Using deep learning predictions to study the development of drawing behaviour in children

Benjamin Beltzung, Marie Pelé, Lison Martinet, Elliot Maître, Jimmy Falck, Cedric Sueur

Using deep learning predictions to study the development of drawing behaviour in children

Benjamin Beltzung[1], Marie Pelé[2], Lison Martinet[1], Jimmy Falck[3], Elliot Maitre[4], Cédric Sueur[1,5]

1 Université de Strasbourg, CNRS, IPHC UMR 7178, Strasbourg, France

2 ANTHROPO-LAB – ETHICS EA 7446, Université Catholique de Lille, F-59000 Lille, France

3 Laboratoire lorrain de recherche en informatique et ses applications (Loria – CNRS/Université de Lorraine/Inria), Nancy, France

4 Université de Toulouse - IRIT UMR5505, 31400, Toulouse, France

5 Institut Universitaire de France, Paris, France

Corresponding author : Cédric Sueur, cedric.sueur@iphc.cnrs.fr; +33388107453; IPHC UMR 7178, 23 rue Becquerel 67087 Strasbourg, France

Abstract: Drawing behaviour in children provides a unique window into their cognitive development. This study uses Convolutional Neural Networks (CNNs) to examine cognitive development in children's drawing behavior by analyzing 386 drawings from 193 participants, comprising 150 children aged 2 to 10 years and 43 adults from France. CNN models, enhanced by Bayesian optimization, were trained to categorize drawings into ten age groups and to compare children's drawings with adults'. Results showed that model accuracy increases with the child's age, reflecting improvement in drawing skills. Techniques like Grad-CAM and Captum offered insights into key features recognized by CNNs, illustrating the potential of deep learning in evaluating developmental milestones, with significant implications for educational psychology and developmental diagnostics.

Keywords: Artificial Intelligence, Cognitive Development, Explicability, Interpretability, Drawing Behaviour

Introduction

In children, drawing behaviour appears around the age of 18 months-old. During a lifetime, drawing is an important mode of communication driven not only by cognitive aspects but also by cultural ones. For instance, by utilising human figure representations in drawings, [1] demonstrated variations in body size and shape among young adults from Israel and Thailand when asked to depict themselves. Similarly, through analysing drawings created by children from different cultures, Restoy and colleagues [2] showcased how the level of individualism in countries could influence the size and number of human figures depicted. Of course, culture is not the sole factor influencing drawing behaviour and numerous cognitive aspects have been explored through it. Luquet [3] firstly suggested that the development of the drawing behaviour can be seen as a four-stage process. The first stage occurs when the child lacks intention to represent reality and discovers by chance a shape analogy between an object and its initially nonsignificant trace (*fortuitous realism*). Around four or five years old, the child tries to produce realistic drawings but does not have the entire abilities (motor and representative skills) to do so (*missed realism*). Then, the child uses her knowledge of objects' components to represent them but some misconceptions linked for example to transparency or perspective representation are still present (*intellectual realism*). Eventually, the last stage proposed by Luquet [3] is *visual realism*, when every step is completed.

Since, many other developmental theories on drawing behaviour have been proposed. For example, Adi-Japha and colleagues [4] proposed three different steps in the development of drawing behaviour in children. First, *action representation* occurs when the drawing is associated with verbalisation. For example, when the child represents a moving object like a car or a train and produces the corresponding sound. *Romancing* occurs when the child is able to name her drawing but it remains challenging for another individual to interpret it. Then, the

guided elicitation phase occurs when the child is able to produce a figurative drawing helped by an adult. Thus, Adi-Japha and collaborators [4] focused on *how* the drawing is produced better than the drawing as a result. On the other hand, Baldy [5] proposed a classification only based on the morphological development of the human figures in the drawing (e.g. tadpoles, filiform figures, tube-shaped, etc.). Doing so, Baldy's classification appears more focused on the drawing as a product and then, is close to Luquet's classification.

Even if all these developmental theories are relevant and not intrinsically exclusive, they consider different concepts and focused on different aspects of the drawing behaviour: its process and its result. Without forgetting that it may be difficult to visually interpret the corresponding stage for a given drawing. Traditionally, drawings are analysed by defining and extracting a set of features such as the number and size of figures or the number of used colours, etc. While this approach can be insightful, two challenges arise. First, the amount of information contained in a drawing is substantial in nature, and using predefined features significantly limits the amount of extracted information. Then, as each of these feature focuses on a single aspect of the drawing, this approach does not consider the holistic aspect of drawings (i.e. considering drawing as a whole, and not only its different parts). Moreover, studies in toddlers showed that their drawings may have some meanings even if they do not have figurative aspects on it [6,7]. For these reasons, such methods may not be sufficient to benefit from the information contained in a drawing to its fullest extent. A possible way to mitigate these issues is to ask children about what they intended to represent. However, they may not be directly conscious about the deep meaning of their drawing, and this could not be applied to scribbles drawn by very young children who are not able to verbally communicate yet or children with pathologies that make them unable to communicate.

To minimise these biases, a potential candidate is the use of artificial intelligence and more precisely deep learning. Over the past decade, important advances in deep learning models

(neural networks) have been made when considering images, video or audio processing, substantially improving the predictive accuracy and outperforming state-of-the-art methods in many fields, such as system health management [8], face recognition [9], or even speech recognition [10]. The most popular type of deep learning models is Convolutional Neural Networks (CNNs) [11], which is known to provide a high accuracy for tasks involving image analyses. While CNNs architecture can vary according to the task, some key concepts remain. Different types of layers exist in CNN and play different roles. In convolutional layers, a convolution is applied on the image: a filter, representing a feature, slides over the image, and results in a feature map. Each value of the feature map is the degree of activation of the filter on the corresponding part of the image. Depending on the depth of the convolutional layers, filters can detect either low-level features (e.g. lines or curves) or high-level features (e.g. objects).

While such models usually provide a high accuracy, they are widely considered as black boxes with regard to the decision-making process [12,13], as it is not possible to straightly understand the process that led the model to predict a particular output. Indeed, the complexity and number of parameters that can attain billion, make such models difficult to interpret. However, from this complexity also comes a strength. Deep learning models allow for analysing every pixel of a given image, extracting a large amount of information contained in it, and can therefore potentially grasp all the relevant features. The features learning produced by neural networks are also complex, and allow for an objective feature representation.

Although deep learning has already been successfully used to analyse drawings [14–17], even in children [18,19]; the development of the drawing behaviour has been, to our knowledge, only poorly analysed through the lens of deep learning [20–23].

In this paper, we first build and train a CNN by using Bayesian optimisation to classify drawings according to the age of the individuals (i.e. 10 age categories, from children to adults).

By using the same method, we then trained multiple models to classify children subcategories versus adults' drawings to compute the accuracy and predict drawings not belonging to the classes considered in the models. We hypothesised that the accuracy of the models should increase with the age difference between children and adults. Indeed, as children grow, their drawing skills improve and their drawings become closer to what an adult could produce.

Material and methods

a. Dataset

The data consist of 386 drawings produced by 150 children (from two to ten years-old) and 43 adults (novice and experts) (detail in Table S1). All the participants were given a touch-screen displaying a white background and could choose among a panel of ten colours by having access to an overlay on the bottom of the screen. Examples of drawings are shown in Figure 1. From three years-old to adults, the drawings were collected using the following protocol. Each participant was asked to produce two drawings, one under a free condition, where the participant was not given a particular instruction, and one under a self-portrait condition, where the drawer was asked to draw himself. Two categories of adults were defined prior to the data collection. First, a group of adults who had never taken drawing classes and did not have drawing as a hobby, that will be here considered as novices. Then, an expert category, including art school students and professional illustrators. The data distribution is presented in Table S1. For more information about these datasets, please refer to [24].

As two-year-old participants were not verbally able to communicate and to understand the instructions, no particular task was given for these productions. For this reason, they will be considered as produced in free condition. In this category, 6 subjects participated and produced a total number of 30 drawings, ranging from 3 to 8 drawings per individual (respectively 3, 4, 4, 5, 6, 8 drawings).

All drawings were of dimensions 2732 × 2048, and were resized to a 224 × 224 square with 3 channels for the colours to conduct analyses.

b. Ethics

We ensured the confidentiality of drawings collected from human participants, adhering strictly to the ethical guidelines of our research institutions. The study received approval from the Strasbourg University Research Ethics Committee (Unistra/CER/2019-11). Informed consent was secured from all adult participants and from a parent or legal guardian for minors. Additionally, consent for the publication of any identifying images in an online open-access format was obtained, safeguarding participant privacy and adhering to ethical standards for research.

c. Transfer learning from deep learning

To analyse the development of the drawing behaviour through the age, we compared drawings from different age categories. To do so, we used deep learning, and more precisely transfer learning. Transfer learning is a method used in machine learning and consists in using the knowledge of an already trained model for another task [25]. This technique is particularly useful for small datasets, which is the case here. For this reason, we also used data augmentation, more particularly horizontal flips, as this transformation does not distort the image and the result remains realistic. For this study, we used the architecture of VGG19 [26] pretrained on the ImageNet dataset [27], as VGG models have already been widely used for drawing analyses [20,28]. VGG19 is a CNN consisting of 16 convolutional layers and 3 fully connected layers. The last fully connected layer was removed, as the ImageNet classes are not of interest in this study.

By using VGG19 architecture and ImageNet weights, we first trained a model with 10 classes (i.e. the age categories). The architecture of all models is described below. Then, to refine these analyses, we considered, for children only, new age categories by grouping each class with the following one. For example, drawings produced by 3 and 4 years-old are gathered in a new category called '3–4 years-old', drawings produced by 4 and 5 years-old will be in a new category called '4–5 years-old', thus approximately doubling the number of drawings per category. This process was done up to and including 10 years-old. Novices and experts' drawings were independently grouped in a new category simply called 'adults', in order to compare drawings from children to those of adults. To do so, we trained 7 models, each classifying drawing from a following pair of children's categories against the adults' drawings. We did exactly the same protocol with ResNet18 model [29]. ResNet18 is a model from the Residual Network (ResNet) family consisting of 18 layers. We obtained results with ResNet18 similar to the ones of VGG19. Codes and scripts are available on Github: https://github.com/cedricsueur/drawinganalyses.

d. Bayesian optimization

For the training, we independently built models and tuned hyperparameters. Multiple strategies exist to find the best architecture and hyperparameters. For example, grid search consists in defining a set of vectors for each hyperparameter and training a model for every possible combination. Another possibility is a random search, taking into account the fact that hyperparameters may not all have the same impact.

Here, we used Bayesian optimisation, a technique using the information of past evaluations to iteratively find the best combination from a given parameters space. The model is first trained with a set of parameters subjectively defined by the user. This method then allows for using prior knowledge (i.e. the knowledge of the past trials) to select a new set of parameters that are

expected to improve the accuracy. The number of iterations of this process is defined beforehand.

The architecture of our model is based on VGG19, which takes an image with a (224,224,3) shape as input. Before flattening the result of the last convolution, we considered 3 pooling options through the Bayesian optimisation process: average pooling, max pooling, or no pooling. After flattening, we considered a dropout layer through Bayesian optimisation, with a value between 0 and 0.4 and a step (i.e. the smallest meaningful distance between two values) of 0.1, potentially followed by a Batch Normalisation layer. Then, from 1 to 3 fully connected layers are added, each containing from 32 to 512 units with a step of 32, each followed by a relu activation. The last fully connected layer is eventually followed by the classification layer using sigmoid function and binary crossentropy loss. The learning rate is also optimised between 1^{e-5} and 1^{e-1} through log sampling with stochastic gradient descent (SGD) optimiser. The tuner is running on 50 trials (i.e. testing 50 different combinations of parameters) to minimise the validation loss. For each trial, the model is trained on 15 epochs, with a batch size varying from 16 to 64 with a step of 16.80% of the data are used for the training and 20% for validation, with an early stopping if the validation accuracy did not improve for the last 3 epochs. Once all the trials have been evaluated, the optimal parameters are saved. The best architecture with the best hyperparameters is then trained 10 times to compute the validation accuracy and finally the mean validation accuracy for each model. To assess the relative importance of colours, the same procedure was conducted after converting the images into grayscale.

e. Predictions matrices

To assess the similarity across different age groups in our study, we employed a methodical approach by generating prediction matrices for each model. This process involves calculating

the average prediction values for age categories that do not directly correspond to the predefined classes of the model. Specifically, we utilise the sigmoid function in each model to generate a probability score between 0 and 1, where drawings by children are systematically labelled as 0 and those by adults as 1. We adopt a threshold of 0.5 to differentiate between these categories: a prediction score above 0.5 indicates that the model categorises the drawing as an adult's work, whereas a score below 0.5 suggests it belongs to a child. This threshold-based approach allows us to gauge the model's confidence in its predictions. For instance, in a model differentiating between drawings by 3-4-year-olds and adults, a score nearing 0 denotes high confidence in identifying the drawing as child-produced (specifically, by the 3–4 years age group). Conversely, a score approaching 0.5 signifies uncertainty in classification, indicating the model's difficulty in distinguishing between the age groups. While it is common to evaluate model accuracy by examining predictions within the model's designated classes, analysing predictions for drawings outside these classes offers additional insights. For example, analysing how a model, trained to differentiate between 3-4-year-olds and adults, predicts the age category of a drawing made by a 10-year-old child can provide valuable information. Such an analysis not only helps in understanding the model's perceptual boundaries between age groups but also offers a quantitative perspective on its classification behaviour across a broader spectrum of ages, thereby enriching our understanding of the model's interpretive capabilities.

f. Explicability

In our method, we also employed Captum, a model interpretability library for PyTorch, alongside Grad-CAM for explicability purposes. Deep learning models are known to be very efficient for image classification, however, most models remain as black boxes, and disentangling features which played a role in the classification remains a challenging task. This complexity arises from the difficulty in interpreting such models, as highlighted in recent

studies [15]. In the present case, it is of interest to understand which features were discriminative. However, given that a significant proportion of drawings are nonfigurative, directly answering this question proves to be challenging. Instead, a viable approach involves examining the regions of the images that played an important role in the classification.

To this end, Grad-CAM [30] offers a powerful method. For any given image, Grad-CAM generates a heatmap that highlights the regions important for a specific class. This becomes particularly insightful when applied to the predicted class to discern why the model categorised the input image as belonging to this class. The algorithm computes the gradient using the activation of the last convolutional layer, which captures high-level features, for a given class.

Integrating Captum [31] into this workflow enhances the explicability further by providing a comprehensive toolkit for model interpretability. Captum supports various interpretability algorithms, including Grad-CAM, allowing researchers to not only visualise important regions but also understands the attribution of each input features to the model's output. By applying Captum's Grad-CAM visualisations on the validation data for every model, we can gain deeper insights into the discriminative features recognised by the models. This integration facilitates a more nuanced understanding of model predictions, particularly in complex cases where direct interpretation of features is not straightforward.

Results

a. 10-classes model

The optimal model for classifying drawings across ten age categories achieved the accuracy of 40%, as illustrated in the confusion matrix presented in Figure 2. This performance significantly surpasses the 10% accuracy expected from a model making predictions at random. Although the presence of a diagonal in the confusion matrix indicates correct classifications, a considerable number of drawings were incorrectly predicted by the model to belong to ages

other than their true categories. To enhance our understanding of the model's performance and address its limitations, we delved deeper into the analysis of the model. This involved examining the patterns and characteristics of the misclassifications to identify potential areas for improvement and gain insights into the model's decision-making process.

To gain deeper insights into the distinctions between our age groups, we trained models to differentiate between drawings made by adults and those made by children of various age categories. The selection of the optimal models was facilitated through Bayesian optimisation, with the chosen parameters and their mean accuracy detailed in Table S2. This mean accuracy was derived from the validation accuracy of the optimal model, which was trained ten times. The performance of models trained on grayscale drawings is specifically outlined in Table S3. An interesting trend observed is the decrease in model accuracy with increasing age for both RGB and grayscale models, as depicted in Figure 3. Notably, the grayscale models generally exhibit higher mean accuracy than their RGB counterparts, with the exception of the model for the 3–4 years old category. The models exhibit a strong capability to distinguish between the drawings of 2-3-year-old children and those of adults. However, this differentiation accuracy diminishes progressively as the age increases, eventually stabilising at a plateau for the drawings produced by children in the 5-7-year-old age group. This suggests that while the models are highly effective at identifying the distinctive characteristics of very young children's drawings compared to those of adults, the differences become less pronounced or harder to detect in the artwork of older children, particularly those aged 5 to 7 years. This plateau indicates a point where the models no longer significantly improve in distinguishing between the drawings of children in this age range and adults, reflecting a nuanced challenge in capturing the gradual development of drawing skills as children grow older.

b. Models by pair

Further analysis involved predicting the age categories between children and adults to compute the mean prediction for each model and class. The outcomes of these predictions are illustrated in Figure 4 for both RGB and grayscale models, along with an example interpretation of the matrices. To remind, we employed probability score ranging from 0 to 1, where a score of 0 represents approximation to drawings made by younger children and a score of 1 corresponds to drawings more similar to those made by adults. On the other hand, a score near the midpoint of 0.5 reveals a level of ambiguity, showing the model's challenge in clearly separating the artworks by these age groups. In the prediction matrices for RGB images, the first row illustrates that all mean predictions surpass the 0.5 threshold, indicating that drawings made by children aged from 4 to 10 years more closely resemble those made by adults than those of the 2-3-year-old category. Notably, in this first row and subsequent ones, the mean prediction values for each specific age group rise in conjunction with the age of the predicted category, from younger children to adults, offering a quantitative measure of the development of drawing skills. For example, the mean prediction value for 4-year-olds is 0.67, suggesting their drawings are more adult-like than those of 2-3-year-olds, as 0.67 significantly exceeds 0.5. Conversely, in the model for 3-4-year-olds, the prediction for 5-year-olds is 0.54, barely distinguishable from a random classification since this value hovers near 0.5. Focusing on the transition from drawings by 2-3-year-olds to those by adults, stabilisation in mean prediction values is observed from ages 7 to 10 years, as opposed to the notable increase observed from 4 to 7 years. This might indicate a plateau in the evolution of drawing skills within this age range. For models analysing children from 4 to 5 years old and older, the prediction values approximate 0.5, highlighting challenges in differentiating drawings, which may suggest a convergence in drawing styles. The model focusing on 5-7-year-olds indicates stabilisation, with mean prediction values approaching 0.5. This suggests that drawings by children aged 8 to 10 years are perceived as similarly adult-like to those by 5-7-year-olds. In the final two

models, which compare children's drawings to those of 7–8 and 8-9-year-olds, predictions fall below 0.5 for the 8–9 and 9-10-year-old categories, potentially indicating a slowdown in the development of drawing behaviour from ages 7 to 10. Based on these observations of prediction value increases, stabilisation, and their positions relative to the 0.5 uncertainty threshold, we can delineate three stages in drawing: 2 to 4 years old, about 4 to 7 years old, and above 7 years old.

c. Interpretability models

Grad-CAM (Gradient-weighted Class Activation Mapping) and Captum, tools for model interpretability, were specifically applied to the validation datasets to investigate what features the VGG19 and ResNet18 convolutional neural networks (CNNs) identify and utilise to characterise and classify drawings. The interpretability of the Grad-CAM outputs varies significantly across different models and is also influenced by the representativeness of the drawings under examination, as illustrated in Figure 5 and Figure 6.

This variability is particularly evident when comparing the analysis of figurative drawings to that of non-figurative ones. In the case of figurative art, it is relatively straightforward to discern what the model recognises as key features, such as faces or eyes. However, for non-figurative drawings, which lack clear representational content, it becomes challenging to understand what aspects of the drawing are being recognised and how these contribute to the classification decision. Furthermore, even when specific elements like rainbows or flowers are identified within a drawing, it remains unclear how the CNNs interpret these elements in the context of classifying the drawings.

A manual evaluation of the interpretability models, specifically analysing their performance through Grad-CAM heatmaps and Captum pixel attributions, revealed that 72% of human faces in drawings were accurately recognised as faces. This indicates that the models

tend to focus more intensely on the areas depicting faces in drawings that contain them, as demonstrated by more pronounced heatmaps or pixel attributions in these regions. However, this assessment is inherently subjective, reflecting the inherent challenges in quantifying model interpretability. The models encountered difficulties in accurately identifying faces under certain conditions: when multiple faces appear within a single drawing, when faces are intertwined or merged with other lines or shapes, or when the drawings include animal faces. These challenges likely stem from the training dataset's composition, predominantly featuring self-portraits that contain a single, clearly delineated human face.

Discussion

Drawings have long been recognised as a reflective mirror to the inner workings of the human mind, especially in children. Yet, the interpretation of these drawings by adults, whether experts or not, may not always align with the child's original intent. This discrepancy underscores the potential for bias, particularly when interpretations rely on subjective judgment rather than objective or mathematical definitions [24,32,33]. In response to this challenge, we advocate for the adoption of novel and objective methodologies capable of decoding the rich tapestry of information encapsulated within drawings [15]. Our research harnesses the power of deep learning to navigate the complex landscape of children's and adults' drawings, offering new insights into the evolution of drawing behaviour across different ages.

The initial model, tasked with categorising drawings into ten distinct age groups, achieved 40% accuracy, far surpassing the 10% expected from random guesswork. However, the preponderance of drawings classified by mistake as belonging to 7-year-olds suggests an anomaly likely attributed to the model's internal 'black box' mechanics rather than to any tangible psychological or developmental rationale. By adopting a paired comparison approach, we found that models distinguishing between the youngest children (2 and 3 years old) and

adults were particularly effective, achieving accuracy rates exceeding 85%. This accuracy diminishes with age, plateauing around 60% for models comparing 7 and 8-year-olds to adults– a trend that intuitively mirrors the increasing sophistication of children's drawings at this age. Intriguingly, converting images to grayscale improved accuracy across almost all models, with the notable exception of the 3-4-year-old category, where colour appears to play a pivotal role in the expressiveness of drawings, possibly serving as a tool for exploration rather than representation [24]. Our analysis further may identify three potential stages within the prediction matrices, representing the trajectory of drawing behaviour. This stage not only corroborate previous findings [4,7,34] but also may offer a potential framework for understanding the progression from scribbles to more sophisticated artistic expressions, aligning with established theories like those proposed by Luquet [3].

The interpretability of deep learning models [35,36], especially when examined through the lenses of Grad-CAM and Captum, presents a complex landscape that is as varied as it is intriguing. These tools, designed to provide a window into the 'thought processes' of neural networks, generate heatmaps or pixel graphs that can sometimes clearly demarcate the features deemed important by the model, such as human faces in drawings. This capability is remarkable, suggesting that, to some extent, models are capable of 'seeing' and prioritising elements in images that humans also find significant. This endeavour could be significantly advanced by increasing the sample size and refining the models' ability to identify specific facial features such as eyes, nose, mouth, etc. [2,5,37]. Enhancing the granularity with which the models recognise and interpret these elements could lead to a deeper understanding of the nuanced ways in which neural networks process visual information, offering more detailed insights into their interpretive capabilities. However, the clarity and utility of these heatmaps are not uniform across all types of drawings. When these interpretability tools are applied to non-figurative drawings that do not directly represent visible objects or scenes–their output

often becomes cryptic. This enigmatic nature of the heatmaps in such contexts highlights a fundamental challenge in artificial intelligence: understanding how deep learning models process and interpret abstract visual content. Unlike figurative productions, where the presence of recognisable shapes and forms can guide the interpretation of heatmaps, nonfigurative ones lack these anchors, making the model's focus and decision-making process harder to decipher.

This variability in interpretability underscores a broader issue within the field of AI– despite the advanced capabilities of neural networks, their decision-making processes can sometimes be as opaque as they are sophisticated [13,38–40]. The challenge lies not only in achieving high accuracy in tasks such as drawing classification but also in making these processes transparent and understandable. This is particularly crucial when AI is used in domains where understanding the 'why' behind decisions is as important as the decisions themselves. Moreover, even with we got respectable results to classify drawings according to age, the discrepancy in heatmap's clarity between figurative and non-figurative drawings raises questions about the training of these models. Neural networks learn to prioritise certain features over others based on the datasets on which they are trained. If these datasets are predominantly composed of figurative images, the models may develop a bias towards recognising and interpreting features that are present in such images, at the expense of understanding more abstract, nonfigurative content. This suggests that diversifying training datasets to include a broader range of artistic expressions could be key to enhancing model interpretability across a wider spectrum of drawings.

Furthermore, the interpretability challenge also points to the need for developing more sophisticated tools and techniques that can provide deeper insights into the workings of neural networks. As the field of AI continues to evolve, the quest for models that are not only accurate but also interpretable will likely remain a central theme, driving advancements in technology and methodology. Moreover, factors such as the emotional state and motivation of the child, as

well as the conditions under which the drawings were produced, can introduce variability into our analysis. Despite these challenges, ensuring uniform conditions across all age groups helps mitigate potential biases, suggesting that future research protocols could benefit from a more controlled drawing environment. To further refine our models and enhance their predictive accuracy, we propose expanding the scope of our trials and exploring alternative architectural frameworks. Additionally, comparing machine-generated classifications with human judgments could provide valuable insights into the interpretability and applicability of these models in real-world contexts.

In conclusion, our research highlights how deep learning models can classify even with some difficulties drawings across age groups. Work still has to be done but this is an important methodological step in our understanding of drawing behaviour. Indeed, AI and more particularly deep learning can now be considered as a new tool in our pre-existing drawing comprehension 'tool box' including other devices as fractals, PCA, etc. [15]. This allows new perspectives of interdisciplinary work and underscores the potential of deep learning to uncover the subtle nuances of human expression and perception. Indeed, such approach could revolutionise in the diagnosis of mental health conditions, such as depression [41,42] and developmental disorders like autism [43,44], by providing nuanced insights into patients' mental states through their drawings. Also, our findings offer a new lens to understand how human cultures and societies influence drawing behaviour and vice versa [1,2,45]. Eventually, deep learning could enrich discussions on how drawing - and by extension art - may reflect societal values and experiences. We hope our study paves the way for future explorations into the multidimensional expression that is drawing.

Acknowledgements

We thank the school director and the teachers who gave us access to their classrooms, proving their interest in our research project. We are grateful to all the participants and to the parents of

all the children, who accepted with enthusiasm to contribute to our study. Thanks also to Sarah Piquette, who provided help regarding the ethical components of this project.

Funding

This study was made possible with funding support from PNRIA and MITI (80Prime), which facilitated the research process and enabled the investigation to be conducted effectively.

Data availability: Codes and scripts are available on https://github.com/cedricsueur/drawinganalyses. Data is available on Zenodo: https://doi.org/10.5281/zenodo.11097174

References

[1] B. Binson, D.J. Federman, R. Lev-Wiesel, Do Self-Figure Drawings Reveal the Drawer's Cultural Values? Thais and Israelis Draw Themselves, Journal of Humanistic Psychology (2019) 0022167819831082. https://doi.org/10.1177/0022167819831082.

[2] S. Restoy, L. Martinet, C. Sueur, M. Pelé, Draw yourself: How culture influences drawings by children between the ages of two and fifteen, Frontiers in Psychology 13 (2022) 940617.

[3] G.-H. Luquet, Le dessin enfantin.(Bibliothèque de psychologie de l" enfant et de pédagogie.)., (1927).

[4] E. Adi-Japha, I. Levin, S. Solomon, Emergence of representation in drawing: The relation between kinematic and referential aspects, Cognitive Development 13 (1998) 25–51.

[5] R. Baldy, Dessin et développement cognitif, Enfance 57 (2005) 34–44.

[6] M.V. Cox, Children's drawings of the human figure, Psychology Press, 2013.

[7] N.H. Freeman, Drawing: Public instruments of representation., (1993).

[8] S. Khan, T. Yairi, A review on the application of deep learning in system health management, Mechanical Systems and Signal Processing 107 (2018) 241–265. https://doi.org/10.1016/j.ymssp.2017.11.024.

[9] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep Face Recognition, in: Procedings of the British Machine Vision Conference 2015, British Machine Vision Association, Swansea, 2015: p. 41.1-41.12. https://doi.org/10.5244/C.29.41.

[10] M.D. Hassan, A.N. Nasret, M.R. Baker, Z.S. Mahmood, Enhancement automatic speech recognition by deep neural networks, Periodicals of Engineering and Natural Sciences 9 (2021) 921–927. https://doi.org/10.21533/pen.v9i4.2450.

[11] R. Chauhan, K.K. Ghanshala, R. Joshi, Convolutional neural network (CNN) for image detection and recognition, in: IEEE, 2018: pp. 278–282.

[12] T. Lei, Z. Shi, D. Liu, L. Yang, F. Zhu, A novel CNN-based method for question classification in intelligent question answering, in: 2018: pp. 1–6.

[13] R. Shwartz-Ziv, N. Tishby, Opening the black box of deep neural networks via information, arXiv Preprint arXiv:1703.00810 (2017).

[14] B. Beltzung, M. Pelé, J.P. Renoult, M. Shimada, C. Sueur, Using Artificial Intelligence to Analyze Non-Human Drawings: A First Step with Orangutan Productions, Animals 12 (2022) 2761. https://doi.org/10.3390/ani12202761.

[15] B. Beltzung, M. Pelé, J.P. Renoult, C. Sueur, Deep learning for studying drawing behavior: A review, Front. Psychol. 14 (2023) 992541. https://doi.org/10.3389/fpsyg.2023.992541.
[16] S.-Y. Chen, P.-H. Lin, W.-C. Chien, Children's Digital Art Ability Training System Based on AI-Assisted Learning: A Case Study of Drawing Color Perception, Front Psychol 13 (2022) 823078. https://doi.org/10.3389/fpsyg.2022.823078.
[17] A. Philippsen, Y. Nagai, A predictive coding account for cognition in human children and chimpanzees: A case study of drawing, IEEE Trans. Cogn. Dev. Syst. (2020) 1–1. https://doi.org/10.1109/TCDS.2020.3006497.
[18] N. Ali, A. Abd-Alrazaq, Z. Shah, M. Alajlani, T. Alam, M. Househ, Artificial intelligence-based mobile application for sensing children emotion through drawings, Studies in Health Technology and Informatics 295 (2022) 118–121.
[19] L. Kissos, L. Goldner, M. Butman, N. Eliyahu, R. Lev-Wiesel, Can artificial intelligence achieve human-level performance? A pilot study of childhood sexual abuse detection in self-figure drawings, Child Abuse & Neglect 109 (2020) 104755. https://doi.org/10.1016/j.chiabu.2020.104755.
[20] B. Long, J.E. Fan, M.C. Frank, Drawings as a window into developmental changes in object representations, in: Proceedings of the 40th Annual Conference of the Cognitive Science Society., 2018.
[21] J. Moon, M.-J. Kim, S.-O. Lee, Y. Yu, A deep learning model based on triplet losses for a similar child drawing selection algorithm, Journal of the Korea Industrial Information Systems Research 27 (2022) 1–9. https://doi.org/10.9723/jksiis.2022.27.1.001.
[22] D. Pysal, S.J. Abdulkadir, S.R.M. Shukri, H. Alhussian, Classification of children's drawing strategies on touch-screen of seriation objects using a novel deep learning hybrid model, Alexandria Engineering Journal 60 (2020) 115–129.
[23] Y. Yuan, J. Huang, X. Ma, K. Yan, Children's Drawing Psychological Analysis using Shallow Convolutional Neural Network, in: 2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), IEEE, 2020: pp. 692–698.
[24] L. Martinet, C. Sueur, S. Hirata, J. Hosselet, T. Matsuzawa, M. Pelé, New indices to characterize drawing behavior in humans (Homo sapiens) and chimpanzees (Pan troglodytes), Scientific Reports 11 (2021) 3860. https://doi.org/10.1038/s41598-021-83043-0.
[25] L. Torrey, J. Shavlik, Transfer Learning, in: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, IGI Global, 2010: pp. 242–264. https://doi.org/10.4018/978-1-60566-766-9.ch011.
[26] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556 [Cs] (2015). http://arxiv.org/abs/1409.1556 (accessed November 10, 2021).
[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, F.-F. Li, Imagenet: A large-scale hierarchical image database., In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), in: 2009: pp. 248–255. http://dx.doi.org/10.1109/CVPR.2009.5206848.
[28] A. Theodorus, M. Nauta, C. Seifert, Evaluating CNN interpretability on sketch classification, in: Twelfth International Conference on Machine Vision (ICMV 2019), SPIE, 2020: pp. 475–482. https://doi.org/10.1117/12.2559536.
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016: pp. 770–778.
[30] R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-CAM: Why did you say that?, arXiv:1611.07450 [Cs, Stat] (2017). http://arxiv.org/abs/1611.07450 (accessed April 26, 2022).
[31] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, Captum: A unified and generic model interpretability library for pytorch, arXiv Preprint arXiv:2009.07896 (2020).
[32] B. Beltzung, L. Martinet, A.J. Macintosh, X. Meyer, J. Hosselet, M. Pelé, C. Sueur, To Draw Or Not To Draw: Understanding The Temporal Organization Of Drawing Behavior Using Fractal Analyses, Fractals 31 (2023) 2350009.

[33] C. Sueur, L. Martinet, B. Beltzung, M. Pelé, Making drawings speak through mathematical metrics, Human Nature 33 (2022) 400–424.
[34] J. Matthews, Children drawing: Are young children really scribbling?, Early Child Development and Care 18 (1984) 1–39. https://doi.org/10.1080/0300443840180101.
[35] A. Schöttl, A light-weight method to foster the (Grad) CAM interpretability and explainability of classification networks, in: IEEE, 2020: pp. 348–351.
[36] Q. Zhang, S.-C. Zhu, Visual interpretability for deep learning: a survey, Frontiers of Information Technology & Electronic Engineering 19 (2018) 27–39.
[37] R. Baldy, Fais-moi un beau dessin: regarder le dessin de l'enfant, comprendre son évolution, In Press, 2011.
[38] M. Carabantes, Black-box artificial intelligence: an epistemological and critical analysis, AI & Society 35 (2020) 309–317.
[39] E. Duede, Deep learning opacity in scientific discovery, Philosophy of Science 90 (2023) 1089–1099.
[40] F. Faries, V. Raja, Black Boxes and Theory Deserts: Deep Networks and Epistemic Opacity in the Cognitive Sciences, (2022).
[41] L. Eytan, D.L. Elkis-Abuhoff, Indicators of depression and self-efficacy in the PPAT drawings of normative adults, The Arts in Psychotherapy 40 (2013) 291–297. https://doi.org/10.1016/j.aip.2013.04.003.
[42] J. Kim, S. Chung, Drawing Test Form for Depression: The Development of Drawing Tests for Predicting Depression Among Breast Cancer Patients, Psychiatry Investig 18 (2021) 879–888. https://doi.org/10.30773/pi.2021.0044.
[43] T. Charman, S. Baron-Cohen, Drawing development in autism: The intellectual to visual realism shift, British Journal of Developmental Psychology 11 (1993) 171–185. https://doi.org/10.1111/j.2044-835X.1993.tb00596.x.
[44] A. Lee, R.P. Hobson, Drawing self and others: How do children with autism differ from those with learning difficulties?, British Journal of Developmental Psychology 24 (2006) 547–565. https://doi.org/10.1348/026151005X49881.
[45] P. Bozzato, C. Longobardi, Cross-cultural evaluation of children's drawings of gender role stereotypes in italian and cambodian students, Journal of Psychological and Educational Research 29 (2021) 97–115.

FIGURES

Figure 1. Examples of drawings to illustrate the variability in artistic expression across different age groups and conditions. Each column showcases drawings from distinct age categories: 3-year-olds (a), 7-year-olds (b), and expert adults (c). Within the figure, two rows are used to differentiate the drawing conditions. The first row features drawings created under a free condition, where the participants had the liberty to draw whatever they chose, reflecting their spontaneous creativity and imagination. The second row displays drawings produced in a self-portrait condition, where participants were asked to draw themselves, providing insight into

their self-perception and ability to represent human features. This juxtaposition of ages and conditions offers a visual comparison of developmental progression in artistic skills and conceptual understanding from early childhood through to expert adult levels.

Figure 2. Confusion matrix for the model that classifies drawings into 10 distinct categories. In this matrix, each cell represents the number of drawings that have been classified by the model. The column labels indicate the categories predicted by the model for these drawings, while the row labels denote the true categories to which the drawings actually belong. The diagonal cells, where the predicted category matches the true label, show the number of correctly classified drawings for each category. The off-diagonal cells reveal the instances of misclassification, where the model predicted a category different from the true category. This matrix provides a detailed view of the model's performance across all categories, highlighting its accuracy and areas where confusion between categories occurs.

Figure 3. Accuracy of various models trained to distinguish between children's drawings and adults' drawings. Each model, corresponding to different age groups of children, was trained multiple times — specifically, 10 iterations — to ensure reliability and to account for any variability in the training process. The graph plots the mean accuracy achieved by each of these models across their training iterations, providing a visual representation of how well each model performs in differentiating between the artistic expressions of children at various developmental stages and those of adults. This comparative analysis not only highlights the overall effectiveness of the models but also allows for the examination of how drawing characteristics and discernibility evolve with age.

Figure 4. mean prediction values for each model across various age categories, with separate analyses for (a) RGB and (b) grayscale models. These values stem from the models' evaluations of drawings, reflecting the perceived similarity between drawings from different age groups

and the target categories defined by the models (children vs. adults). The provided matrices offer a quantifiable measure of this similarity, where each cell denotes the average model prediction for drawings belonging to a specific age group. For instance, an analysis of the RGB matrix (a) first row allows us to understand how the model distinguishes between drawings by 4 to 10-year-olds, 2 to 3-year-olds, and adults. Comparison values range from 0, indicating that drawings from the examined age category are more similar to those of 2 to 3-year-olds, to 1, suggesting closer similarity to adult drawings. Specifically, on the first row's second-coloured row, a mean prediction value of 0.84 for drawings by 5-year-olds implies that, on average, the model perceives these drawings as more similar to adult drawings (closer to 1) than to those of younger children (further from 0). This interpretation of the matrices facilitates a detailed understanding of how the models discern differences and similarities in drawings across age groups, effectively quantifying the developmental progression in drawing skills from the perspective of the models' classifications.

Figure 5. Visual exploration of Grad-CAM (Gradient-weighted Class Activation Mapping) heatmaps for drawings from the validation set of different models, showcasing how the model focuses on specific areas of the drawings to make its classifications. Grad-CAM is a technique used to highlight the regions of an input image that are important for predictions from a convolutional neural network model. Panel a) features a heatmap overlay on a drawing by a 5-year-old, analysed within the 4–5 years-old model framework. Despite the drawing being incorrectly classified as an adult's work with a predicted value of 0.6885, the heatmap interestingly highlights the face as a significant feature for its decision, indicating the model's reliance on facial features for classification, even though the overall prediction was inaccurate. Panel c) presents a heatmap for a drawing by a 10-year-old, evaluated by the 9–10 years-old model. This drawing is correctly classified, yet the heatmap appears nonsensical, failing to

highlight discernible features that justify its classification. This suggests that while the model's prediction was correct, the rationale behind its focus is unclear, raising questions about the interpretability of the model's decision-making process. The heatmaps for correctly classified adults' drawings, as seen in panels b) and d) and computed using the 2–3 years-old model, demonstrate varying degrees of focus. In b), the heatmap seems to concentrate on specific features, possibly contributing to a high-confidence prediction. Conversely, in d), despite a high prediction value, the heatmap does not highlight any particular feature, indicating that the model's decision-making process might not always align with human-intuitive feature recognition. These examples illustrate the complexity and variability in how deep learning models interpret and classify drawings. While Grad-CAM heatmaps offer valuable insights into the regions of interest that models use for their predictions, the interpretability of these visual explanations can vary significantly, from being seemingly logical to puzzling, highlighting the challenges in understanding and improving model accuracy and reliability.

Figure 6. Examples of Captum recognition, where black pixels represent elements deemed most important in the models' age classification. (a) and (b) showcase instances where faces are accurately recognised. (c) illustrates an example where the body, but not the face, is identified. (d) highlights the significance of the eye in a wolf's drawing. (e) displays Captum's analysis of a scribble. (f) demonstrates a drawing with multiple faces, of which only two are identified. (g) exhibits an example where three faces are melded or overlapped with other elements, posing a challenge for recognition. (h) shows a case where a cat's face is not recognised. (i) depicts an example where flowers are identified as the most important elements by Captum.

Figure 1

Figure 2

Click here to access/download;Figure;FIgure 2.tif

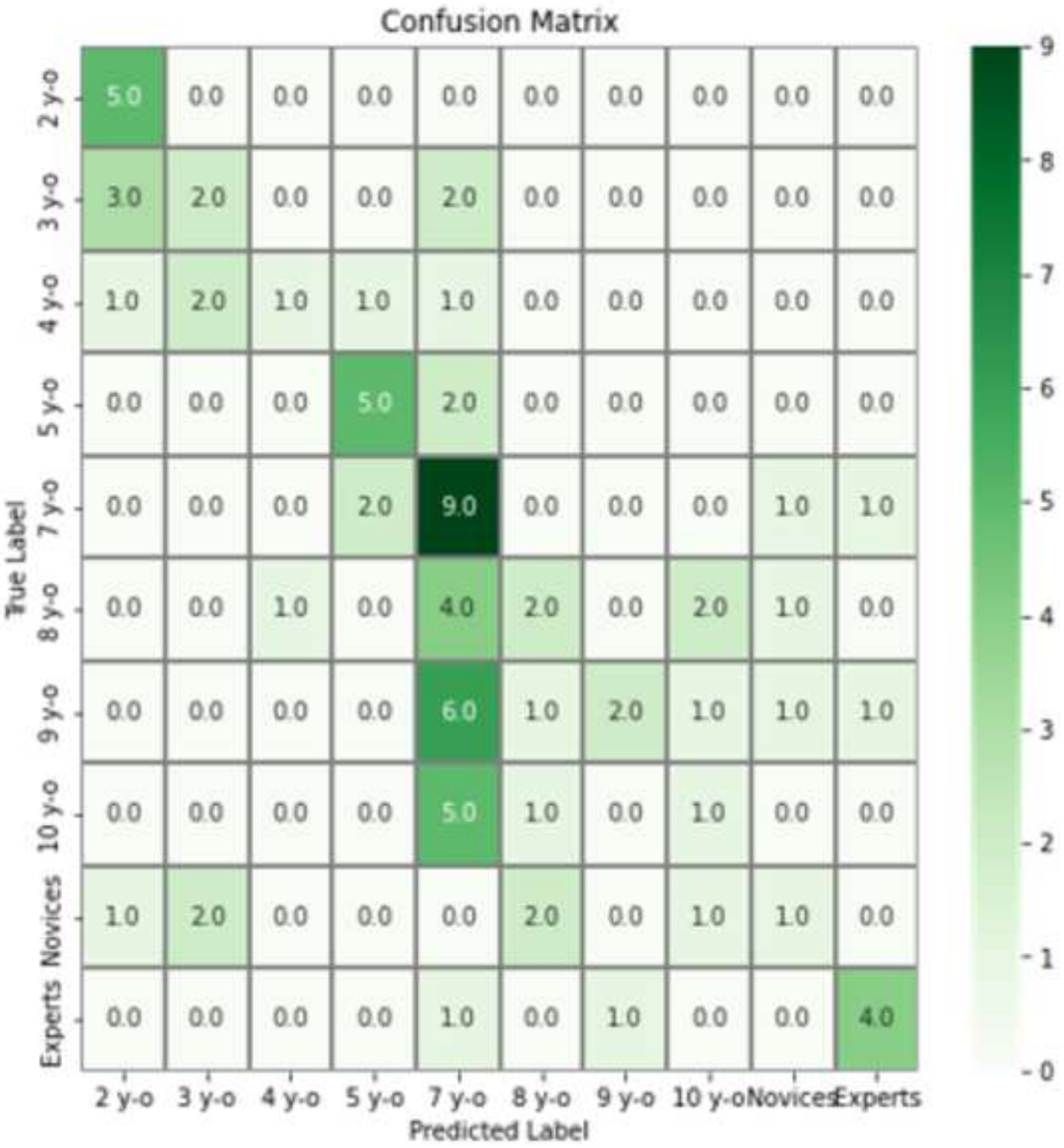


Figure 3

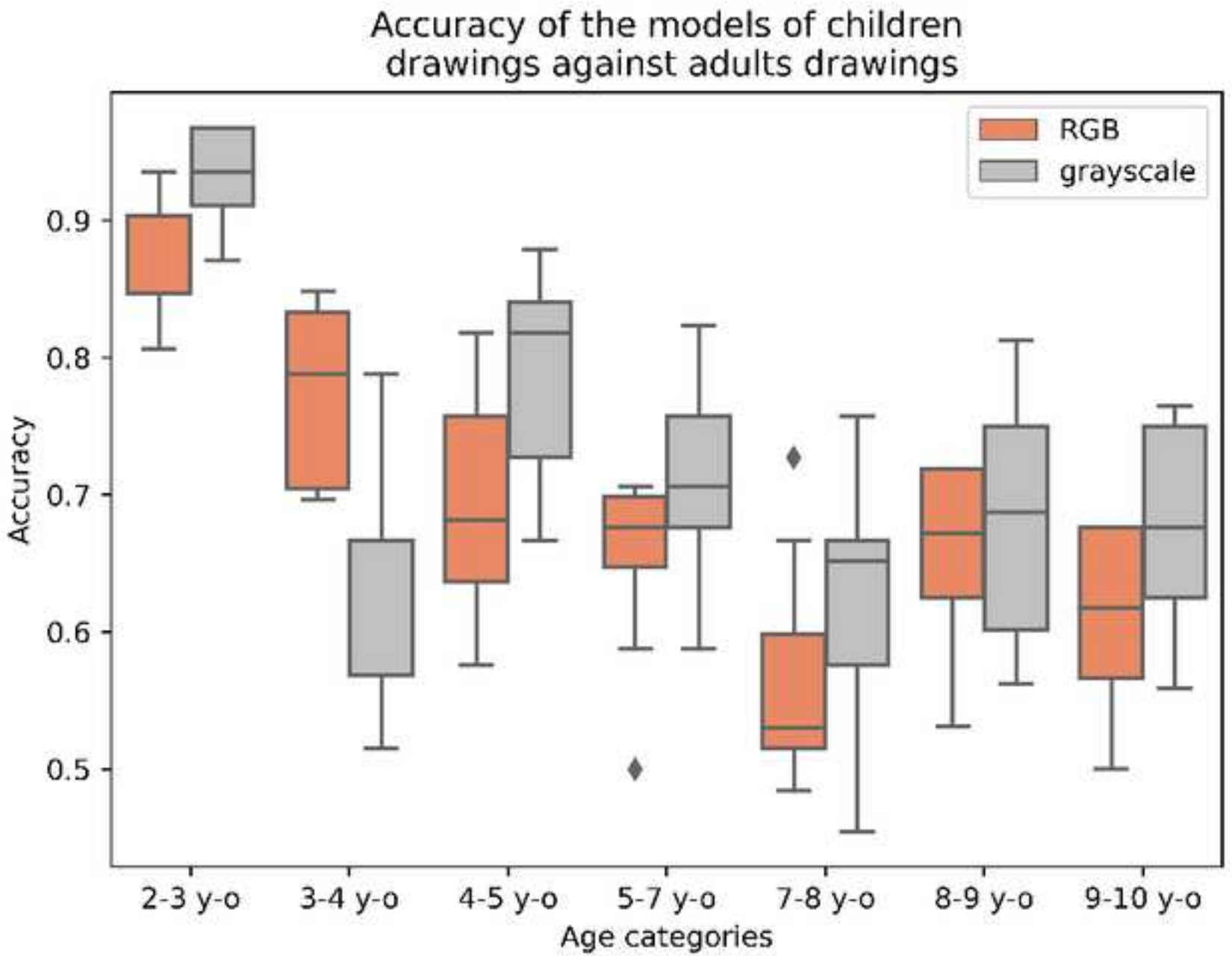


Figure 4

Click here to access/download;Figure;figure 4.tif

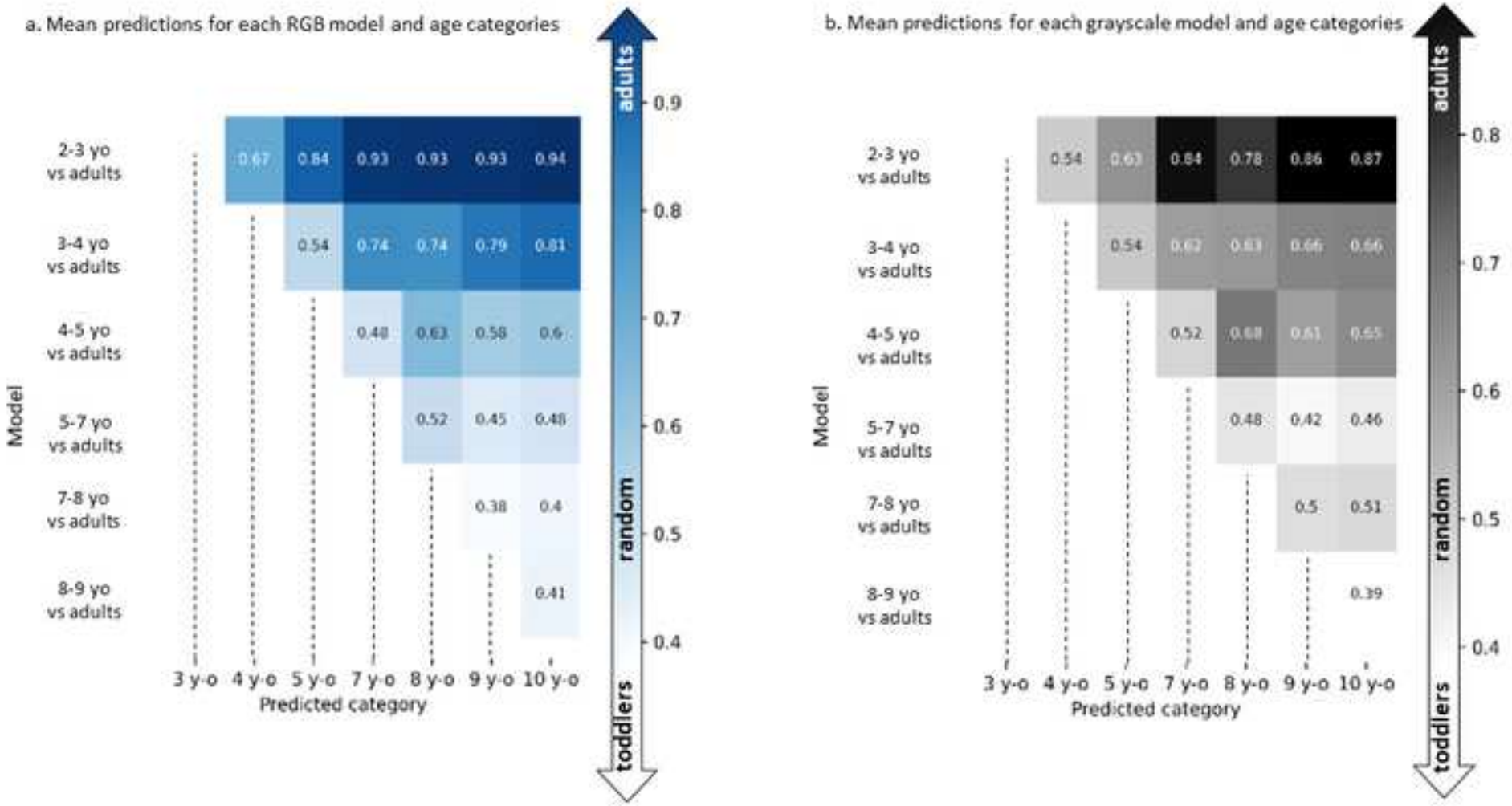


Figure 5

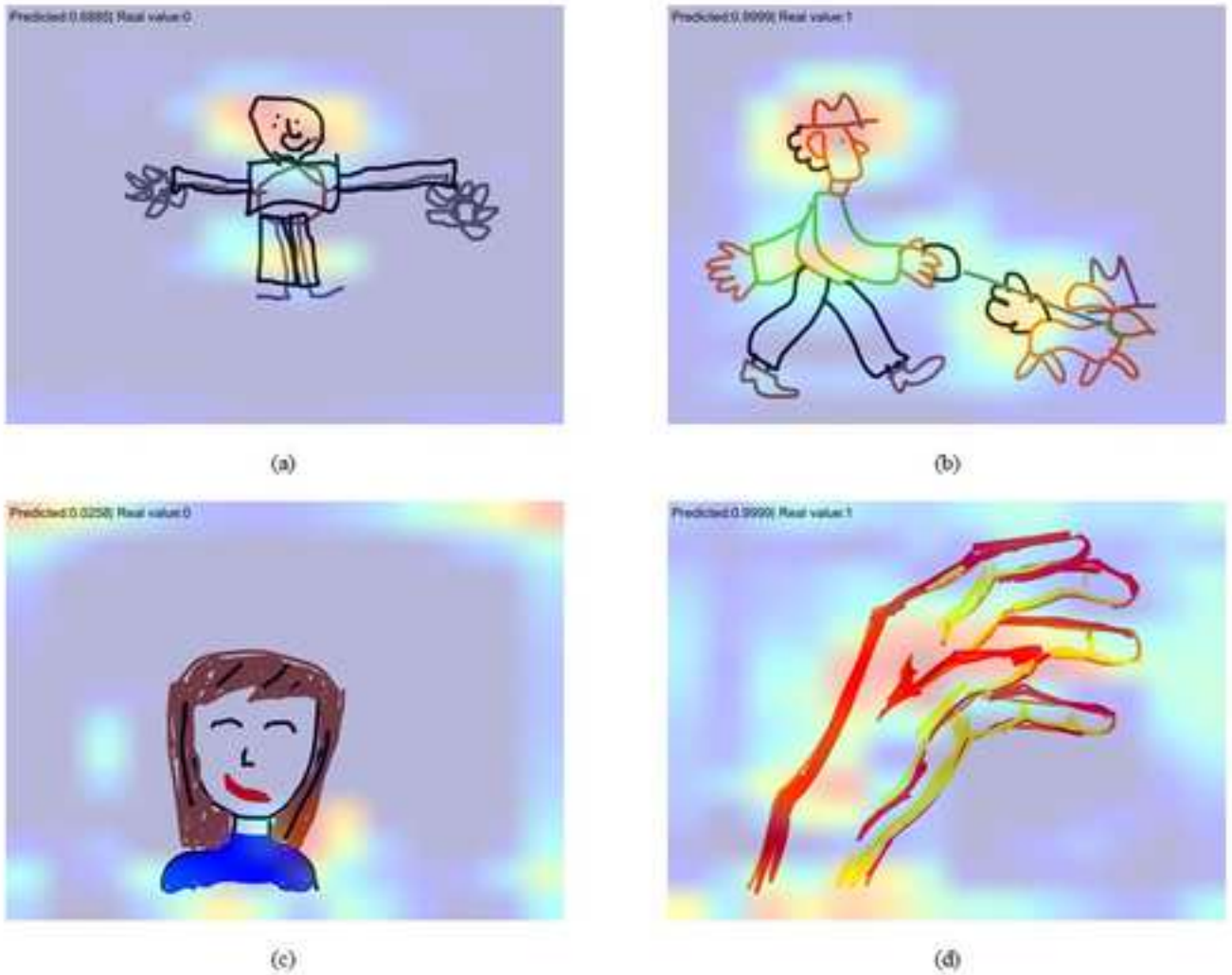
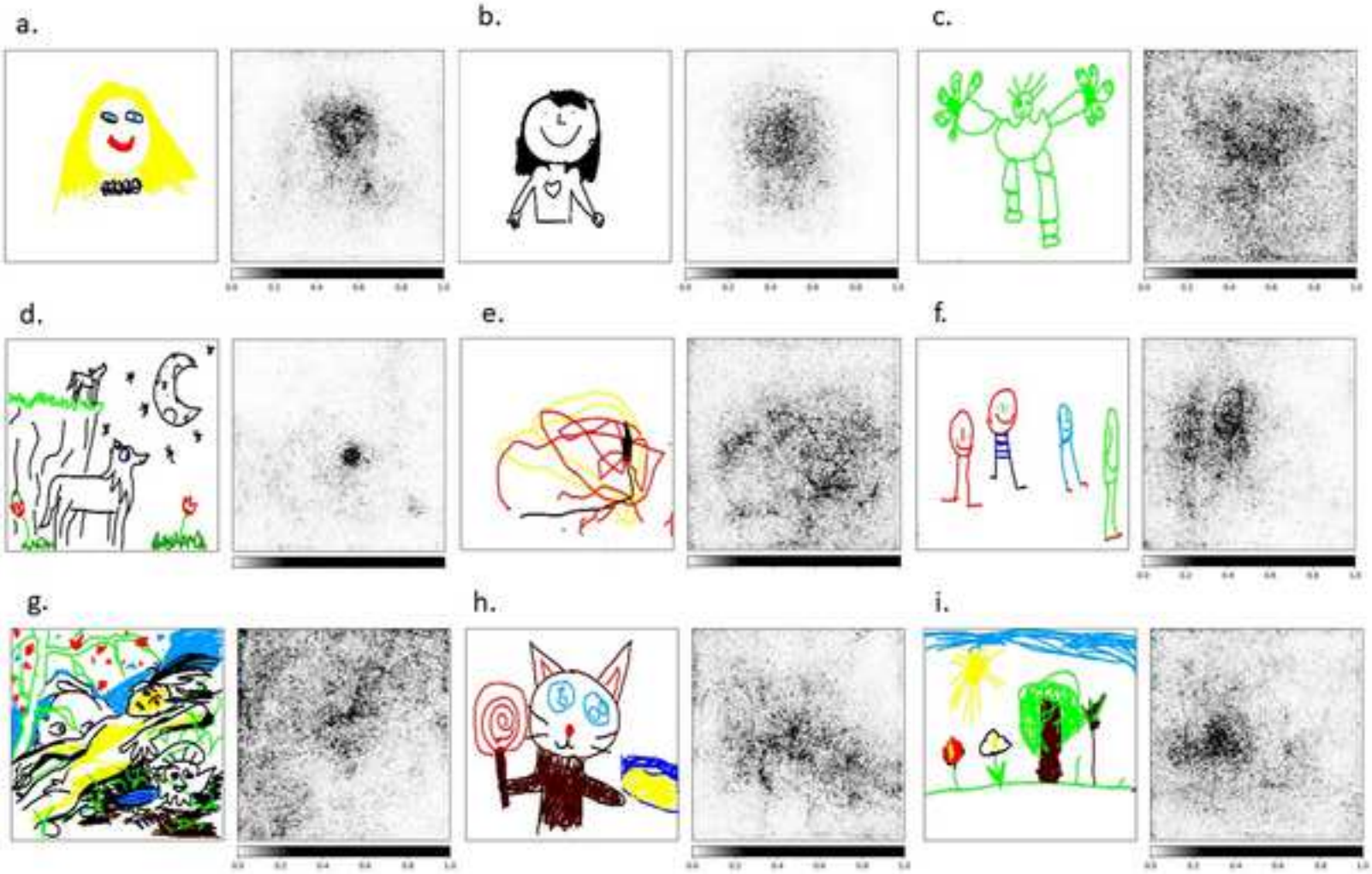


Figure 6

Click here to access/download;Figure;Figure 6.tif



Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: