Neural Computing and Applications (2020) 32:12909–12933
https://doi.org/10.1007/s00521-020-04735-8

ORIGINAL ARTICLE

Deformation modeling and classification using deep convolutional neural networks for computerized analysis of neuropsychological drawings

Momina Moetesum[1] · Imran Siddiqi[1] · Shoaib Ehsan[2] · Nicole Vincent[3]

Received: 24 June 2019 / Accepted: 10 January 2020 / Published online: 20 January 2020

Abstract
Drawing-based tests are cost-effective, noninvasive screening methods, popularly employed by psychologists for the early detection and diagnosis of various neuropsychological disorders. Computerized analysis of such drawings is a complex task due to the high degree of deformations present in the responses and reliance on extensive clinical manifestations for their inferences. Traditional rule-based approaches employed in visual analysis-based systems prove insufficient to model all possible clinical deformations. Meanwhile, procedural analysis-based techniques may contradict with the standard test conduction and evaluation protocols. Leveraging on the increasing popularity of convolutional neural networks (CNNs), we propose an effective technique for modeling and classifying dysfunction indicating deformations in drawings without modifying clinical standards. Contrary to conventional sketch recognition applications where CNNs are trained to diminish intra-shape class variations, we employ deformation-specific augmentation to enhance the presence of specific deviations that are defined by clinical practitioners. The performance of our proposed technique is evaluated using Lacks' scoring of the Bender-Gestalt test, as a case study. The results of our experimentation substantiate that our proposed approach can represent domain knowledge sufficiently without extensive heuristics and can effectively identify drawing-based biomarkers for various neuropsychological disorders.

Keywords Neuropsychological drawings · Deformation classification · Deep visual features · Bender-Gestalt test

1 Introduction

Drawings are one of the earliest known modes of human communication. To date, drawings are a preferred form of expression in a variety of situations like procedural flowcharts, engineering or architectural plans, electronic circuit diagrams and freehand sketches. In clinical psychology, drawings give deep insight into the mental state and behavior of the individuals who made them. This information is employed in several neuropsychological screening tests to detect the early signs of various neurological disorders both in children and adults. Some of the popular tests include the clock draw test (CDT) [1], Rey–Osterrieth complex figure (ROCF) drawing test [2] and Bender-Gestalt test (BGT) [3], as shown in Fig. 1.

Conventionally, these tests are interpreted by trained practitioners who visually assess the responses produced by different subjects. A common practice is to measure the extent of deviation(s) from the standard template(s) introduced by subjects while drawing. These deviations are scored based on standard rules, and the presence or absence of a particular disorder is then determined. For instance, 'Micrographia' and 'Tremor' are two common Parkinsonian conditions that cause irregularity and tightness while

✉ Imran Siddiqi
imran.siddiqi@bahria.edu.pk

Momina Moetesum
momina.buic@bahria.edu.pk

Shoaib Ehsan
sehsan@essex.ac.uk

Nicole Vincent
nicole.vincent@mi.parisdescartes.fr

[1] Bahria University, Islamabad, Pakistan

[2] University of Essex, Essex, UK

[3] Paris Descartes University, Paris, France

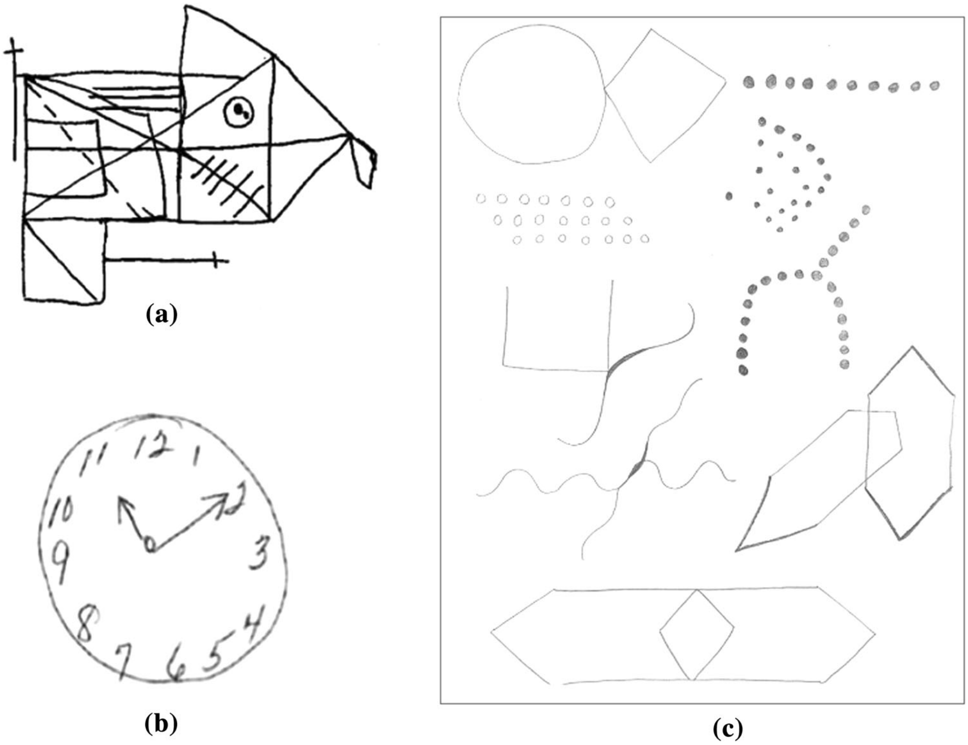


Fig. 1 Samples of drawing task-based neuropsychological tests: **a** ROCF drawing sample [4], **b** CDT drawing sample [5] and **c** BGT drawing sample [6]

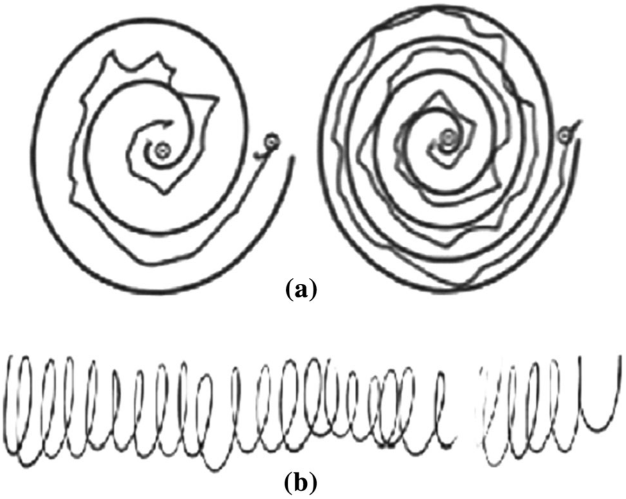


Fig. 2 Irregularity and tightness of loops and spirals as biomarker for motor dysfunction [7]

drawing spirals and loops as shown in Fig. 2a, b, respectively.

Due to extensive scoring standards, manual scoring and analysis of neuropsychological drawings are tedious and time-consuming tasks, even for a trained practitioner. Furthermore, a scorer's experience and bias, and a patient's profile can introduce undesirable factors like inter-scorer variability. Computerized analysis of such drawings can facilitate practitioners in diagnostic decision-making, test validation and standardization. However, designing a system that can identify specific deformations in a hand-drawn shape is a challenging task.

Despite various challenges, the research in computerized analysis of neuropsychological drawings is gaining rapid popularity. The work done in this area can broadly be categorized as 'Visual analysis-based techniques' and 'Procedural analysis-based techniques.' The visual analysis-based techniques mostly employ shape-specific static features extracted from the offline samples of drawings. On the contrary, the procedural analysis-based techniques rely on dynamic features (like kinematics, pen pressure and stroke sequences) captured during the online drawing responses. Both strategies tend to determine the most discriminating features by either employing heuristics or supervised machine learning algorithms. Nonetheless, effective measurement of clinical deviations is highly dependent on the translation of domain knowledge into computable feature space.

Due to the highly unconstrained nature of the hand-drawn responses of these tests, the rule-based approaches often prove insufficient to represent all possible scenarios [6]. On the contrary, the supervised machine learning-based techniques require rich information to discriminate between the expected and the deformed drawing samples. This may require modifications in the standard templates and test conduction protocols [8]. For this reason, researchers are motivated to explore various interdisciplinary solutions, to bridge the gap between modern research and conventional practices [9].

Recently, deep learning-based methods [10] have gained immense popularity in the domain of feature representation and classification. Convolutional neural networks (CNNs) [11–13], a branch of deep learning, have urged researchers to revisit many popular computer vision and pattern recognition problems including 'Sketch recognition' [14]. Like the neuropsychological drawings, sketches are highly deformed variations in a particular shape class, as shown in Fig. 3. A significant advantage of employing CNNs for sketch recognition is their ability to generalize a wide variety of deformations as one shape class. CNNs achieve this by minimizing the intra-class variations and by enhancing the inter-class variations.

Nevertheless, despite their success in the conventional sketch recognition, the applicability of CNNs for analysis of neuropsychological drawings has been limited due to several challenges. A key challenge in neuropsychological drawing analysis is that particular intra-shape class variations (clinical deformations) need to be enhanced instead of being minimized, as shown in Fig. 4. However, the variations within each deformation class must again be generalized. This means that CNNs must be trained to be deformation-specific instead of being shape-specific. To the best of our knowledge, this has not been attempted previously.

In this study, we present a method to employ CNNs for the computerized analysis of neuropsychological drawings. Several clinical deformations across multiple shape templates are generalized and modeled by training a number of deformation-specific CNNs. Each deformation-specific CNN will extract features from a drawn shape response. The extracted features will then be fed to a classifier to determine its presence or absence. Once all deformations present in a particular drawing sample are determined, a quantified score will be generated for the psychologist to assess the condition of the subject. By applying this approach, we intend to address the issues of heuristic insufficiency and feature validity. Furthermore, by employing the offline samples already collected by the domain experts, we intend to avoid modifications in the standard test conduction protocols as well. Practitioners can use the results of our system to validate their findings and decide the next step of the treatment.

The key objective of this study is to identify the key challenges that characterize the computerized analysis of neuropsychological drawings and then to address those. A number of significant contributions have been made in this regard. More specifically, this study intends to:

1. Investigate the potential of CNNs to model clinical deformations as a replacement to an extensive rule-based approach. To achieve this, we aim to generalize a wide range of intra-deformation class deviations by employing deformation-specific augmentations instead of shape-specific ones. Consequently, our proposed technique should be able to:
   - Classify the common deformations across multiple shape classes. (To the best of our knowledge, this has not been attempted before.)
   - Classify multiple deformations independently within the same shape class. This will enable a practitioner to validate his/her findings not only at

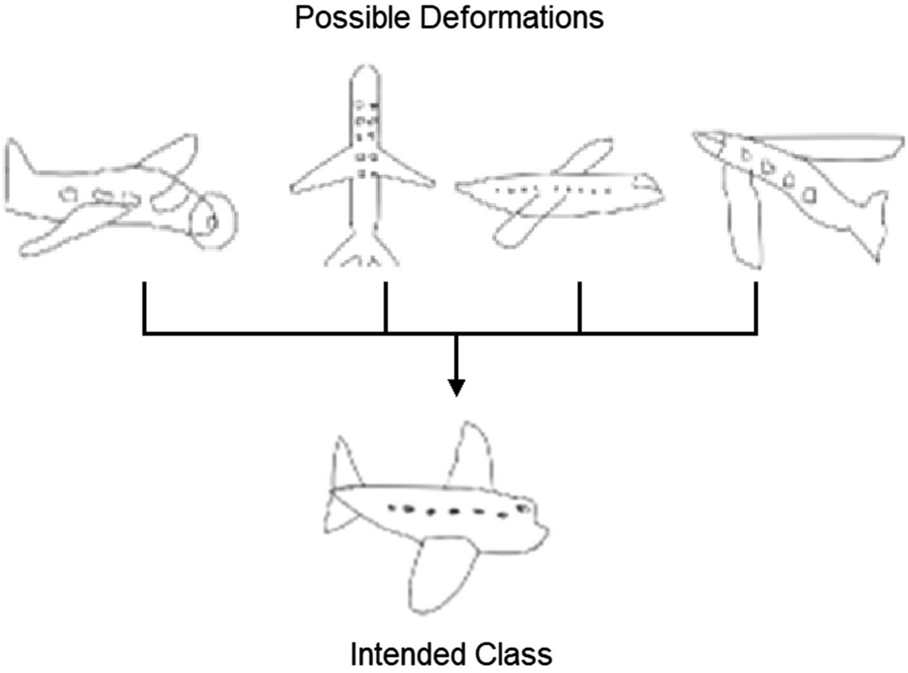


Fig. 3 Example of deformation classification in conventional sketch recognition systems [14]

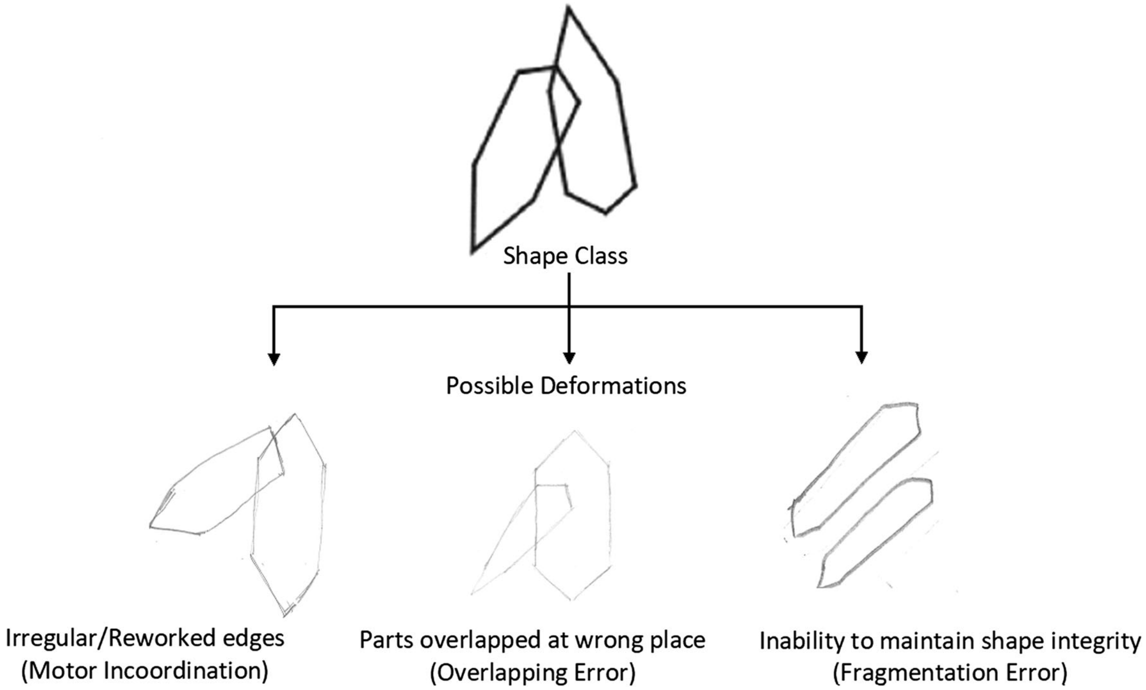


Fig. 4 Example of deformation classification intended for our proposed system (three deformations scored during manual assessment of a standard BGT test)

the image level but the feature level as well. This will mitigate the issue of feature validation.

2. Provide a comprehensive empirical study of the impact of cross-domain transfer learning in the area of drawing analysis. It is a vital exploratory contribution for future studies as the lack of sufficient training data has also been a hindering factor in the applicability of CNNs in this domain. The results of our experimentation can be employed as a baseline by future researchers interested in this area.
3. Propose a taxonomy of techniques employed in the domain of computerized analysis of neuropsychological drawings. It is an important contribution from a research perspective and can facilitate novice researchers in this area.

The organization of the rest of the paper is as follows. Section 2 presents a taxonomy of the work done in the computerized analysis of neuropsychological drawings and highlights some open issues. Section 3 describes the design and the methodology of our proposed solution. The experimental protocol designed to evaluate the performance of our technique is discussed in Sect. 4. The empirical analysis of our findings is reported in Sect. 5. Section 6 concludes the paper and discusses some of the future research directions.

2 State of the art

Over the past few decades, the computerized analysis of handwriting has been performed for a variety of tasks like writer identification, writer verification, handwriting recognition and handwriting interpretation [15]. The recent literature [8, 9, 16–19] suggests a paradigm shift toward the use of computerized handwriting analysis as a diagnostic tool for screening of various neuropsychological and neurological disorders like learning disabilities in children or degenerative diseases (i.e., Parkinson's, Alzheimer's and Dementia) in adults. Several works have been presented in this direction, which we have generalized into two broad categories, i.e., *Visual Analysis-Based Techniques* and *Procedural Analysis-Based Techniques*.

2.1 Visual analysis-based techniques

The visual analysis-based techniques primarily rely on various attributes of the final drawn outcome to distinguish between a healthy and a patient response. Such techniques attempt to analyze the quality of the final drawing by estimating the extent of deviation from the expected template. Figure 5 shows two responses of templates drawn by a healthy subject and a patient suffering from 'Visuospatial Neglect (VSN).' As expected, the samples drawn by the patient are severely deformed as compared to the ones drawn by the healthy subject. Although an easy task for a

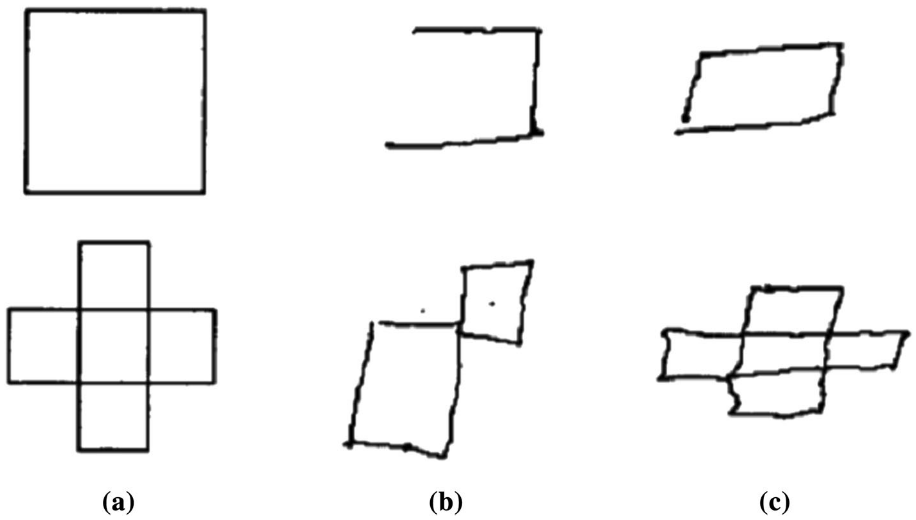


Fig. 5 **a** Intended templates, **b** response samples drawn by a patient suffering from VSN and **c** response samples drawn by a healthy subject [20]

human expert, it is a significantly difficult one for the machine to determine the degree of deformation based on visual analysis. Two common approaches have been employed in the literature for this purpose, (a) to assess the deviations at the local level (i.e., stroke or primitive component level) and (b) to assess the drawing as a whole by extracting shape-based global features. Nonetheless, a common attribute in both approaches is the use of extensive heuristics for the classification purpose. A summary of some of the prominent studies employing visual analysis-based approaches is presented in Table 1.

2.1.1 Component-level drawing analysis

A vital preprocessing step in a component-level drawing analysis is to localize or segment the constituent parts of a drawing. The basic idea is to assess the quality of the whole figure by estimating the deformations in its constituent parts independently. A similar approach is employed in [21], where spatial features are extracted from the primitive components of various geometric shapes and are then compared against the corresponding stimuli to assess the quality of the entire drawing. Similar techniques have also been proposed in [22, 23] to score the responses of the ROCF drawing test (Fig. 1a). According to the defined clinical standards, there are eighteen scoring sections in a ROCF drawing. This requires the system to first localize these sections in the drawing; however, due to the unconstrained nature, these drawings, localization and segmentation of individual scoring sections become a highly challenging task. Authors employ a fuzzy logic-based heuristic approach to localize a small subset of ROCF

Table 1 Summary of studies employing visual analysis-based techniques

References	Task	Samples	Analysis	Conclusion
Canham et al. [23]	ROCF	37 Offline	Localization and assessment of scoring regions using fuzzy logic-based heuristics	78% accuracy on three out of eighteen scoring sections is reported
Fairhurst et al. [20]	Two geometric templates for assessing VSN	26 Online	A set of task-specific geometric features are extracted for analysis	Signifies online sample acquisition and analysis; however, no quantitative score is presented
Bennasar et al. [24]	CDT	648 Offline	Geometric features for digit quality and spatial features for organization are classified using cascaded classifiers	77.78% and 74.38% accuracies for stage three and stage four dementia diagnosis is achieved
Moetesum et al. [6]	Nine BGT drawings	18 Offline	Shape-specific geometric features are extracted and assessed using a heuristic-based approach	Six out of eleven clinical deformations are classified across a small subset of shapes with accuracy ranging from 63.8 to 94.2%
Pereira et al. [28]	Archimedean spiral and meander	55 Offline	Mean relative tremor (MRT) is computed and is used to train a number of classifiers (NB, OPF and SVM)	78.9%, 77.1% and 75.8% accuracies are reported for NB, OPF and SVM, respectively
Harbi et al. [26]	CDT	165 Online	CNN-based features for digit recognition and ontology-based heuristics for classification are used	99% and 95.7% accuracies are reported while classifying drawings of controls and dementia patients, respectively

scoring sections. Once localized, spatial features are extracted from the regions which are then assessed to classify a subset of clinician deformations present in the drawing. Although the complete scoring of the ROCF drawing is not achieved due to localization difficulties, nevertheless, the authors report an accuracy of 78% on a dataset of 31 drawing samples of children attending a special school.

2.1.2 Complete drawing analysis

Recently, there have been attempts to analyze and interpret the samples of CDT drawings (Fig. 1b). CDT is a popular screening test employed for the indication of cognitive disorders like dementia [5]. Similar to ROCF, the CDT drawings have a complex scoring criterion that requires the deduction of inferences not only from the presence/absence of essential shape components but also from their organization (i.e., correct placement of the digits and clock hands). As a result, an independent assessment of the individual components cannot aid in the complete interpretation of the CDT drawing. Authors in [24] present a scheme to classify the offline CDT drawing samples of 648 subjects (healthy and dementia patients) by computing the spatial variations and placement order of the clock components with respect to each other and the clock circumference. A set of spatial and geometric features is extracted for this purpose. Feature selection is then applied to extract the most effective features, which are then fed to a network of cascaded classifiers. The study suggests that the selected features not only prove effective in discriminating between the healthy and patient samples but also in classifying stages of the disease. Classification rates of 77.78% and 74.38% are reported for stage three and stage four dementia diagnosis, respectively. In another attempt [25, 26], authors suggest the use of online sample acquisition methods to facilitate the segmentation and analysis of individual drawing components. An extensive ontology-based heuristic technique is then employed for each component's quality assessment and to determine the correlation between different components.

The 'Archimedean Spiral' drawing test [27] (Fig. 2a) is another popular tool among neurologists for the detection of various Parkinsonian symptoms like tremor. In [28], authors measure the mean relative tremor from the offline spiral samples of 55 subjects to identify the patients suffering from Parkinson's disease (PD). The extracted features are fed to three different classifiers including Naive Bayes (NB), optimum-path forest (OPF) and support vector machines (SVM) to achieve classification rates of 78.9%, 77.1% and 75.8%, respectively.

In [6], a pilot study is conducted to analyze the nine shapes of BGT test for visuo-perceptual dysfunctions indicated by the Lack's scoring system [29]. By extracting a set of geometric features from each of the nine BGT shape classes, a small subset of clinical deformations is classified using a heuristic-based approach. Although the promising results are achieved on a limited sample set, nonetheless, the study concludes that a heuristic-based approach is insufficient to model all possible deformations that are commonly observed by clinical practitioners while analyzing the BGT drawings.

2.2 Procedural analysis-based techniques

With the advent of technology (i.e., digitizer tablets, electronic pens and wearable sensors), researchers find it easier to incorporate the assessment of motor and cognitive skills of a subject, in addition to perceptive ones, while analyzing their drawings. This led to the idea of procedural analysis-based systems, where more focus is given to the way a subject draws, rather than the final outcome itself. Although effective in discriminating the samples drawn by a healthy subject and a patient, yet such systems [8, 17, 30], suggest modifications in the standard test protocols conventionally practiced by the domain experts. This causes issues like feature validation and test acceptability among clinical practitioners [31]. Nonetheless, it is important to discuss such systems for the completeness of the literature review. Table 2 summarizes the literature survey of studies employing various procedural analysis-based techniques.

2.2.1 Drawing movement analysis

Kinematic features are one of the most popularly employed attributes of drawing and handwriting for discriminating between the samples of healthy subjects and patients of various cognitive diseases. In a series of related studies [32, 33], the authors employ several kinematic features to identify drawings of patients suffering from VSN and 'Dyspraxia.' Simple geometric shapes, like the Necker's cube [34], are used as templates. For differential analysis, linear regression and artificial neural networks (ANNs) are employed. In a similar attempt [35], the acceleration sequences captured by using a digitizer tablet while performing the 'Cube Drawing Test' (Fig. 6) are fed to multiple classifiers. A graph-based genetic programming technique is employed to find the best-evolved classifier. A multi-class 'Area Under the Curve (AUC)' score of 0.70 on both the training and the test data is achieved. A total of 120 drawings from 40 subjects (multiple drawings from each subject) are used. In another study, Heinik et al. in [36] extract a combination of kinematic, pressure and spatiotemporal features from the online CDT samples of 20 healthy individuals and 20 subjects suffering from 'Major

Table 2 Summary of studies employing procedural analysis-based techniques

References	Task	Samples	Analysis	Conclusion
Chindaro et al. [50]	Square and cross templates	120 Online	A set of static and dynamic features are extracted to model stroke sequences and analyzed using HMMs	55.9% and 69.5% classification accuracies are achieved for square and cross, respectively
Heinik et al. [36]	CDT	40 Online	Kinematic, pressure and spatiotemporal features are extracted and analyzed	81.1% accuracy is achieved while classifying samples of MDD patients
Tabatabaey et al. [57]	Polygon shapes	178 Online	Drawing sequences are classified using SVM to find correlation between preferred drawing strategy and expected outcome	63.48% accuracy is achieved in classifying correlation between drawing strategy and drawn response
Periera et al. [46]	Archimedean spiral and meander	308 Online	Pen-based pressure signals are converted into images and are classified using CNNs	87.14% accuracy is achieved in classifying samples of healthy subjects and PD patients
Mucha et al. [48]	Archimedean spiral	35 Online	Fractional derivatives of kinematic features are used to train RF and SVM classifiers	72.38% accuracy is achieved in classifying samples of healthy subjects and PD patients

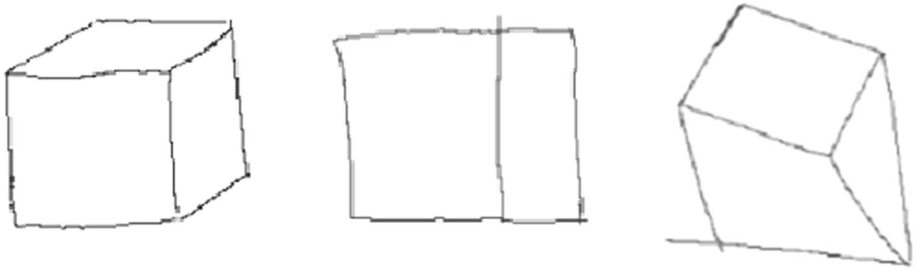


Fig. 6 Samples of Cube Drawing Test drawn by subjects described in [35]

Depressive Disorder (MDD)' and achieve a classification accuracy of 81.1%.

Kinematic and pressure features have also been explored for the identification of early signs of Alzheimer's disease in a number of studies like [37–40]. However, in most of these studies, handwritten words and sentences are considered to be more effective templates for the relevant kinematic feature acquisition. Online handwriting models have also been used to assess various learning disabilities in children and degenerative disorders in the elderly [16, 41, 42]. It is particularly interesting to observe that the selection of a template plays a vital role in capturing the most effective dynamic features for the diagnosis of a specific disorder. This is evident from studies like [43–48] that attempt to highlight the discriminating abilities of online features of handwriting in the prediction and differential diagnosis of PD. According to these studies, the existing templates (like the Archimedean spiral) may not be suitable for capturing these new features. Consequently, template modification is required which might contradict with the clinical procedures already in use.

2.2.2 Drawing strategy analysis

The latest trend in the computerized analysis of neuropsychological drawings is to observe the drawing strategy adopted by the subject. Although the idea is not so recent, it has gained renewed popularity due to the availability of online data capturing systems. Such analysis focuses on the behavior and preferred drawing strategy of the subject while drawing a stimulus. Instead of analyzing the end product or the hand movement involved in the drawing action, constructional sequence order and planning of the constituent parts of the drawing are observed.

Remi et al. propose a similar technique in [49], for the assessment of hand-drawn samples of children with learning difficulties. A template consisting of handwritten sentences and a set of geometrical shapes is used. The constituent parts of each drawn stimulus are localized using shape-based clustering and a collection of drawing sequences. Deformation-based features are then extracted to determine the signs of learning and writing difficulties. In a similar attempt, authors in [50] employ syntax analysis on the online samples of children's drawings. Selective patterns of different drawing-based tests are used as templates to determine the gender and handwriting ability of each participant.

A more comprehensive technique of modeling the sketching gestures of subjects is proposed in [51]. The drawing order of the shape primitives is observed in addition to their organizational relation with the neighboring primitives. Based on predefined rules (like specific angle, sequence and the number of primitives required for each shape class), the quality of the produced sketch is evaluated. The system is tested on both hand-drawn samples and synthetic shapes from HHreco [52] dataset. The drawing gestures of patients with disabilities are also analyzed in a series of related studies [53–55], using an optoelectronic system. In another series of related works [56, 57], an attempt is made to determine the correlation between the preferred polygon drawing strategy of school children with their handwriting skill development. Drawing sequences of 178 school children (ages 6–7 years), captured using a digitizer tablet are employed. A 63.48% success in correlating the estimated drawing strategy and the predicted handwriting performance is achieved using an SVM classifier. A similar approach is previously reported in [58].

2.3 Key inferences and motivation

After an extensive review of the techniques employed in the literature, the following inferences can be derived regarding both approaches.

- Visual analysis-based techniques that determine the quality of the whole shape by assessing the individual constituent parts independently are marked by the challenges of localization and segmentation. In most cases, localization of the intended part is impossible due to the highly unconstrained nature of the drawing itself. On the contrary, visual analysis-based techniques derive inference from the entire drawing, not only require to assess the quality of the constituent parts but their correlation with each other as well. This requires an extensive rule-based approach that is impractical and at times insufficient.
- Procedural analysis-based techniques observing the hand movement or preferred drawing strategy do not depend on the visual feedback of the final response but rely on the features that represent the motor and cognitive functionality. Although it is a promising approach, nevertheless, it not only requires a sophisticated acquisition equipment but also requires the modification of the test administration and inference procedure. As discussed earlier, this raises issues like test and feature validity and may face resistance from the target users, i.e., clinical practitioners.

In this study, we attempt to address the aforementioned gaps in the literature by employing CNN-based deformation modeling. The basic motivation behind the use of CNNs is that they can extract discriminating features from various regions of the complete drawing without localization or segmentation. The convolutional filters employed on the different layers of a deep CNN produce activations for specific discriminating patterns. CNNs can generalize features across a wide variety of deviations, and thus, the need to explicitly define heuristics for each deviation is no longer necessary. Instead, visual samples (training data) for each type of deformation are required for modeling. In case, if the focus has to be given to a particular region in the drawing, we can employ data augmentation specific to that deformation while training the classification layer. For instance, to extract tremor features from the shape edges, we employ a median residual resultant representation of the raw image. The median residual resultants are known to highlight the fine irregularities in the drawing and thus can be employed for this purpose.

By utilizing the offline samples collected by psychologists, we resolve the need for any modifications in the original test conduction protocol. Furthermore, since the ground truth has also been prepared with the help of the domain experts, feature validation issues are addressed as well. The complete methodology and the proposed system architecture are explained in the upcoming sections.

3 Methodology

In this section, we explain the proposed methodology for employing CNNs for the identification of clinical deformations present in neuropsychological drawing samples. The key concepts involved are transfer learning and data augmentation that are discussed in accordance with our proposed system. Although our study aims to present a generic framework for the computerized analysis of neuropsychological drawings, however, for the purpose of performance evaluation, the proposed framework is customized to analyze and score the BGT drawings. The BGT test and its scoring criteria are also discussed in detail.

3.1 Deep convolutional neural networks and feature extraction

CNNs (popularly known as ConvNets) are nonlinear models capable of learning nonlinear features [12]. A typical deep CNN architecture primarily comprises of a series of convolutional layers followed by some intermediate layers [13]. The connectivity between a pair of consecutive layers is designed to facilitate the detection of distinctive local patterns in an input image. These patterns can then be employed for various visual classification-based tasks including deformation identification. To effectively employ CNNs for the problem under consideration, we must first understand the processing of the fundamental constituent layers.

A convolutional layer is designed to take a high-order tensor (order 3 for an RGB image) as an input. It then transforms the input tensor into feature maps. A feature map is obtained by applying convolution with a matrix of weights (or kernel), followed by the addition of bias. Each layer has a number of kernels that replicate the process to generate a set of feature maps as an output. The processing of a particular convolutional layer l in a deep CNN consisting of L layers can be generalized by Eq. 1, where $l = 1, \ldots, L$.

$$X_l^{(m)} = f\left(\sum_{n=1}^{N} W_l^{(n,m)} * X_{l-1}^{(n)} + B_l^{(m)}\right) \tag{1}$$

$X_{l-1}^{(n)}$ represents one of the N input feature maps from the previous layer, while $X_l^{(m)}$ represents an output feature map of the current layer, where $m = 1, \ldots, M$. Each input feature map is convolved with the kernels $W_l^{(n,m)}$ of the layer l, resulting in a sum of N convolutions denoted by $\sum_{n=1}^{N} W_l^{(n,m)} * X_{l-1}^{(n)}$. Each spatial position in a feature map has its own bias which is presented as a matrix $B_l^{(m)}$. f represents an activation function that is applied on the result. In most cases, the *Rectified Linear Unit (ReLU)* [59] (Eq. 2) is the preferred activation function. ReLU activation preserves all the positive values from the resultant feature map and eliminates the negative ones by converting them to zero. This enhances the mapping capability of a ConvNet for the desired attributes.

$$f(x) = max(x, 0) \tag{2}$$

A convolutional layer is parameterized by a large number of features. In order to reduce the dimensionality of the resultant feature maps, a subsampling (or pooling) layer is commonly introduced before the next convolutional layers. The pooling layer applies an aggregation function on the input maps to select only the maximal values from the local subregions.

The convolutional, ReLU and pooling layers together serve as a feature extraction convolutional base. Deep CNNs exploit the combined benefits of the regional convolutions and a layered hierarchy to learn effective representations for specific visual recognition tasks. The depth of a CNN architecture depends on the level of abstraction required. The intermediate convolutional layers are designed to extract distributed representations from the input data. On the contrary, a fully connected layer employs all the features extracted by the previous layer in order to amplify the selected distinctive patterns. There may exist one or more fully connected layers at the end of a deep CNN similar to a conventional 'Multi-layer Perceptron (MLP).'

3.2 Transfer learning and pre-trained ConvNets

A forward run in a CNN model is employed for the prediction of probabilities; nevertheless, before a CNN model is ready to predict, it requires an extensive training. The training process involves running the CNN network in both directions, i.e., backward and forward. For instance, let us consider a training sample x_i given as an input to a CNN model. The model is run in the forward direction, and an output x_L is obtained. x_L is then used to compute the loss z, i.e., the difference between the output and target ground truth. The loss is then used to update the network parameters by employing backward propagation [12]. The training process is repeated until a minimum value for loss is achieved.

Training a CNN model relies on the tuning of millions of parameters, which is only possible in the case of sufficiently large amount of annotated data. The lack of sufficient training samples has been one of the major limiting factors in the use of CNNs for neuropsychological drawing analysis. Nevertheless, recent advances in machine learning have proposed several solutions to overcome this limitation. Transfer learning is one such alternative, which allows a CNN architecture to transfer learned weights across different source and target datasets. Leveraging on the idea that the convolutional base of a CNN model is more generic or task-independent, transfer learning allows us to train these initial layers on a different task with a larger dataset. Once their parameters are optimized using the source dataset, these weights can be frozen and only the fully connected layers can now be trained on the task-specific dataset. This concept is known as *fine-tuning*, and it not only enables the use of deep CNNs on the problems with smaller datasets but also reduces the computational cost and training time significantly.

The features extracted from a pre-trained convolutional base can also be used to train a separate classifier. In other words, rather than tuning the layers of a pre-trained ConvNet on the target dataset (by continuing back propagation), the ConvNet can be employed as a feature extractor mapping the input patterns to feature vectors using a forward pass only and ignoring the fully connected layers. The extracted features can then be used to train another classifier. The CNN-SVM combination, for instance, has been investigated in a number of studies [60–62], for various visual recognition tasks. To fully assess the potential of CNNs in the analysis of neuropsychological drawings, we design our system to employ pre-trained ConvNets for both *shape recognition* and *deformation classification*. The details of both are presented in the subsequent sections.

3.3 Scoring of Bender-Gestalt test: a case study

As discussed earlier, to comprehend the complexity of the problem and to evaluate the effectiveness of the proposed technique, we employ the analysis of the BGT test as a case study. BGT is a popular drawing-based psychometric test employed by clinical psychologists for the screening and differential diagnosis of various neurological disorders [63–67]. The test comprises of a set of nine different templates or gestalts as shown in Fig. 7. The test conduction protocol requires the subject to copy or recall each template on a single sheet of paper using a pencil (Fig. 1c). Since BGT is a visual perception assessment test, focus is on the outcome rather than the procedural strategy involved. Several scoring systems [3, 68–71] have been proposed for the estimation of deformations in BGT drawings; however, 'Lacks' scoring system' [29] is popularly employed by practitioners and is being considered in this study as well.

Lacks' scoring system is based on 11 essential discriminators of brain dysfunction inspired by the Gestalt psychology [29]. It determines the presence/absence of some of these errors using all of the nine BGT templates, while some are determined using a subset. The scored errors and the templates on which they are applicable are presented in Table 3. A brief description of the errors is outlined in Table 4. Figure 8 shows some of the examples of Lacks' deformations (errors) in different BGT shapes.

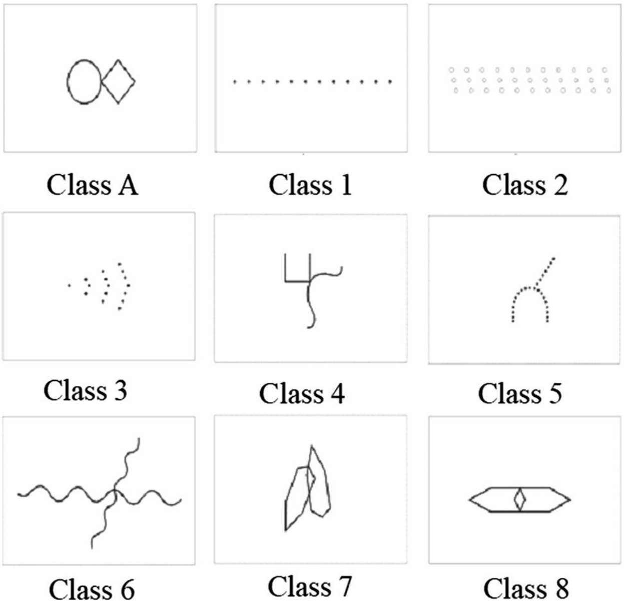


Fig. 7 Nine Bender-Gestalt shapes

Table 3 Scoring sheet using Lacks' scoring system

Error	BGT shape class								
	A	1	2	3	4	5	6	7	8
Rotation	✓	✓	✓	✓	✓	✓	✓	✓	✓
Overlap	X	X	X	X	X	X	✓	✓	X
Simplification	✓	✓	✓	✓	✓	✓	✓	✓	✓
Fragmentation	✓	✓	✓	✓	✓	✓	✓	✓	✓
Retrogression	✓	✓	✓	✓	X	✓	X	✓	✓
Perseveration	X	✓	✓	✓	X	✓	X	X	X
Collision	✓	✓	✓	✓	✓	✓	✓	✓	✓
Closure	✓	X	X	X	✓	X	X	✓	X
Motor Incoordination	✓	✓	✓	✓	✓	✓	✓	✓	✓
Angulation	X	X	✓	✓	X	X	X	X	X
Cohesion	✓	✓	✓	✓	✓	✓	✓	✓	✓

A detailed study of the scoring system under consideration reveals the complexity of the problem at hand. Several conditions determine the presence or absence of a particular deformation in a shape. Deformations are highly shape dependent and may be measured differently across different templates. Some conditions may be considered as an error in one shape while normal in another. Same characteristics may be considered as one type of error in one shape while another type of error in a different shape. Multiple deformations can coexist in a single shape.

3.4 System architecture

This section presents the details of our proposed system for the analysis of BGT responses. An overall system architecture is presented in Fig. 9, while details of each module are discussed in the subsequent sections.

3.4.1 Input acquisition

In order to keep the original test conduction protocol, our proposed system is designed to take offline scanned images of the test samples produced by the subjects (healthy/patient) as raw input. An important consideration at this stage is to decide whether to give the complete image containing multiple templates as an input or to provide each template response separately. As discussed in the Lacks' scoring section, deformations can be shape-specific. Due to this reason, the target shape localization and recognition is an important preprocessing step. Automated BGT shape segmentation has previously been attempted in [72]. Since segmentation is not a prime objective of this study, we will not address it. Furthermore, due to the challenges outlined in [72], we do not want to affect the performance of our

Table 4 Brief description of Lacks' indicators of brain dysfunctions

Errors	Description
Rotation	To rotate a shape beyond 80° but less than 180° while drawing
Overlap	To omit, rework or simplify overlapping portions of a shape
Simplification	To simplify a primitive component, to miss fine details or to draw the joining parts of a shape separately. It is marked differently across shapes
Fragmentation	To omit the primitive components of a shape or change their organization in a way to completely alter the original template
Retrogression	To substitute an advance component with a primitive one. It is marked differently across templates
Perseveration	To repeat a component in the original template beyond its required quantity
Collision	To draw two or more templates with in proximity of each other. The drawn shapes can be touching or overlapping in extreme cases
Closure	To show difficulty in drawing closed shapes like circle and polygons
Motor incoordination	To introduce signs of tremor and jerk while drawing smooth strokes
Angulation	To simplify or completely omit the angular parts of a shape
Cohesion	To introduce significant size disparity between the components of a shape

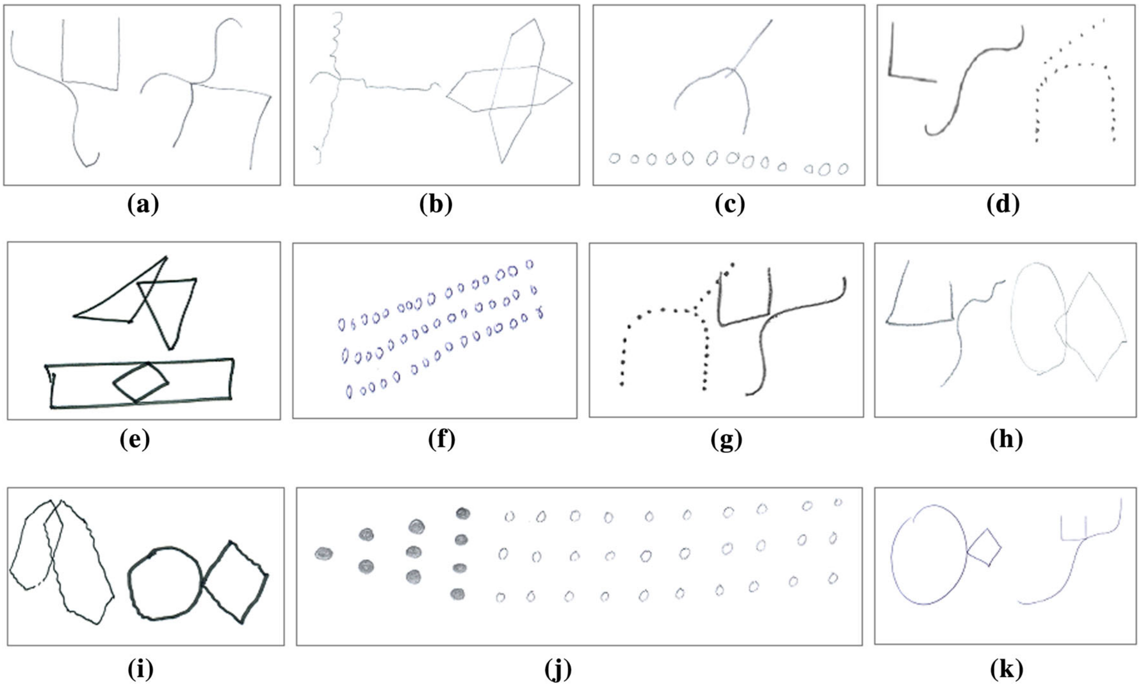


Fig. 8 Examples of deformations: **a** rotation in BGT template 4, **b** Overlap difficulty in BGT template 6 and 7, **c** Simplification error in BGT template 5 and 1, **d** Fragmentation in BGT shapes 4 and 5, **e** Retrogression in BGT shapes 7 and 8, **f** Perseveration in BGT template 2, **g** Collision of BGT shapes 5 and 4, **h** Closure difficulty in BGT shape 4 and A, **i** motor incoordination in BGT template A and 7, **j** angulation in BGT shapes 2 and 3 and **k** cohesion in BGT template A and 4

deformation classification module due to incorrect segmentation. Hence, all the constituent shapes $(I_1, I_2, \ldots, I_N)$ of a BGT sample are independently fed to the system one by one; however, scores from all shapes are accumulated to give a sample-wise decision.

3.4.2 Shape recognition

Shape recognition is vital as some deformations are not applicable to all shapes (as discussed in Lacks' scoring section). Two pilot studies [6, 73] are previously conducted that analyze shape context descriptors and CNN-based

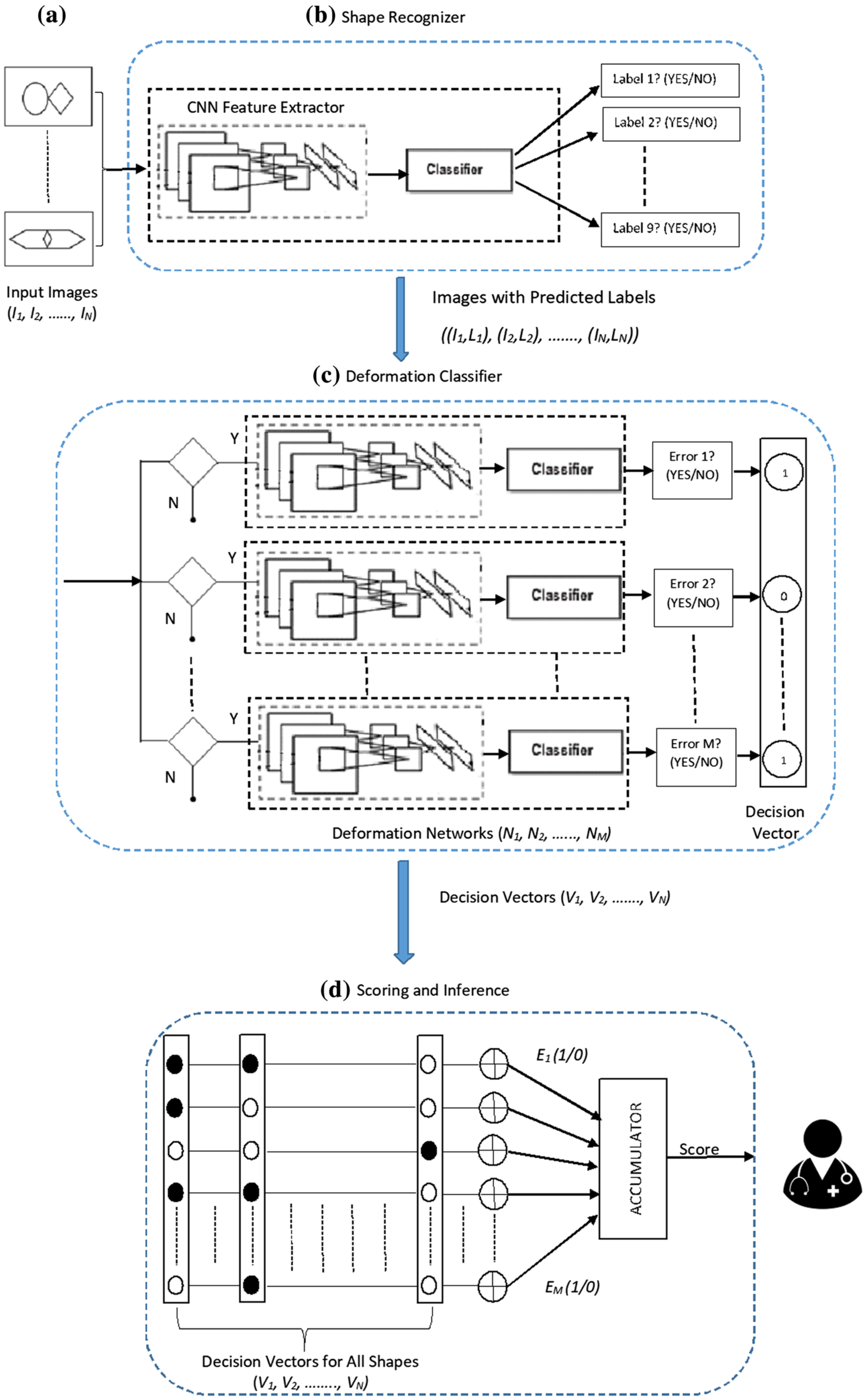

(a)
(b) Shape Recognizer
CNN Feature Extractor
Classifier
Label 1? (YES/NO)
Label 2? (YES/NO)
Label 9? (YES/NO)
Input Images
(I_1, I_2, ..., I_N)
Images with Predicted Labels
((I_1,L_1), (I_2,L_2), ..., (I_N,L_N))
(c) Deformation Classifier
Y
N
Classifier
Error 1? (YES/NO)
Error 2? (YES/NO)
Error M? (YES/NO)
1
0
1
Deformation Networks (N_1, N_2, ..., N_M)
Decision Vector
Decision Vectors (V_1, V_2, ..., V_N)
(d) Scoring and Inference
E_1(1/0)
E_M(1/0)
ACCUMULATOR
Score
Decision Vectors for All Shapes
(V_1, V_2, ..., V_N)

Fig. 9 Proposed system architecture for deformation modeling and classification of BGT shapes: **a** individual segmented shapes from each BGT sample are given as input, **b** features extracted from each shape are fed to a classifier to determine the shape class, **c** recognized image is then fed to each deformation network to determine the presence of the corresponding deformation, and **d** decision vectors from each sample are used to generate the final score

shape recognizers, respectively, for the classification of the nine BGT shape classes. As expected, CNN-based features outperformed handcrafted descriptors; therefore, we employ a similar architecture as that used in [73]. The architecture as shown in Fig. 9b consists of a pre-trained convolutional base as a feature extractor. Each input image $(I_1, I_2, \ldots, I_N)$ given to the feature extractor is resized to match the input layer of the respective ConvNet employed.

The extracted features are then fed to a classifier (e.g., SVM, LDA, etc.) as demonstrated in [60–62, 74], which then predicts the shape class label $(L_1, L_2, \ldots, L_N)$. We conduct an in-depth empirical analysis on the performance of various pre-trained ConvNets in combination with a number of classifiers, to determine the best CNN-classifier combination for the BGT shape recognition task. The architectural details of the pre-trained ConvNets and the hyper-parameter specifications of the classifiers employed are presented in Sect. 4 of the paper.

3.4.3 Deformation modeling and classification

Figure 9c demonstrates the conceptual model of our proposed deformation classification module. In order to model deformations, we again employ pre-trained CNN architectures to extract deformation-specific features which are then used to train a classifier. The outcome of the classifier is a binary decision regarding the presence/absence of a particular deformation. The main design issue in this module was to decide whether to train a generic model for each deformation, for instance, one network to model the rotation error across all the shapes or to train individual shape-based deformation models (i.e., separate rotation models for each of the nine BGT shapes). However, to avoid similar customization that is criticized in tailored rule-based approaches, a generic model for each deformation across all the shapes is designed to provide generality. Another reason to avoid a shape-specific deformation modeling approach is the scarcity of training data for a particular deformation across all templates.

Feature extraction is further enhanced by using augmented data customized specifically to represent a certain deformation. The details of our deformation-specific augmentation techniques are discussed in the next section. Once all the deformation models are trained, we feed the pre-segmented shapes with shape labels $((I_1, L_1), (I_2, L_2), \ldots, (I_N, L_N))$ to our proposed deformation classification module. The module consists of all the deformation-specific networks $(N_1, N_2, \ldots, N_M)$. Each input shape is assessed independently by all of the applicable deformation networks. The decision from each deformation model is then stored in a decision vector. Each decision vector consists of the results [error (1)/no error (0)] of all the deformation models applicable to one BGT shape. The decision vector also contains a 'not-applicable' indicator for all the deformations that are not applicable to any particular shape.

3.4.4 Deformation-specific data augmentation

Data augmentation is a common practice to overcome the data scarcity, overfitting and class imbalance issues. In our proposed methodology, deformation-specific augmentation is primarily being performed to provide the maximum possible deviations of each deformation class. It is mentioned in the previous sections that the availability of shape-wise samples for each deformation is not feasible in a real-life scenario. Due to this reason, some deformations classes do not have considerable representation of each shape-wise data. Although shape-wise deformation modeling is not being performed, nonetheless, to generate some samples of the missing shape-wise deformations, we employ deformation-specific transformations on the non-erroneous shape samples.

Several techniques have been proposed in the literature to augment images [75]. Some techniques (e.g., Geometric transformations) augment images in the data space, while others (e.g., SMOTE [76]) augment data in the feature space. In a domain-specific problem like clinical deformation classification, the preservation of class labels after augmentation is a critical concern. Although feature space-based techniques can provide better post-augmentation data cleaning, nevertheless, we require the domain expert to label the augmented data as well. For this purpose, a data space-based approach is applied, where different transformations are applied on the existing training samples to generate similar examples. Our deformation-specific transformations can be categorized as 'generic' and 'shape-specific.'

All deformations except 'Simplification,' 'Retrogression' and 'Perseveration' have common characteristics across all of the shapes on which they are applicable. Due to this reason, augmentation techniques for these deformations are relatively generic. A brief description of the transformations applied for a generic deformation across all of the shapes is given below:

- To generate data with 'Rotation' error, all shapes from the original training samples are rotated 2° apart to achieve the rotated copies of the original shape between 80° to 180° or mirror image (Fig. 10a). However, caution is taken to ensure that no shape with original rotation error becomes error-free. Also for some BGT shapes, mirror image produces the same shape as the original (e.g., BGT shape 1 and 8, etc.). For such shapes, 180° rotation is not considered. Like rotation, the generation of angulation examples is achieved by rotation of the original images of template 2 and 3, at the angles between 45° and 80°.
- Samples for the organization-based deformations like 'Overlapping difficulty,' 'Collision' and 'Closure difficulty' are generated by employing a controlled translation of the constituent parts of a shape. For instance, the Overlapping difficulty error for BGT shapes 6 and 7 is generated by translating and merging the individual

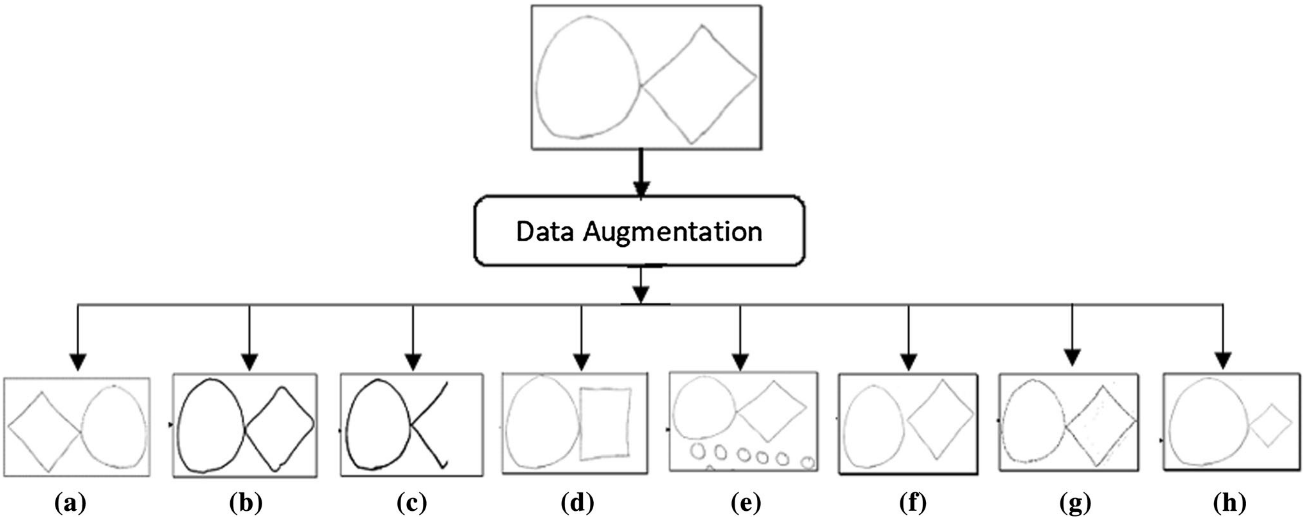


Fig. 10 Example of the deformation-specific augmentation results for BGT shape A **a** mirror image produced by rotation, **b** Simplification of sharp angles of diamond into curves using morphological operations, **c** Fragmentation introduced by converting part of foreground image into background, **d** replacement of constituent diamond with square to produce Retrogression example, **e** translation of BGT shape A and 2 to produce Collision, **f** significant separation of circle and diamond for Closure difficulty, **g** inverted median residual of original BGT shape A for motor incoordination and **h** resizing of diamond to produce cohesion

hand-drawn samples of their constituent parts in a way to produce an incorrect or missing overlap. Similarly, translation is also applied to the separately drawn constituent parts of BGT shape A, 4 and 7, to join them at the wrong points, to represent the Closure difficulty as shown in Fig. 10f. In the case of 'Collision,' different BGT shape templates are translated within proximity to each other. For some scenarios, two shapes are translated as shown in Fig. 10e, while for other three or more BGT shapes are translated close to each other.

- Besides rotation and translation, other meaningful representations of the raw data are also evaluated for our selective augmentation technique. For instance, the median residual of the hand-drawn samples of PD patients has been employed in [9], to detect tremors by highlighting the fine irregularities present in the shape contours. A similar technique is employed in our study to represent 'Motor Incoordination.' After selecting shapes representing the motor incoordination error, we generate their median residuals. Both representations (i.e., raw and median residual) of the examples are used to generate relevant features as training data. For illustration purposes, Fig. 10g demonstrates the inverted image of the median residual of a sample of BGT shape A.
- To produce 'Fragmentation' data, a $0.25r \times 0.25c$ sized window is randomly placed on the original shape image of size $r \times c$. If the window contains the foreground pixels, then they are converted into background and the image is saved. Consequently, several copies of the original shape image with the missing details are created. Figure 10c shows an example of automatically introduced Fragmentation error in BGT shape A.
- Size imbalance of the main constituent parts of a shape represents the 'Cohesion' error. Such shapes whose parts are already separated or can easily be separated by erosion are mostly used to generate the examples for cohesion. Some of the separated constituent parts are scaled up, while others are scaled down (Fig. 10h) and merged to generate a shape with disproportionately sized components.

As mentioned earlier, errors like 'Simplification,' 'Retrogression' and 'Perseveration' are marked by different characteristics across different shapes. Due to this reason, their data generation is highly shape-specific, as discussed below:

- To generate the Perseveration examples, extra row(s) or column(s) of dots or circles in BGT shapes 1, 2 and 3 are added by replicating and merging the constituent parts of the shapes. In case of replacing circles with dots in BGT shapes 3 and 5, we apply the morphological technique of hole filling followed by erosion with a disk-shaped structuring element of an appropriate size. Both techniques represent Perseveration data.
- Retrogression error is scored when a constituent part of a BGT shape is replaced by a primitive shape, i.e., dots with dashes, circles with loops, triangle, square or rectangle for a diamond or hexagon. One method to generate this error is to create synthetic geometric shapes and merge them accordingly. However,

synthetic shapes may not represent the imperfections of a hand-drawn shape. Therefore, to generate a near-realistic data, we asked some subjects to draw primitive geometric shapes like triangles, rectangles, squares and circles on separate sheets of paper. The individual shapes are then segmented and merged as the deformed BGT shapes (e.g., Fig. 10d). Same augmentation technique is applied to all other errors where another shape replaces the original constituent part.

- Generally, 'Simplification' is marked whenever the fine details of the drawing are distorted. For instance, angles are curved, overlapping is missed by a great distance, and fine dots are replaced by tiny circles. For overlapping shapes, we applied the same technique which was used for the Overlapping difficulty, i.e., the constituent parts are translated in a manner to allow maximum separation between them. BGT shapes A, 4, 6 and 8 are simplified by applying various morphological operations to convert their sharp angles into smooth curves (Fig. 10b). For BGT shape 1, where Simplification means the replacement of dots with circles, we applied dilation with caution to avoid joining dots together. Once dilated and enhanced, a boundary extraction technique is applied to form circles from blobs. We also used hand-drawn samples of small circles to generate a row at different distances.

Figure 10 shows examples of deformation-specific augmented data, generated for BGT template 'A,' by employing some of the proposed techniques.

3.5 Scoring and inference

As discussed in the previous sections, the decision vector for each shape template contains the results of all the deformations applicable to it. These decision vectors $(V_1, V_2, \ldots, V_M)$ provide useful information to the practitioner regarding the sample drawn by a particular subject. A practitioner can use the decision vectors to validate his/her decision or can apply them to draw various statistics regarding the frequency of a particular deformation as well. Nevertheless, for this case study, we apply inference rules outlined in the Lacks' scoring manual.

According to Lacks' scoring, the occurrence of an error is more important than the frequency of the error. As a result, for all the instances of a particular type of error across all BGT shapes, a score of one is generated. For instance, if the rotation error exists in all of the nine BGT shapes, the score for the rotation error will be considered as one, irrespective of its frequency of occurrence. The same will be considered even if only one of the BGT shapes is rotated between 80° and 180°. Similarly, scores for other deformations are also generated. The final score of the sample will be the sum of all the deformations which are independently scored. The practitioner then decides a threshold value to determine whether the score lies above or below it. Any score above the threshold indicates a sign of brain dysfunction. The flexibility of the threshold value is due to the demographics of the subjects taking the test. To automate the generation of a final score, in the same way, we apply 'logical OR' on each of the error decisions present across all the decision vectors and feed the results to an accumulator. The system then displays a final score of the test as practiced by a clinical expert.

4 Experimental protocol

In this section, we describe the experimental protocol employed in our study. As discussed earlier, the prime objective of our study is to assess whether the CNN-based visual features can represent clinician deformations sufficiently to be employed for a high-level domain knowledge representation problem like the computerized assessment of neuropsychological drawings. To assess that, we further evaluate the applicability of the pre-trained CNN models:

- As feature extractors for inter-class shape recognition even with a high degree of deformations and a limited amount of training data.
- To enhance intra-class variations and inter-class similarities using selective augmentation for deformation classification.

In order to assess the performance of our proposed *shape recognition* and *deformation classification* methodologies, we conduct separate experiments. All experiments are conducted on a CUDA-enabled NVIDIA GPU with computing capability of 6.0 using Keras and Scikit-Learn libraries. The ConvNets employed in the experiments have already been pre-trained on ImageNet [77], and their training is not a part of this study.

4.1 Sample acquisition and ground truth labeling

As discussed earlier, we have selected the analysis of BGT shapes for the evaluation of our proposed system. Nonetheless, to the best of our knowledge, there is no publicly available dataset of BGT drawings. Due to this reason, BGT drawing samples of 60 (30 control/ 30 patient) participants were collected with the help of the Institute of Professional Psychology, Bahria University, Islamabad, Pakistan. According to Lacks' scoring criteria, the maximum attainable score on a BGT test is 11 (not considering the scoring for the time taken), while the minimum score is zero. The presence of an error is scored as one point irrespective of its frequency of occurrence across the different shape classes. The total score in a BGT test is the quantification of the different types of errors present in the complete sample. An accumulated score of 5 is commonly considered as a cutoff in Lacks' scoring manual. A higher score is an indication of a possible brain dysfunction. The same criteria are considered for the inclusion of the subject population into the control and patient groups, respectively. The range of the scores obtained in the control group is [0–5], while that in the patient group is [6–11]. Table 5 outlines the demographic- and education-level data of the participant groups along with their mean BGT test performance scores.

All the participants were already enrolled in an ongoing research being conducted by the Institute of Professional Psychology in collaboration with a local school. Although disease diagnosis is not a part of the study under consideration, nevertheless, the patient group originally consisted of the children with various levels of learning disabilities. The evaluation for the learning disabilities was conducted by employing a test battery consisting of three tests: *(a) Bangor Dyslexia Test (BDT)* [78], *(b) Wide Range Achievement Test-4 (WRAT-4)* [79] and *(c) Bender-Gestalt Test (BGT)*. Since our prime concern is the scoring of the BGT test, only the samples of the BGT test were collected from the Institute of Professional Psychology after the approval of the university's ethic committee. The selected samples were originally scored by three experts with an inter-rater reliability of $M_r = 0.89$, suggesting that all scorers agreed on 89% overall BGT test scores. However, a higher variability was observed in the individual shape error scoring. To normalize this variability, majority voting was applied while creating the ground truth for this study.

4.2 Training and test distribution for shape recognition

For shape recognition, the segmented drawing samples of all 60 subjects are employed, where each sample consists of the nine BGT shape classes, resulting in a total of 540 shape samples (60 samples for each BGT shape class). Fivefold cross-validation is employed and in each fold, and the data are divided into training and test sets accordingly, with each set containing an equal representation of samples from the two subject groups (healthy/patient) under study. For fair evaluation, caution is taken while dividing the data for training and testing. All the segmented shapes selected for training belong to the samples originally selected for training, and the same is done for the testing samples.

4.3 Training and test distribution for deformation classification

Data distribution for deformation training and testing was not so straightforward. Despite an equal number of samples of children from both groups, the individual deformation examples were highly imbalanced. Figure 11 shows the percentage of each type of deformation in the original dataset. Out of the 540 individual shape samples, 299 contained deformations that too was not evenly distributed across all shapes. This was one of the motivating factors for training independent deformation networks instead of a single multi-class one. This enables classification of all deformation classes independently from each other, thus reducing the impact of an overall class imbalance. After separating the classification for each deformation, the data scarcity issue for some of the classes is overcome by the data augmentation techniques already discussed in the previous sections. After sufficient data for each class are ensured using the deformation-specific augmentation, each of the deformation class datasets is further divided into fivefold for cross-validation.

4.4 Pre-trained CNN architectures employed

We investigated a number of pre-trained CNN architectures for shape recognition and deformation classification tasks. These include AlexNet [13], VGG16 and 19 [80], SqueezeNet [81], GoogLeNet (Inception v1) [82], Inception v3 [83], ResNet50 and 101 [84] and DenseNet201 [85]. AlexNet, VGG16 and VGG19 architectures consist of a series of alternating blocks of convolutional (with ReLU activation) and pooling layers. Each of these

Table 5 Demographic, education and BGT performance levels of the participants

Gender	Number	Mean age (years)	Education level (grade)	Mean BGT score
Patient group				
Males	15	13.2 ± 2.0	6–10	8.46 ± 1.98
Females	15	12.7 ± 2.2	6–10	8.20 ± 2.01
All	30	12.9 ± 2.1	6–10	8.33 ± 1.99
Control group				
Males	15	13.6 ± 2.1	6–10	2.26 ± 1.48
Females	15	12.9 ± 2.5	6–10	2.02 ± 1.40
All	30	13.2 ± 2.3	6–10	2.14 ± 1.44

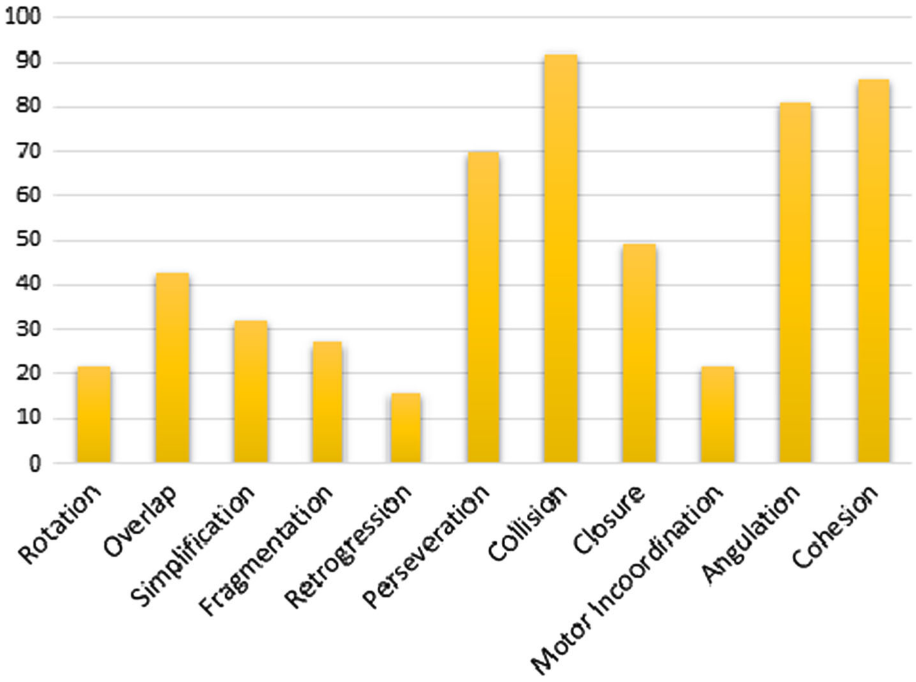


Fig. 11 Percentage of examples of each deformation class in the original dataset

has three fully connected layers, out of which, the last layer is used for the classification purposes. To extract the pre-trained features from these models, the last fully connected layer is removed. SqueezeNet is a CNN microarchitecture with compressed filter sizes. The architecture employs *Fire modules* comprising of a *squeezed* convolutional layer (with 1x1 filters) feeding into an *expanded* layer (with a combination of 1x1 and 3x3 filters). The fire modules allow network learning with a comparatively lesser number of parameters.

The series of Inception architectures introduced the concept of *Inception blocks*, in which different convolutional filters and pooling layers are concatenated to enhance learning. ResNet architectures (ResNet50 and ResNet101) represent the residual networks, where the layers contain direct, additive connections referred as *Skip connections*. The idea of skip connections has been extended to connect the subsequent blocks of densely connected layers in DenseNet architectures. The dense blocks alternately comprise of 1x1 convolutional filters and max-pooling layers in order to reduce the number of tunable parameters. Contrary to the skip connections in the residual networks, the output of the dense blocks is not added but instead concatenated.

All the networks employed in our experiments have been pre-trained on the ImageNet dataset. Features learned by different CNN layers impact the classification performance [86]. Studies like [87] empirically support the notion that features from the deeper layers are more effectively transferable on the target dataset. Due to this reason, we have also employed the features learned from the deeper layers (mainly the last fully connected layer). Table 6 enlists the architectural details of the pre-trained ConvNet models employed in our study; the depth of each network, the input image size and the layers from which the learned features have been extracted are mentioned. Since we employ pre-trained ConvNets as feature extractors rather than training them from scratch, the hyper-parameters involved in their training are not mentioned.

4.5 Multi-class and binary classifiers

To assess the potential of CNN-based features in both scenarios, i.e., shape recognition and deformation classification, we have employed a number of popular supervised learning classifiers. Features extracted from the pre-trained models are fed to train these classifiers independently to observe the impact of the classifiers on the performance. It is worth mentioning that shape recognition is a multi-class (i.e., nine BGT shapes) classification problem, while deformation classification is a binary class (i.e., error or no error) problem. Therefore, each classifier is trained accordingly. For shape recognition, four classifiers are employed. These include the support vector machines (SVM), linear discriminant analysis (LDA), Naive Bayes (NB) and decision trees (DT). Brief details of the classifiers and the respective hyper-parameters involved in training are given below.

- Discriminant analysis is a statistical method that facilitates decision-making by employing dimensionality reduction on the input data to rely on only the most

Table 6 Summary of pre-trained CNN architectures employed

Model	Source dataset	Depth	Input image size	Feature extraction layer	Feature dimensions
AlexNet	ImageNet	8	(227 × 227)	fc7	4096
VGG16	ImageNet	16	(224 × 224)	fc7	4096
VGG19	ImageNet	19	(224 × 224)	fc7	4096
SqueezeNet	ImageNet	18	(227 × 227)	pool10	1000
GoogLeNet	ImageNet	48	(224 × 224)	loss3-classifier	1000
InceptionV3	ImageNet	22	(299 × 299)	predictions	1000
ResNet50	ImageNet	50	(224 × 224)	fc1000	1000
ResNet101	ImageNet	101	(224 × 224)	fc1000	1000
DenseNet201	ImageNet	201	(224 × 224)	fc1000	1000

discriminant values. A linear discriminant model is applied on the extracted CNN-based features. An LDA attempts to minimize the variance between the input features of a class in such a way that it maximizes the distance between the means of the distinct classes. Ranking threshold is an important hyper-parameter while applying an LDA. It is the value that determines the inclusion and exclusion of an instance in the feature space. A threshold value of 0.0001 is selected after an extensive empirical analysis.

- SVM is a non-probabilistic classifier that models a hyperplane to separate the labeled classes. A linear, one-versus-all SVM is trained for the shape recognition task. The tolerance value is set to 0.0001, and cost parameter is set to 1.
- NB is a probabilistic classification technique. We trained a multinomial NB model on the extracted CNN feature vectors. A Laplacian smoothing prior is applied to prevent the impact of zero probabilities on the decision.
- DT is a predictive model that can be used for classification. An important parameter in a classification tree is the number of splits (k) which controls the depth of the tree. The value selected for the tree splits is k=50 for shape recognition.

The selection of the hyper-parameter values is carefully done after an extensive empirical analysis. For the deformation classification task, we employed an LDA classifier with similar hyper-parameter values.

5 Results and analysis

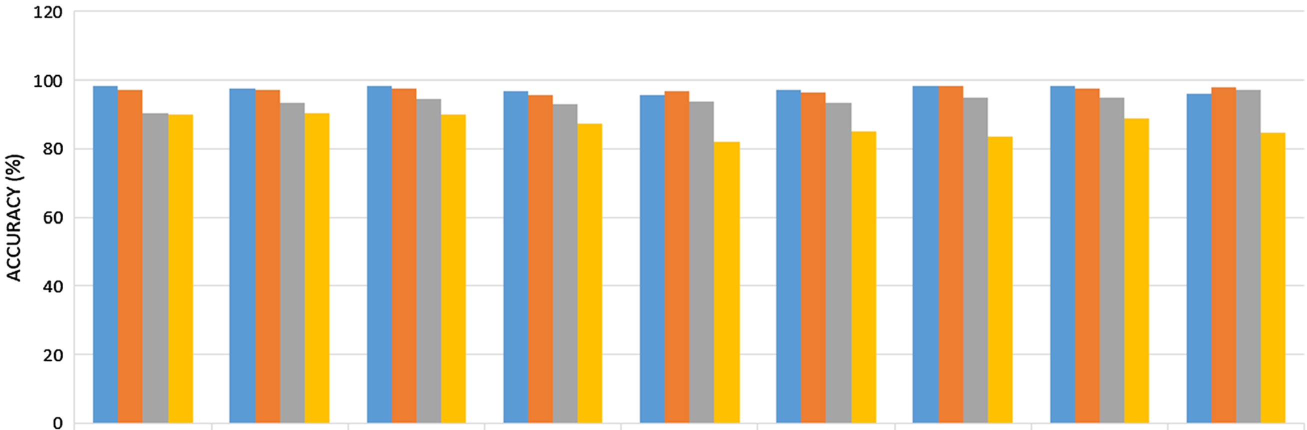
This section discusses the results of our proposed empirical analysis for the BGT shape recognition and deformation classification methodologies presented in this study.

5.1 Shape classification results

To assess the effectiveness of the proposed shape recognition technique, we first evaluate the overall classification accuracies achieved by the combination of each CNN architecture employed with the aforementioned classifiers. As mentioned earlier, a total of 540 shape data (with 60 samples of each shape) are divided into training and testing using fivefold cross-validation. For each fold, the overall shape accuracy is computed as $\frac{t_p+t_n}{t_p+t_n+f_p+f_n}$, where t_p, t_n, f_p and f_n represent the total number of true positives, true negatives, false positives and false negatives, respectively, achieved by the system for all the classes. Mean classification accuracy is then computed from the accuracies achieved by each of the fivefold. Figure 12 presents the mean classification accuracies for each (CNN-classifier) combination.

It is observed that the performance of the features extracted from each pre-trained ConvNet employed is comparable with one another when fed to the same classifier. This supports our initial claim that the pre-trained CNN architectures can be successfully employed to a limited shape class dataset, even without augmentation. Furthermore, it is observed that the highest classification accuracy (i.e., 98.33%) is achieved by training an LDA with features extracted from AlexNet, VGG19, ResNet50 and ResNet101, independently. This shows that shape classification is not significantly affected by the choice of the CNN architecture employed. However, the choice of the classifier is important as both LDA and SVM outperformed NB and DT, significantly. Both LDA and SVM reported comparable accuracies across each CNN architecture, with LDA slightly outperforming in most cases.

To get a deeper insight, the shape-wise classification results of our proposed technique using AlexNet-LDA combination are reported as a confusion matrix in Table 7. From the confusion matrix, it is evident that almost all



	AlexNet	VGG16	VGG19	SqueezeNet	InceptionV3	GoogleNet	ResNet50	ResNet101	DenseNet201
🟦 LDA	98.33	97.41	98.33	96.67	95.74	97.22	98.33	98.33	95.93
🟧 SVM	97.04	97.22	97.39	95.74	96.85	96.48	98.15	97.59	97.78
⬜ NB	90.37	93.52	94.63	93.15	93.7	93.33	95	94.82	97.04
🟨 DT	90.17	90.37	90	87.22	81.85	85.19	83.52	88.89	84.62

Fig. 12 Overall shape classification accuracies achieved by each CNN architecture in combination with classifiers employed

Table 7 Confusion matrix of shape-wise classification results obtained by AlexNet-LDA combination

Class	1	2	3	4	5	6	7	8	A
1	60	0	0	0	0	0	0	0	0
2	1	59	0	0	0	0	0	0	0
3	0	1	58	0	1	0	0	0	0
4	0	0	0	60	0	0	0	0	0
5	0	0	1	0	59	0	0	0	0
6	0	0	0	1	0	59	0	0	0
7	0	0	0	0	0	0	60	0	0
8	0	0	0	0	0	0	2	58	0
A	0	0	0	0	0	1	1	0	58

shape classes are successfully recognized. Few instances of misclassification between BGT shape 1 and 2, 3 and 5, 7 and 8 and 7 and ‘A’ are observed, which may result due to the deformations introduced by the subjects while drawing these shapes.

5.2 Deformation classification results

In this section, we discuss the deformation classification results. It is important to mention that due to the scarcity of examples of each deformation across all BGT shapes, shape-wise comparison is not feasible. Therefore, the reported results of deformation classification are sample specific. Table 8 reports the mean classification accuracies of fivefold for each deformation using the given CNN architectures in combination with LDA classifier. Several important observations are made while analyzing the outcomes reported in Table 8.

- Contrary to their conventional use, CNN-based features can also be employed to enhance (rather than diminish) intra-class variations and inter-class similarities.
- Deformation-specific augmentation enables pre-trained CNN architectures to achieve considerable deformation classification accuracies despite a limited dataset. It is evident from the results of ‘Retrogression’ and ‘Rotation’ classification (i.e., 97.61% and 96.19%, respectively). Both these classes obtained one of the best classification accuracies, whereas both contributed the least in the original dataset, as shown in Fig. 11.
- Unlike shape recognition, where the depth or the width of a ConvNet does not have a significant impact on the classification, in deformation classification, the choice of a suitable architecture can enhance performance.
- Deeper networks (i.e., VGG19, ResNet101 and DenseNet201) appear to outperform wider networks (i.e., GoogLeNet and InceptionV3), whereas in general, ResNet101 outperforms the rest.

For a deeper insight, we further evaluate the performance of our system by computing the ‘Specificity,’ ‘Sensitivity’ and ‘Precision,’ in addition to accuracy. ‘Sensitivity’ measures the ability of the system to correctly classify the deformation and is calculated by the ratio $\frac{t_p}{t_p+f_n}$, while ‘Specificity’ measures the ability of the system to correctly classify the non-erroneous example and is defined as $\frac{t_n}{t_n+f_p}$. ‘Precision’ is the true positive relevance rate and is defined as $\frac{t_p}{t_p+f_p}$, (t_p, t_n, f_p and f_n represent the true positives, true negatives, false positives and false negatives, respectively).

Table 8 Overall deformation classification accuracies achieved by each CNN architecture in combination with LDA classifier

	CNN architecture								
Deformation	AlexNet (%)	VGG16 (%)	VGG19 (%)	SqueezeNet (%)	InceptionV3 (%)	GoogLeNet (%)	ResNet50 (%)	ResNet101 (%)	DenseNet201 (%)
Rotation	90.47	89.52	92.38	91.66	89.52	86.66	89.52	96.19	95.23
Overlap	70.83	62.5	79.16	58.33	54.16	62.5	66.66	79.16	75.0
Simplification	85.84	81.13	87.73	84.90	86.79	83.01	86.79	90.56	87.73
Fragmentation	78.50	84.11	85.04	83.17	79.43	73.83	81.30	85.98	81.37
Retrogression	94.18	96.51	96.51	95.34	94.18	95.34	94.18	97.61	95.34
Perseveration	79.16	81.25	81.25	79.16	82.05	81.25	79.16	83.33	81.25
Collision	84.25	86.11	85.18	79.62	87.96	85.18	75.92	90.74	89.81
Closure	58.06	67.74	61.29	54.83	58.06	61.29	70.96	80.64	74.19
Motor incoordination	83.33	80.55	82.40	79.62	85.18	84.54	86.11	87.96	82.40
Angulation	66.66	70.83	75.0	79.16	75.0	62.5	66.66	83.33	70.83
Cohesion	61.45	70.83	73.95	58.06	67.74	64.51	70.96	80.64	75.0

Table 9 Sensitivity, specificity and precision achieved by ResNet101-LDA combination

Deformation	Metric		
	Sensitivity (%)	Specificity (%)	Precision (%)
Rotation	90.0	96.84	90.0
Overlap	73.33	88.88	73.33
Simplification	82.60	92.77	82.60
Fragmentation	80.0	87.35	59.25
Retrogression	75.0	98.75	75.0
Perseveration	72.72	86.48	72.72
Collision	68.18	96.51	83.33
Closure	75.0	84.21	75.0
Motor Incoordination	81.25	89.13	56.52
Angulation	80.0	85.71	80.0
Cohesion	75.0	84.21	75.0

Table 9 details the sensitivity, specificity and precision values of the deformation classification module using a ResNet101-LDA combination.

It is observed that in most cases, our proposed deformation classification scheme achieves the promising results while considering the sensitivity of the system. However, in some cases, lower values of sensitivity are also obtained. For instance, in the case of ‘Collision,’ ‘Perseveration’ and ‘Overlap,’ the sensitivity of the system is below 75.0%. An interesting observation is that some of the deformation classification modules, such as ‘Collision,’ that achieved comparatively lower sensitivity values consisted of sufficient training samples in the original dataset. On the contrary, lesser training samples existed for ‘Overlap’ and required augmentation. Nevertheless, even after sufficient training, these deformation classes could not be modeled with much success. This shows that the ability of the system to identify deformations greatly depends on the challenging nature of the deformation itself. For instance, Fig. 13a shows an example of a correctly drawn BGT shape 7, while Fig. 13b, c gives examples of the BGT shape 7 drawn with Overlapping difficulty. Although Fig. 13b is scored as ‘Overlapping difficulty’ in the ground truth by trained psychologists, yet it is very similar to Fig. 13a, our system identified it as ‘Non-Erroneous,’ whereas Fig. 13c is much more deformed and therefore is correctly identified by the system as ‘Erroneous.’

As mentioned earlier, Lack’s scoring standard emphasizes the importance of the occurrence of the error rather than the frequency of the error; hence, if the system missed a deformation in one shape template, it was captured in

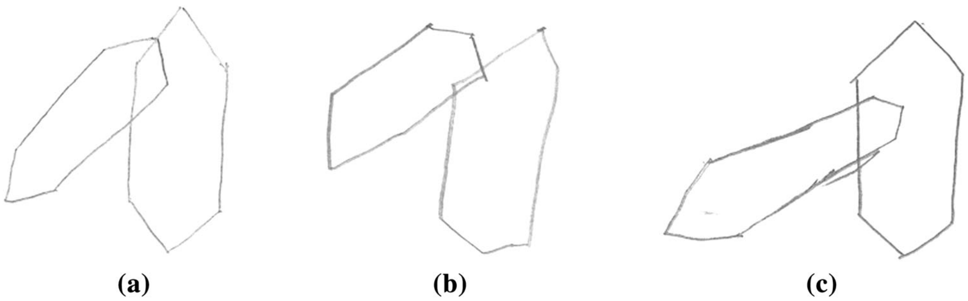


Fig. 13 Three examples of BGT shape 7 assessed by our proposed system **a** sample with no Overlapping difficulty, correctly identified as sample with no Overlapping difficulty, **b** sample with Overlapping difficulty, incorrectly identified as sample with no Overlapping difficulty, and **c** sample with Overlapping difficulty, correctly identified as sample with Overlapping difficulty

another. Therefore, the final scoring of test samples is in coherence with the ground truth.

5.3 Overall comparative analysis of proposed technique with state of the art

A comparative analysis with the state of the art gives a better insight into the effectiveness of any proposed technique. However, despite a rich and extensive literature review, it is observed that a direct feature-wise comparison cannot be deduced with any of the studies outlined. The primary reason is that there is a paradigm shift from visual-based analysis (offline) to signal-based analysis (online), due to the lack of better representation of the visual features. Due to this reason, neuropsychological tests which involve the analysis of the handwriting movement or cognition (Spiral and CDT) are mostly addressed in the literature. On the contrary, tests like BGT which provide information about the visual perceptual orientation of a subject received very little attention from the relevant pattern recognition community, despite its popularity among psychologists. In a previously conducted pilot study [6], the visual analysis of BGT shapes was performed using handcrafted geometric features. However, it was observed that it required an exhaustive rule-based approach to estimate all possible deformations across each BGT shape. Furthermore, it can still prove insufficient on a bigger sample set. As a result, a small set of BGT drawings (18 samples) were used to model only six deformations (Simplification, Overlap, Rotation, Perseveration, Closure and Cohesion) in [6]. On the contrary, in the present study, by employing deep CNN-based features, we are able to classify 11 deformations (Simplification, Fragmentation, Overlap, Rotation, Perseveration, Closure, Cohesion, Angulation, Collision, Motor incoordination and Retrogression), on a relatively bigger dataset (60 samples) with the promising results. Nonetheless, from the perspective of completeness, we attempt to compare the performance of our proposed drawing analysis system with some of the recent works in this domain. A comparison summary is outlined in Table 10. Although, a direct comparison is difficult due to different study objectives and datasets employed, nonetheless, it can be seen that our proposed scheme for the computerized analysis of neuropsychological drawings produces the comparative results.

6 Conclusion and future work

This study addresses some of the challenges of computerized analysis of neuropsychological drawings by proposing a conceptual model of a semiautonomous system. One of the key challenges addressed is the transfer of domain knowledge into computational feature space. Contrary to the conventional approaches which either focus on extensive heuristics or suggest test modifications, the proposed system employs a top-down approach, where clinician manifestations are modeled individually and offline responses of the subjects are assessed against the models to determine their presence. To effectively represent a wide variety of clinical deformations without extensive heuristics, pre-trained CNN architectures are employed. Transfer learning is enhanced by using deformation-specific augmentation. To evaluate the performance of our proposed technique, the scoring of a popular multi-template BGT drawing test is considered as a case study. A customized dataset of 60 subjects (30 healthy/30 patients) is collected and employed for evaluation purposes. Eleven clinical indicators (suggested by Lacks' scoring manual) are modeled as deformations. The test provides an opportunity to assess the performance of CNN-based features for both shape recognition and deformation classification.

By employing pre-trained ConvNets as feature extractors, we significantly reduce the computations and the time involved in training a deep CNN architecture from scratch. Pre-trained ConvNets also overcome the issue of data scarcity which is commonly observed in health-related problems like the one under consideration. To assess the

Table 10 Performance comparison with some recent works in computerized analysis of neuropsychological drawings

References	Task and samples	Study objective	Technique employed	Overall performance
Bennasar et al. [24]	648 Offline CDT drawings	Differential diagnosis of different stages of dementia	A set of geometric and spatial features are extracted and used to train cascaded classifiers	Stage three Acc. = 77.78% Stage four Acc. = 74.38%
Moetesum et al. [6] (a pilot study)	152 Offline BGT (nine shape) drawings	To classify 11 clinical deformations for Visuo-perceptual dysfunctions	Shape-specific geometric features are extracted and assessed using a heuristic-based approach	Only 6/11 deformations are classified across a small subset of shapes with accuracy ranging from 63.8 to 94.2%
Periera et al. [46]	308 Online Meander drawings	To discriminate between samples of healthy subjects and PD patients	Images generated from pen-based pressure signals are used to train a CNN classifier	Acc. = 87.14%
Harbi [26]	165 Online CDT drawings	To discriminate between samples of healthy subjects and dementia patients	Clock components (digit and hands) are recognized using CNNs, and clock drawings are assessed using ontology-based heuristics	Average Acc. = 97.3%
Mucha et al. [48]	35 Online Archimedean spiral drawings	To discriminate of healthy subjects and PD patients	Fractional derivatives between samples of healthy subjects and PD patient computed from various kinematic features are used to train Random Forest and SVM classifiers	Highest Acc. = 72.38%
Proposed technique	540 Offline BGT (nine shape) drawings	To classify 11 clinical deformations for Visuo-perceptual dysfunctions	Deformation-specific CNN-based visual features are extracted from raw images and classified using an LDA classifier	All 11/11 deformations are classified across all shapes with accuracies ranging from 79.1 to 97.6%

discriminating power of the CNN-based features, a number of popularly employed classifiers are trained independently. All experiments are conducted using fivefold cross-validation. A mean classification accuracy of **98.33%** is achieved using AlexNet-LDA, VGG19-LDA, ResNet50-LDA and ResNet101-LDA combinations for shape recognition, thus indicating that the depth of the CNN model does not have a significant impact on shape recognition.

Similarly, to assess the ability of a pre-trained CNN architecture to model deformations, an identical experimental protocol is adopted. The best results are achieved by employing the ResNet101-LDA combination for all the deformations. The mean accuracies ranged from 79.16 to 97.61%. Deformation classification is further evaluated for parameters like sensitivity, specificity and precision, and it shows that our proposed system can indeed successfully model and classify deformations, although for some deformations (Collision, Perseveration and Overlap) further improvements can be made. Overall, the findings of this study validate our preliminary hypothesis that CNN-based visual features can represent domain knowledge sufficiently without an extensive rule-based approach.

The prime objective of this study is to create a benchmark for future studies in this direction. For this reason, pre-trained CNNs are employed as off-the-shelf feature extractors. In the future extensions of this study, fine-tuning-based transfer learning will be analyzed. Transfer learning using CNN architectures other than those trained on ImageNet will also be explored. Impact of a template on capturing of deformation-specific features is another interesting direction that must be explored for better representation of deformation models. Shape-wise deformation classification can also be pursued. Observing the frequency of a particular deformation can also provide an important insight into the behavior of the subjects with potential brain dysfunctions. This can provide a very useful exploratory direction for researchers in clinical psychology. Other BGT scoring manuals can also be modeled and evaluated as a future extension of this work.

The proposed study presents a direction worth exploring for various interdisciplinary communities working to integrate technological solutions in health sector. Our proposed conceptual model can provide a solid basis for an end-to-end system for the analysis of neuropsychological drawings which can be employed by psychologists for standardization, validation of the results and other diagnostic purposes.

References

1. Mainland BJ, Shulman KI (2013) Clock drawing test. In: Larner AJ (ed) Cognitive screening instruments. Springer, New York, pp 79–109

2. Shin M-S, Park S-Y, Park S-R, Seol S-H, Kwon JS (2006) Clinical and empirical applications of the Rey–Osterrieth complex figure test. Nat Protoc 1(2):892–899
3. Bender L (1938) A visual motor gestalt test and its clinical use. Research monographs. American Orthopsychiatric Association, Chicago
4. Coates DR, Wagemans J, Sayim B (2017) Diagnosing the periphery: using the Rey–Osterrieth complex figure drawing test to characterize peripheral visual function. i-Perception 8(3):1–20
5. Price CC, Cunningham H, Coronado N, Freedland A, Cosentino S, Penney DL, Penisi A, Bowers D, Okun MS, Libon DJ (2011) Clock drawing in the montreal cognitive assessment: recommendations for dementia assessment. Dement Geriatr Cogn Disord 31(3):179–187
6. Moetesum M, Siddiqi I, Masroor U, Djeddi C (2015) Automated scoring of bender gestalt test using image analysis techniques. In: 13th international conference on document analysis and recognition (ICDAR). IEEE, pp 666–670
7. Lipsman N, Schwartz ML, Huang Y, Lee L, Sankar T, Chapman M, Hynynen K, Lozano AM (2013) MR-guided focused ultrasound thalamotomy for essential tremor: a proof-of-concept study. Lancet Neurol 12(5):462–468
8. Drotár P, Mekyska J, Rektorová I, Masarová L, Smékal Z, Faundez-Zanuy M (2015) Decision support framework for Parkinson’s disease based on novel handwriting markers. IEEE Trans Neural Syst Rehabil Eng 23(3):508–516
9. Moetesum M, Siddiqi I, Vincent N, Cloppet F (2018) Assessing visual attributes of handwriting for prediction of neurological disorders—a case study on Parkinson’s disease. Pattern Recognit Lett 121:19–27
10. Hinton GE, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
11. LeCun Y, Boser B, Denker J, Henderson D, Howard R, Hubbard W, Jackel L (1989) Handwritten digit recognition with a backpropagation network. In: Neural information processing systems (NIPS)
12. Yann Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
13. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
14. Battiato S, Farinella GM, Giudice O, Puglisi G (2016) Aligning shapes for symbol classification and retrieval. Multimed Tools Appl 75:5513–5531
15. Plamondon R, Srihari SN (2000) Online and off-line handwriting recognition: a comprehensive survey. IEEE Trans Pattern Anal Mach Intell 22(1):63–84
16. Khalid PI, Yunus J, Adnan R (2010) Extraction of dynamic features from hand drawn data for the identification of children with handwriting difficulty. Res Dev Disabil 31(1):256–262
17. Rosenblum S, Samuel M, Zlotnik S, Erikh I, Schlesinger I (2013) Handwriting as an objective tool for Parkinson’s disease diagnosis. J Neurol 260(9):2357–2361
18. Likforman-Sulem L, Esposito A, Faundez-Zanuy M, Clémençon S, Cordasco G (2017) EMOTHAW: a novel database for emotional state recognition from handwriting and drawing. IEEE Trans Hum Mach Syst 47(2):273–284
19. Naseer A, Rani M, Naz S, Razzak MI, Imran M, Xu G (2019) Refining Parkinson’s neurological disorder identification through deep transfer learning. Neural Comput Appl. https://doi.org/10.1007/s00521-019-04069-0
20. Fairhurst MC, Linnell T, Glenat S, Guest R, Heutte L, Paquet T (2008) Developing a generic approach to online automated analysis of writing and drawing tests in clinical patient profiling. Behav Res Methods 40(1):290–303
21. Smith SL, Hiller DL (1996) Image analysis of neuropsychological test responses. In: Medical imaging 1996: image processing. International Society for Optics and Photonics, vol 2710, pp 904–916
22. Canham R, Smith S, Tyrrell A (2000) Automated scoring of a neuropsychological test: the Rey–Osterrieth complex figure. In: Proceedings of the 26th Euromicro conference. IEEE, vol 2, pp 406–413
23. Canham R, Smith S, Tyrrell A (2005) Location of structural sections from within a highly distorted complex line drawing. IEE Proc Vis Image Signal Process 152(6):741–749
24. Bennasar M, Setchi R, Hicks Y, Bayer A (2014) Cascade classification for diagnosing dementia. In: IEEE international conference on systems, man and cybernetics (SMC). IEEE, pp 2535–2540
25. Harbi Z, Hicks Y, Setchi R (2016) Clock drawing test digit recognition using static and dynamic features. Procedia Comput Sci 96:1221–1230
26. Harbi Z, Hicks Y, Setchi R (2017) Clock drawing test interpretation system. Procedia Comput Sci 112:1641–1650
27. Pullman SL (1998) Spiral analysis: a new technique for measuring tremor with a digitizing tablet. Mov Disord 13(S3):85–89
28. Pereira CR, Pereira DR, da Silva FA, Hook C, Weber SA, Pereira LA, Papa JP (2015) A step towards the automated diagnosis of parkinson’s disease: analyzing handwriting movements. In: IEEE 28th international symposium on computer-based medical systems (CBMS). IEEE, pp 171–176
29. Lacks P (1999) Bender Gestalt screening for brain dysfunction. Wiley, New York
30. Glenat S, Heutte L, Paquet T, Guest R, Fairhurst M, Linnell T (2008) The development of a computer-assisted tool for the assessment of neuropsychological drawing tasks. Int J Inf Technol Decis Mak 7(04):751–767
31. Song Y, Davis R, Ma K, Penney DL (2016) Balancing appearance and context in sketch interpretation. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence. AAAI Press, pp 2625–2632
32. Smith SL, Cervantes BR (1998) Dynamic feature analysis of vector-based images for neuropsychological testing. In: Medical imaging 1998: physiology and function from multidimensional images. International Society for Optics and Photonics, vol 3337, pp 304–314
33. Garbi A, Smith SL, Heseltine D, Thomson P (1999) Automated and enhanced assessment of unilateral visual neglect. In: IET conference proceedings, pp 382–386(4)
34. Kornmeier J, Bach M (2005) The Necker cube—an ambiguous figure disambiguated in early visual processing. Vis Res 45(8):955–960
35. Smith SL, Lones MA (2009) Implicit context representation Cartesian genetic programming for the assessment of visuo-spatial ability. In: IEEE congress on evolutionary computation, pp 1072–1078
36. Heinik J, Werner P, Dekel T, Gurevitz I, Rosenblum S (2010) Computerized kinematic analysis of the clock drawing task in elderly people with mild major depressive disorder: an exploratory study. Int Psychogeriatr 22(3):479–488
37. Werner P, Rosenblum S, Bar-On G, Heinik J, Korczyn A (2006) Handwriting process variables discriminating mild Alzheimer’s disease and mild cognitive impairment. J Gerontol Ser B Psychol Sci Soc Sci 61(4):P228–P236
38. Hazell A, Smith SL (2008) Towards an objective assessment of alzheimer’s disease: the application of a novel evolutionary algorithm in the analysis of figure copying tasks. In: Proceedings of the 10th annual conference companion on Genetic and evolutionary computation. ACM, pp 2073–2080

39. 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 disease and healthy controls. Curr Alzheimer Res 14(9):960–968
40. El Yacoubi M, Garcia-Salicetti S, Kahindo C, Rigaud A-S, Cristancho-Lacroix V (2018) From aging to early-stage Alzheimer's: Uncovering handwriting multimodal behaviors by semi-supervised learning and sequential representation learning. Pattern Recognit 86:08
41. Plamondon R, O'Reilly C, Ouellet-Plamondon C (2014) Strokes against stroke—strokes for strides. Pattern Recognit 47(3):929–944
42. O'Reilly C, Plamondon R, Lebrun L-H (2014) Linking brain stroke risk factors to human movement features for the development of preventive tools. Front Aging Neurosci 6:150
43. Drotár P, Mekyska J, Rektorová I, Masarová L, Smékal Z, Faundez-Zanuy M (2013) A new modality for quantitative evaluation of parkinson's disease: in-air movement. In: 13th IEEE international conference on bioinformatics and bioengineering (BIBE). IEEE, pp 1–4
44. Drotár P, Mekyska J, Rektorová I, Masarová L, Smékal Z, Faundez-Zanuy M (2014) Analysis of in-air movement in handwriting: a novel marker for Parkinson's disease. Comput Methods Programs Biomed 117(3):405–411
45. Drotár P, Mekyska J, Rektorová I, Masarová L, Smékal Z, Faundez-Zanuy M (2016) Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease. Artif Intell Med 67:39–46
46. Pereira CR, Weber SA, Hook C, Rosa GH, Papa JP (2016) Deep learning-aided Parkinson. In: 29th SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). IEEE, pp 340–346
47. Taleb C, Khachab M, Mokbel C, Likforman-Sulem L (2017) Feature selection for an improved Parkinson's disease identification based on handwriting. In: 1st international workshop on Arabic script analysis and recognition (ASAR). IEEE, pp 52–56
48. Mucha J, Zvoncak V, Galaz Z, Faundez-Zanuy M, Mekyska J, Kiska T, Smekal Z, Brabenec L, Rektorova I, Lopez-de Ipina K (2018) Fractional derivatives of online handwriting: a new approach of parkinsonic dysgraphia analysis. In: 41st international conference on telecommunications and signal processing (TSP). IEEE, pp 1–4
49. Rémi C, Frélicot C, Courtellemont P (2002) Automatic analysis of the structuring of children's drawings and writing. Pattern Recognit 35(5):1059–1069
50. Chindaro S, Guest R, Fairhurst M, Potter J (2004) Assessing visuo-spatial neglect through feature selection from shape drawing performance and sequence analysis. Int J Pattern Recognit Artif Intell 18(07):1253–1266
51. Renau-Ferrer N, Rémi C (2010) A method for visuo-spatial classification of freehand shapes freely sketched. In: 14th international conference on image processing, computer vision, & pattern recognition, p 216
52. Beuvens F, Vanderdonckt J (may 2012) Usigesture: An environment for integrating pen-based interaction in user interface development. In: 6th International conference on research challenges in information science (RCIS), pp 1–12
53. De Pandis MF, Galli M, Vimercati S, Cimolin V, De Angelis MV, Albertini G (2010) A new approach for the quantitative evaluation of the clock drawing test: preliminary results on subjects with Parkinson's disease. Neurol Res Int 2010:283890
54. Galli M, Vimercati SL, Stella G, Caiazzo G, Norveti F, Onnis F, Rigoldi C, Albertini G (2011) A new approach for the quantitative evaluation of drawings in children with learning disabilities. Res Dev Disabil 32(3):1004–1010
55. Vimercati S, Galli M, De Pandis M, Ancillao A, Stella G, Albertini G (2012) Quantitative evaluation of graphic gesture in subjects with Parkinson's disease and in children with learning disabilities. Gait Posture 35:S23–S24
56. Tabatabaey-Mashadi N, Sudirman R, Guest RM, Khalid PI (2013) An evaluation of children's structural drawing strategies. J Teknol [Sci Eng] 61(2):27–32
57. Tabatabaey-Mashadi N, Sudirman R, Guest RM, Khalid PI (2015) Analyses of pupils' polygonal shape drawing strategy with respect to handwriting performance. Pattern Anal Appl 18(3):571–586
58. Khalid PI, Yunus J, Adnan R, Harun M, Sudirman R, Mahmood NH (2010) The use of graphic rules in grade one to help identify children at risk of handwriting difficulties. Res Dev Disabil 31(6):1685–1693
59. Nair V, Hinton GE (2010) Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp 807–814
60. Szarvas M, Yoshizawa A, Yamamoto M, Ogata J (2005) Pedestrian detection with convolutional neural networks. In: Proceedings of the IEEE intelligent vehicles symposium. IEEE, pp 224–229
61. Mori K, Matsugu M, Suzuki T (2005) Face recognition using SVM fed with intermediate output of CNN for face detection. In: MVA, pp 410–413
62. Lauer F, Suen CY, Bloch G (2007) A trainable feature extractor for handwritten digit recognition. Pattern Recognit 40(6):1816–1824
63. Sheikhi S (2007) Clinical use of bender-gestalt test in brain lesions diagnosis and its comparison with magnetic resonance imaging (MRI). J Urmia Nurs Midwifery Fac 5(1):15–21
64. Santos AA, Jorge LM (2007) Bender test with dyslexics: comparison of two systems of punctuation. Psico-USF 12(1):13–21
65. Allen RA, Decker SL (2008) Utility of the bender visual-motor gestalt test second edition in the assessment of attention-deficit/hyperactivity disorder 1. Percept Motor Skills 107(3):663–675
66. Ferreira RB, Feil CF, Nunes MLT (2009) Bender visual-motor gestalt test in the children's clinical assessment. Psico-USF 14(2):185–192
67. Keppeke L, Cintra I, Schoen TH (2013) Bender visual-motor gestalt test in adolescents: relationship between visual-motor development and the tanner stages 1. Percept Motor Skills 117(1):257–275
68. Koppitz EM (1964) The bender gestalt test for young children. Grune & Stratton, New York
69. Cavanaugh JJ (1973) Preventing reading failure: prediction, diagnosis, intervention. Pediatrics 52(1):157–158
70. Hutt ML, Monheit S (1985) Effectiveness of the hutt adaptation of the bender-gestalt test configuration scale in differentiating emotionally disturbed adolescents. Psychol Rep 56(2):439–443
71. Brannigan GG, Decker SL, Madsen DH (2004) Innovative features of the Bender–Gestalt Ii and expanded guidelines for the use of the global scoring system. Bender visual-motor Gestalt test, second edition assessment service bulletin, 1
72. Moetesum M, Siddiqi I, Masroor U, Vincent N, Cloppet F (2016) Segmentation and classification of offline hand drawn images for the BGT neuropsychological screening test. In: 8th international conference on digital image processing (ICDIP 2016). International Society for Optics and Photonics
73. Nazar H, Moetesum M, Ehsan S, Siddiqi I, Khurshid K, Vincent N, McDonald-Maier KD (2017) Classification of graphomotor impressions using convolutional neural networks—an application to automated neuropsychological screening tests. In: 14th international conference on document analysis and recognition (ICDAR)

74. Niu X-X, Suen CY (2012) A novel hybrid CNN-SVM classifier for recognizing handwritten digits. Pattern Recognit 45(4):1318–1325
75. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6(1):60
76. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
77. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 248–255
78. Miles TR (1997) The Bangor dyslexia test. Learning Development Aids, Wisbech
79. Wilkinson GS, Robertson GJ (2006) Wide range achievement test (WRAT4). Psychological Assessment Resources, Lutz, FL
80. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
81. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K (2016) Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5 mb model size. arXiv preprint arXiv:1602.07360
82. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1–9
83. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2818–2826
84. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778
85. Iandola F, Moskewicz M, Karayev S, Girshick R, Darrell T, Keutzer K (2014) DenseNet: implementing efficient ConvNet descriptor pyramids. arXiv preprint arXiv:1404.1869
86. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? In: Advances in neural information processing systems, pp 3320–3328
87. Agrawal P, Carreira J, Malik J (2015) Learning to see by moving. In: Proceedings of the IEEE international conference on computer vision, pp 37–45

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.