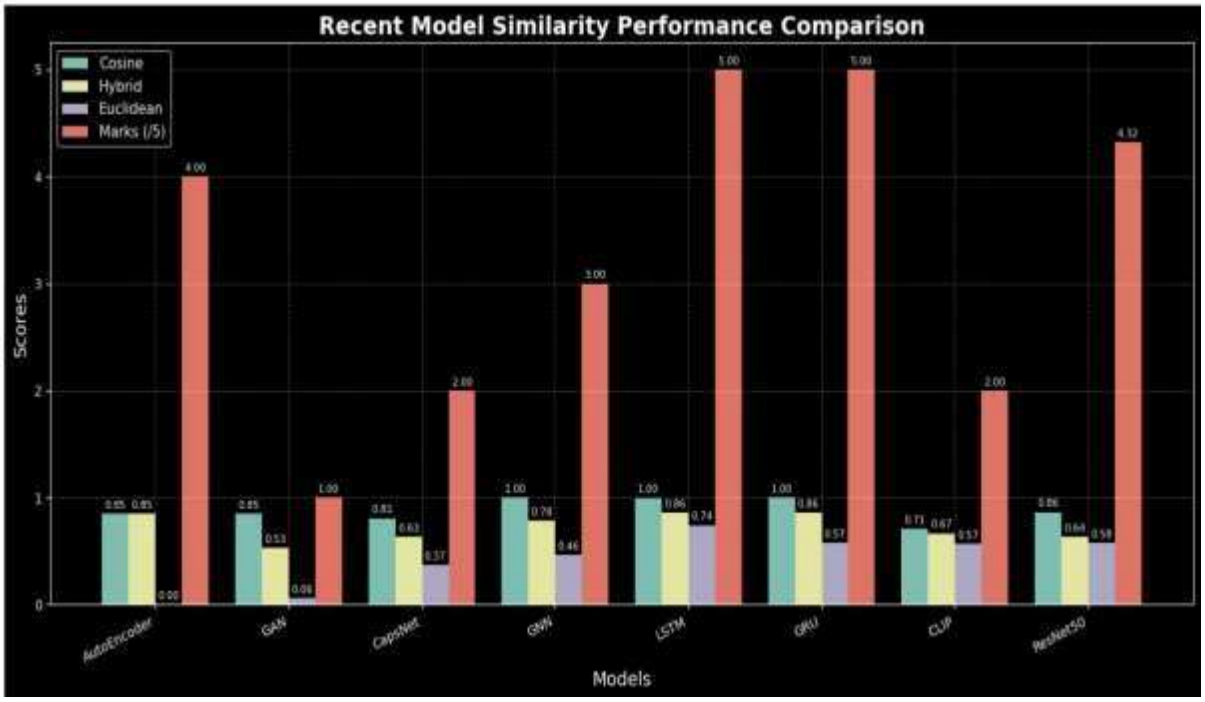


Fig 6. Deep Learning Models performance comparison

The above experimental findings suggest that the model evaluation should not depend exclusively on a single similarity metric. High cosine similarity metric may indicate that the mathematical closeness between embeddings, but it does not certainly correspond to perceptual correctness or structural fidelity. The combination of the feature similarities can be a better evaluation method because it would incorporate several aspects of features in order to obtain the correlation with the quality of the output and the human understanding.
All things considered, the Siamese with ResNet50 model is proven to be the most stable and reliable structure among others based on their performances in various types of evaluations. While some other traditional models have extremely high similarity values, but at the same time unstable or inconsistent structure, the ResNet50 model is the one that provides a perfect balance between all required characteristics.

4. Conclusion

The experimental results confirm that similarity-based cognitive evaluation and scoring, combined with advanced deep learning architectures, provides a robust and objective framework for automated visuospatial cognitive screening. The ability to quantitatively compare multiple Deep learning models ResNet, efficientnet, densenet, inceptionnet, VGG, Swin transformer, convnext and ViT and other than Siamese models Autoencoder, GAN, LSTM, GRU and CLIP, also enables the selection of optimal architectures for real-world deployment. The LSTM, GRU and GNN model achieved extremely high cosine similarity 0.9977, yet obtained the lowest marks. It may generate visually plausible outputs while lacking faithful reconstruction consistency, instability in feature preservation and being insufficient to reflect perceptual quality. This indicates that isolated metric optimization cannot guarantee robust assessment performance. The proposed model Siamese backbone ResNet50 outperforms all other models and provides stable performance with 0.864 similarity and 4 marks out of 5. Our proposed deep learning model enables efficient feature extraction while avoiding vanishing gradient issues. The residual connections help preserve both low-level and high-level visual representations, leading to robust generalization, reliable similarity estimation and scoring reliability. These prominent characteristics are mainly suitable in large-scale cognitive screening, telemedicine applications and remote dementia monitoring. Overall, the proposed methodology supports early detection of visuospatial deficits in addition to provides a scalable, automated cognitive assessment alternative to traditional methods.
In future, the research will primarily focus on enhancing recurrent architecture-based visuospatial assessment frameworks, with particular prominence on improving the mapping between similarity scores and clinical scoring mechanisms. The proposed direction also aims to integrate multiple cognitive assessment domains, including language, memory, recall, and visuospatial abilities, along with response time and aggregated scoring measures, to enable a more comprehensive evaluation of an individual's cognitive status. By combining multimodal cognitive indicators into a unified assessment framework, the system can support more accurate identification and differentiation of cognitively normal individuals and those with cognitive impairment.

REFERENCES:

1. Perez, E., & Weisberg, S. M. (2025). Individual differences in spatial navigation. Reference Module in Neuroscience and Biobehavioral Psychology (Elsevier, 2023). doi, https://doi.org/10.31219/osf.io/9x6a3.
2. Pena MCS, Sobreira EST, Souza CP, Oliveira GN, Tumas V, do Vale FAC. Visuospatial cognitive tests for the evaluation of patients with Parkinson's disease. Dement Neuropsychol. 2008 Jul-Sep; 2(3):201-205. doi: 10.1590/S1980-57642009DN20300007. PMID: 29213571; PMCID: PMC5619466.
3. Berente Dalida Borbala , Kamondi Anita , Horvath Andras Attila, "The Assessment of Visuospatial Skills and Verbal Fluency in the Diagnosis of Alzheimer's Disease", Frontiers in Aging Neuroscience, Volume 13, 2022, DOI=10.3389/fnagi.2021.737104, ISSN 1663-4365.
4. Huang Yan , Zhang Xuehao , Tang Jia , Xia Yuqi , Yang Xiaotong , Zhang Yanmei , Wei Chaogang , Ruan Ruiqi , Ying Hang , Liu Yuhe, "Vestibular cognition assessment system: Tablet-

based computerized visuospatial abilities test battery", Frontiers in Psychology, Volume 14, 2023, DOI=10.3389/fpsyg.2023.1095777, 1664-1078.
5. Rai Pooja, Singh Indramani L., Singh Tara, Tiwari Trayambak, Joshi Deepika, "Comparing visuospatial construction ability of MoCA and MMSE in cognitive assessment of patients with dementia of Alzheimer's type", Alzheimer's & Dementia, 19, 1552-5260, 2023, https://doi.org/10.1002/alz.065307.
6. Cundari M, Vestberg S, Hansson A, Kennberg J, Gustafsson P, Rasmussen A. Sensorimotor functions, visuospatial perception and visuospatial abilities in adult attention deficit hyperactivity disorder and autism spectrum disorder. Journal of the International Neuropsychological Society. 2025;31(3):254-266. doi:10.1017/S1355617725000189
7. Vindhya Vishwanath, Palash K. Malo, Ajith Partha, Divya N M, Abhishek Mensegere Lingegodwa, Albert Stezin, Amitha C M, Rajitha Narayanasamy, Meenakshi Menon, Meghana R, Goutham Velavarajan, Deva Kumar HS, Prathima Arvind, Monisha S, Shafeeq K Shahul Hameed, Sunitha HS, Deepashri Agrawal, Banashree Mondal, Sadhana Singh, Jonas S. Sundarakumar, Thomas Gregor Issac, "Demographic Correlates of Visuo-Cognition in a Healthy Ageing Indian Population - A Cross-Sectional Analysis", Alzheimer's & Dementia, 20, 1552-5260, 2024, https://doi.org/10.1002/alz.091239.
8. Mirchi, Z., Kheirkhah, M., Khosrowabadi, R., Fadardi, J., & Ramezani, M. (2024). Development of a real-world simulated instrument for evaluating visuospatial working memory: a preliminary psychometric study on older adults. BMC Geriatrics, 24. https://doi.org/10.1186/s12877-024-05140-9.
9. Aul, C., Brau, J., Sugarman, A., DeGutis, J., Germine, L., Esterman, M., McGlinchey, R., & Fortenbaugh, F. (2023). The functional relevance of visuospatial processing speed across the lifespan. Cognitive Research: Principles and Implications, 8. https://doi.org/10.1186/s41235-023-00504-y.
10. Rai, P., Singh, I., Singh, T., Tiwari, T., & Joshi, D. (2023). Comparing visuospatial construction ability of MoCA and MMSE in cognitive assessment of patients with dementia of Alzheimer's type. Alzheimer's & Dementia, 19. https://doi.org/10.1002/alz.065307.
11. Mammarella, I., Cardillo, R., & Orefice, C. (2023). Trajectory assessments of cognitive, visuospatial and academic profile in nonverbal learning disability (visuospatial developmental disorder). Research in developmental disabilities, 139, 104540. https://doi.org/10.1016/j.ridd.2023.104540.
12. Shao, X., Wang, K., Zhang, Y., Zhen, X., Dong, F., Tian, H., & Yu, Y. (2023). Outcome of visuospatial dysfunction assessment in patients with Parkinson's disease using mobile application software. Frontiers in Aging Neuroscience, 15. https://doi.org/10.3389/fnagi.2023.1108166.
13. Zhang, X., Lv, L., Min, G., Wang, Q., Zhao, Y., & Li, Y. (2021). Overview of the Complex Figure Test and Its Clinical Application in Neuropsychiatric Disorders, Including Copying and Recall. Frontiers in Neurology, 12. https://doi.org/10.3389/fneur.2021.680474.
14. Latino, F., Cataldi, S., & Fischetti, F. (2021). Effects of a Coordinative Ability Training Program on Adolescents' Cognitive Functioning. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.620440.
15. Haque, R., Manzanares, C., Brown, L., Pongos, A., Lah, J., Clifford, G., & Levey, A. (2019). VisMET: a passive, efficient and sensitive assessment of visuospatial memory in healthy aging, mild cognitive impairment and Alzheimer's disease. Learning & Memory, 26, 93 - 100. https://doi.org/10.1101/lm.048124.118.
16. Massaldjieva, R. (2018). Differentiating Normal Cognitive Aging from Cognitive Impairment No Dementia: A Focus on Constructive and Visuospatial Abilities. Gerontology. https://doi.org/10.5772/intechopen.73385.
17. Burggraaf, R., Frens, M., Hooge, I., & Van Der Geest, J. (2018). Performance on tasks of visuospatial memory and ability: A cross-sectional study in 330 adolescents aged 11 to 20. Applied Neuropsychology: Child, 7, 129 - 142. https://doi.org/10.1080/21622965.2016.1268960.
18. Appalaraju, S., & Chaoji, V. (2017). Image similarity using Deep CNN and Curriculum Learning. ArXiv, abs/1709.08761.

19. Potts, C., Richardson, J., Bond, R., Price, R., Mulvenna, M., Zvolsky, P., Harvey, M., Hughes, C., & Duffy, F. (2021). Reliability of Addenbrooke's Cognitive Examination III in differentiating between dementia, mild cognitive impairment and older adults who have not reported cognitive problems. European Journal of Ageing, 19, 495 - 507. https://doi.org/10.1007/s10433-021-00652-4. https://library.neura.edu.au/schizophrenia/signs-and-symptoms/cognition/visuospatial-ability/index.html
20. Geno Peter, J. Livin, and A. Sherine, "Hybrid optimization algorithm based optimal resource allocation for cooperative cognitive radio network," Array, vol. 12, p. 100093, Dec. 2021, doi: 10.1016/j.array.2021.100093.
21. Chen, X., & He, K. (2020). Exploring Simple Siamese Representation Learning. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15745-15753. https://doi.org/10.1109/cvpr46437.2021.01549.