



HAL
open science

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

► To cite this version:

Benjamin Beltzung, Marie Pelé, Lison Martinet, Elliot Maître, Jimmy Falck, et al.. Using deep learning predictions to study the development of drawing behaviour in children. 2024. <hal-04714749>

HAL Id: hal-04714749

<https://univ-catholille.hal.science/hal-04714749v1>

Preprint submitted on 30 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

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

3
4 Benjamin Beltzung¹, Marie Pelé², Lison Martinet¹, Jimmy Falck³, Elliot Maitre⁴, Cédric
5 Sueur^{1,5}

6 1 Université de Strasbourg, CNRS, IPHC UMR 7178, Strasbourg, France

7 2 ANTHROPO-LAB – ETHICS EA 7446, Université Catholique de Lille, F-59000 Lille,
8 France

9 3 Laboratoire lorrain de recherche en informatique et ses applications (Loria –
10 CNRS/Université de Lorraine/Inria), Nancy, France

11 4 Université de Toulouse - IRIT UMR5505, 31400, Toulouse, France

12 5 Institut Universitaire de France, Paris, France

13

14 Corresponding author : Cédric Sueur, cedric.sueur@iphc.cnrs.fr; +33388107453; IPHC UMR
15 7178, 23 rue Becquerel 67087 Strasbourg, France

16

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

27

28 **Keywords:** Artificial Intelligence, Cognitive Development, Explicability, Interpretability,
29 Drawing Behaviour

30 **Introduction**

31

32 In children, drawing behaviour appears around the age of 18 months-old. During a
33 lifetime, drawing is an important mode of communication driven not only by cognitive aspects
34 but also by cultural ones. For instance, by utilising human figure representations in drawings,
35 [1] demonstrated variations in body size and shape among young adults from Israel and
36 Thailand when asked to depict themselves. Similarly, through analysing drawings created by
37 children from different cultures, Restoy and colleagues [2] showcased how the level of
38 individualism in countries could influence the size and number of human figures depicted. Of
39 course, culture is not the sole factor influencing drawing behaviour and numerous cognitive
40 aspects have been explored through it. Luquet [3] firstly suggested that the development of the
41 drawing behaviour can be seen as a four-stage process. The first stage occurs when the child
42 lacks intention to represent reality and discovers by chance a shape analogy between an object
43 and its initially nonsignificant trace (*fortuitous realism*). Around four or five years old, the child
44 tries to produce realistic drawings but does not have the entire abilities (motor and
45 representative skills) to do so (*missed realism*). Then, the child uses her knowledge of objects'
46 components to represent them but some misconceptions linked for example to transparency or
47 perspective representation are still present (*intellectual realism*). Eventually, the last stage
48 proposed by Luquet [3] is *visual realism*, when every step is completed.

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

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

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

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

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

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

100 Although deep learning has already been successfully used to analyse drawings [14–
101 17], even in children [18,19]; the development of the drawing behaviour has been, to our
102 knowledge, only poorly analysed through the lens of deep learning [20–23].

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

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

110

111 **Material and methods**

112 *a. Dataset*

113 The data consist of 386 drawings produced by 150 children (from two to ten years-old) and
114 43 adults (novice and experts) (detail in Table S1). All the participants were given a touch-
115 screen displaying a white background and could choose among a panel of ten colours by having
116 access to an overlay on the bottom of the screen. Examples of drawings are shown in Figure 1.
117 From three years-old to adults, the drawings were collected using the following protocol. Each
118 participant was asked to produce two drawings, one under a free condition, where the
119 participant was not given a particular instruction, and one under a self-portrait condition, where
120 the drawer was asked to draw himself. Two categories of adults were defined prior to the data
121 collection. First, a group of adults who had never taken drawing classes and did not have
122 drawing as a hobby, that will be here considered as novices. Then, an expert category, including
123 art school students and professional illustrators. The data distribution is presented in Table S1.
124 For more information about these datasets, please refer to [24].

125 As two-year-old participants were not verbally able to communicate and to understand
126 the instructions, no particular task was given for these productions. For this reason, they will be
127 considered as produced in free condition. In this category, 6 subjects participated and produced
128 a total number of 30 drawings, ranging from 3 to 8 drawings per individual (respectively 3, 4,
129 4, 5, 6, 8 drawings).

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

132

133 *b. Ethics*

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

141

142 *c. Transfer learning from deep learning*

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

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

168

169 *d. Bayesian optimization*

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

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

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

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

200

201 *e. Predictions matrices*

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

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

222

223 *f. Explicability*

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

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

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

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

246

247 **Results**

248 *a. 10-classes model*

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

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

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

277

278 *b. Models by pair*

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

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

310

311 *c. Interpretability models*

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

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

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

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

337

338 **Discussion**

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

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

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

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

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

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

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

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

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

426

427 **Acknowledgements**

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

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

432

433 **Funding**

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

436

437 **Data availability:** Codes and scripts are available on
438 <https://github.com/cedricsueur/drawinganalyses>. Data is available on Zenodo:
439 <https://doi.org/10.5281/zenodo.11097174>

440

441

442

443

444 **References**

- 445 [1] B. Binson, D.J. Federman, R. Lev-Wiesel, Do Self-Figure Drawings Reveal the Drawer's
446 Cultural Values? Thais and Israelis Draw Themselves, *Journal of Humanistic Psychology* (2019)
447 0022167819831082. <https://doi.org/10.1177/0022167819831082>.
- 448 [2] S. Restoy, L. Martinet, C. Sueur, M. Pelé, Draw yourself: How culture influences drawings by
449 children between the ages of two and fifteen, *Frontiers in Psychology* 13 (2022) 940617.
- 450 [3] G.-H. Luquet, *Le dessin enfantin*. (Bibliothèque de psychologie de l' enfant et de pédagogie.),
451 (1927).
- 452 [4] E. Adi-Japha, I. Levin, S. Solomon, Emergence of representation in drawing: The relation
453 between kinematic and referential aspects, *Cognitive Development* 13 (1998) 25–51.
- 454 [5] R. Baldy, *Dessin et développement cognitif*, *Enfance* 57 (2005) 34–44.
- 455 [6] M.V. Cox, *Children's drawings of the human figure*, Psychology Press, 2013.
- 456 [7] N.H. Freeman, *Drawing: Public instruments of representation.*, (1993).
- 457 [8] S. Khan, T. Yairi, A review on the application of deep learning in system health management,
458 *Mechanical Systems and Signal Processing* 107 (2018) 241–265.
459 <https://doi.org/10.1016/j.ymssp.2017.11.024>.
- 460 [9] O.M. Parkhi, A. Vedaldi, A. Zisserman, Deep Face Recognition, in: *Proceedings of the British*
461 *Machine Vision Conference 2015*, British Machine Vision Association, Swansea, 2015: p. 41.1-
462 41.12. <https://doi.org/10.5244/C.29.41>.
- 463 [10] M.D. Hassan, A.N. Nasret, M.R. Baker, Z.S. Mahmood, Enhancement automatic speech
464 recognition by deep neural networks, *Periodicals of Engineering and Natural Sciences* 9 (2021)
465 921–927. <https://doi.org/10.21533/pen.v9i4.2450>.
- 466 [11] R. Chauhan, K.K. Ghanshala, R. Joshi, Convolutional neural network (CNN) for image detection
467 and recognition, in: *IEEE*, 2018: pp. 278–282.
- 468 [12] T. Lei, Z. Shi, D. Liu, L. Yang, F. Zhu, A novel CNN-based method for question classification
469 in intelligent question answering, in: 2018: pp. 1–6.
- 470 [13] R. Shwartz-Ziv, N. Tishby, Opening the black box of deep neural networks via information,
471 *arXiv Preprint arXiv:1703.00810* (2017).
- 472 [14] B. Beltzung, M. Pelé, J.P. Renoult, M. Shimada, C. Sueur, Using Artificial Intelligence to
473 Analyze Non-Human Drawings: A First Step with Orangutan Productions, *Animals* 12 (2022)
474 2761. <https://doi.org/10.3390/ani12202761>.

- 475 [15] B. Beltzung, M. Pelé, J.P. Renoult, C. Sueur, Deep learning for studying drawing behavior: A
476 review, *Front. Psychol.* 14 (2023) 992541. <https://doi.org/10.3389/fpsyg.2023.992541>.
- 477 [16] S.-Y. Chen, P.-H. Lin, W.-C. Chien, Children's Digital Art Ability Training System Based on
478 AI-Assisted Learning: A Case Study of Drawing Color Perception, *Front Psychol* 13 (2022)
479 823078. <https://doi.org/10.3389/fpsyg.2022.823078>.
- 480 [17] A. Philippsen, Y. Nagai, A predictive coding account for cognition in human children and
481 chimpanzees: A case study of drawing, *IEEE Trans. Cogn. Dev. Syst.* (2020) 1–1.
482 <https://doi.org/10.1109/TCDS.2020.3006497>.
- 483 [18] N. Ali, A. Abd-Alrazaq, Z. Shah, M. Alajlani, T. Alam, M. Househ, Artificial intelligence-based
484 mobile application for sensing children emotion through drawings, *Studies in Health Technology
485 and Informatics* 295 (2022) 118–121.
- 486 [19] L. Kissos, L. Goldner, M. Butman, N. Eliyahu, R. Lev-Wiesel, Can artificial intelligence achieve
487 human-level performance? A pilot study of childhood sexual abuse detection in self-figure
488 drawings, *Child Abuse & Neglect* 109 (2020) 104755.
489 <https://doi.org/10.1016/j.chiabu.2020.104755>.
- 490 [20] B. Long, J.E. Fan, M.C. Frank, Drawings as a window into developmental changes in object
491 representations, in: *Proceedings of the 40th Annual Conference of the Cognitive Science
492 Society.*, 2018.
- 493 [21] J. Moon, M.-J. Kim, S.-O. Lee, Y. Yu, A deep learning model based on triplet losses for a
494 similar child drawing selection algorithm, *Journal of the Korea Industrial Information Systems
495 Research* 27 (2022) 1–9. <https://doi.org/10.9723/jksii.2022.27.1.001>.
- 496 [22] D. Pysal, S.J. Abdulkadir, S.R.M. Shukri, H. Alhussian, Classification of children's drawing
497 strategies on touch-screen of seriation objects using a novel deep learning hybrid model,
498 *Alexandria Engineering Journal* 60 (2020) 115–129.
- 499 [23] Y. Yuan, J. Huang, X. Ma, K. Yan, Children's Drawing Psychological Analysis using Shallow
500 Convolutional Neural Network, in: *2020 International Conferences on Internet of Things
501 (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber,
502 Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE
503 Congress on Cybermatics (Cybermatics)*, IEEE, 2020: pp. 692–698.
- 504 [24] L. Martinet, C. Sueur, S. Hirata, J. Hosselet, T. Matsuzawa, M. Pelé, New indices to characterize
505 drawing behavior in humans (*Homo sapiens*) and chimpanzees (*Pan troglodytes*), *Scientific
506 Reports* 11 (2021) 3860. <https://doi.org/10.1038/s41598-021-83043-0>.
- 507 [25] L. Torrey, J. Shavlik, Transfer Learning, in: *Handbook of Research on Machine Learning
508 Applications and Trends: Algorithms, Methods, and Techniques*, IGI Global, 2010: pp. 242–264.
509 <https://doi.org/10.4018/978-1-60566-766-9.ch011>.
- 510 [26] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image
511 Recognition, *arXiv:1409.1556 [Cs]* (2015). <http://arxiv.org/abs/1409.1556> (accessed November
512 10, 2021).
- 513 [27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, F.-F. Li, Imagenet: A large-scale hierarchical
514 image database., In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, in:
515 2009: pp. 248–255. <http://dx.doi.org/10.1109/CVPR.2009.5206848>.
- 516 [28] A. Theodorus, M. Nauta, C. Seifert, Evaluating CNN interpretability on sketch classification, in:
517 *Twelfth International Conference on Machine Vision (ICMV 2019)*, SPIE, 2020: pp. 475–482.
518 <https://doi.org/10.1117/12.2559536>.
- 519 [29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016: pp.
520 770–778.
- 521 [30] R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-CAM: Why did
522 you say that?, *arXiv:1611.07450 [Cs, Stat]* (2017). <http://arxiv.org/abs/1611.07450> (accessed
523 April 26, 2022).
- 524 [31] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N.
525 Kliushkina, C. Araya, S. Yan, Captum: A unified and generic model interpretability library for
526 pytorch, *arXiv Preprint arXiv:2009.07896* (2020).
- 527 [32] B. Beltzung, L. Martinet, A.J. Macintosh, X. Meyer, J. Hosselet, M. Pelé, C. Sueur, To Draw Or
528 Not To Draw: Understanding The Temporal Organization Of Drawing Behavior Using Fractal
529 Analyses, *Fractals* 31 (2023) 2350009.

- 530 [33] C. Sueur, L. Martinet, B. Beltzung, M. Pelé, Making drawings speak through mathematical
531 metrics, *Human Nature* 33 (2022) 400–424.
- 532 [34] J. Matthews, Children drawing: Are young children really scribbling?, *Early Child Development
533 and Care* 18 (1984) 1–39. <https://doi.org/10.1080/0300443840180101>.
- 534 [35] A. Schöttl, A light-weight method to foster the (Grad) CAM interpretability and explainability of
535 classification networks, in: *IEEE*, 2020: pp. 348–351.
- 536 [36] Q. Zhang, S.-C. Zhu, Visual interpretability for deep learning: a survey, *Frontiers of Information
537 Technology & Electronic Engineering* 19 (2018) 27–39.
- 538 [37] R. Baldy, Fais-moi un beau dessin: regarder le dessin de l'enfant, comprendre son évolution, In
539 Press, 2011.
- 540 [38] M. Carabantes, Black-box artificial intelligence: an epistemological and critical analysis, *AI &
541 Society* 35 (2020) 309–317.
- 542 [39] E. Duede, Deep learning opacity in scientific discovery, *Philosophy of Science* 90 (2023) 1089–
543 1099.
- 544 [40] F. Faries, V. Raja, Black Boxes and Theory Deserts: Deep Networks and Epistemic Opacity in
545 the Cognitive Sciences, (2022).
- 546 [41] L. Eytan, D.L. Elkis-Abuhoff, Indicators of depression and self-efficacy in the PPAT drawings
547 of normative adults, *The Arts in Psychotherapy* 40 (2013) 291–297.
548 <https://doi.org/10.1016/j.aip.2013.04.003>.
- 549 [42] J. Kim, S. Chung, Drawing Test Form for Depression: The Development of Drawing Tests for
550 Predicting Depression Among Breast Cancer Patients, *Psychiatry Investig* 18 (2021) 879–888.
551 <https://doi.org/10.30773/pi.2021.0044>.
- 552 [43] T. Charman, S. Baron-Cohen, Drawing development in autism: The intellectual to visual realism
553 shift, *British Journal of Developmental Psychology* 11 (1993) 171–185.
554 <https://doi.org/10.1111/j.2044-835X.1993.tb00596.x>.
- 555 [44] A. Lee, R.P. Hobson, Drawing self and others: How do children with autism differ from those
556 with learning difficulties?, *British Journal of Developmental Psychology* 24 (2006) 547–565.
557 <https://doi.org/10.1348/026151005X49881>.
- 558 [45] P. Bozzato, C. Longobardi, Cross-cultural evaluation of children's drawings of gender role
559 stereotypes in italian and cambodian students, *Journal of Psychological and Educational
560 Research* 29 (2021) 97–115.
561

562

563 FIGURES

564

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

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

575

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

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

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

597 and the target categories defined by the models (children vs. adults). The provided matrices
598 offer a quantifiable measure of this similarity, where each cell denotes the average model
599 prediction for drawings belonging to a specific age group. For instance, an analysis of the RGB
600 matrix (a) first row allows us to understand how the model distinguishes between drawings by
601 4 to 10-year-olds, 2 to 3-year-olds, and adults. Comparison values range from 0, indicating that
602 drawings from the examined age category are more similar to those of 2 to 3-year-olds, to 1,
603 suggesting closer similarity to adult drawings. Specifically, on the first row's second-coloured
604 row, a mean prediction value of 0.84 for drawings by 5-year-olds implies that, on average, the
605 model perceives these drawings as more similar to adult drawings (closer to 1) than to those of
606 younger children (further from 0). This interpretation of the matrices facilitates a detailed
607 understanding of how the models discern differences and similarities in drawings across age
608 groups, effectively quantifying the developmental progression in drawing skills from the
609 perspective of the models' classifications.

610

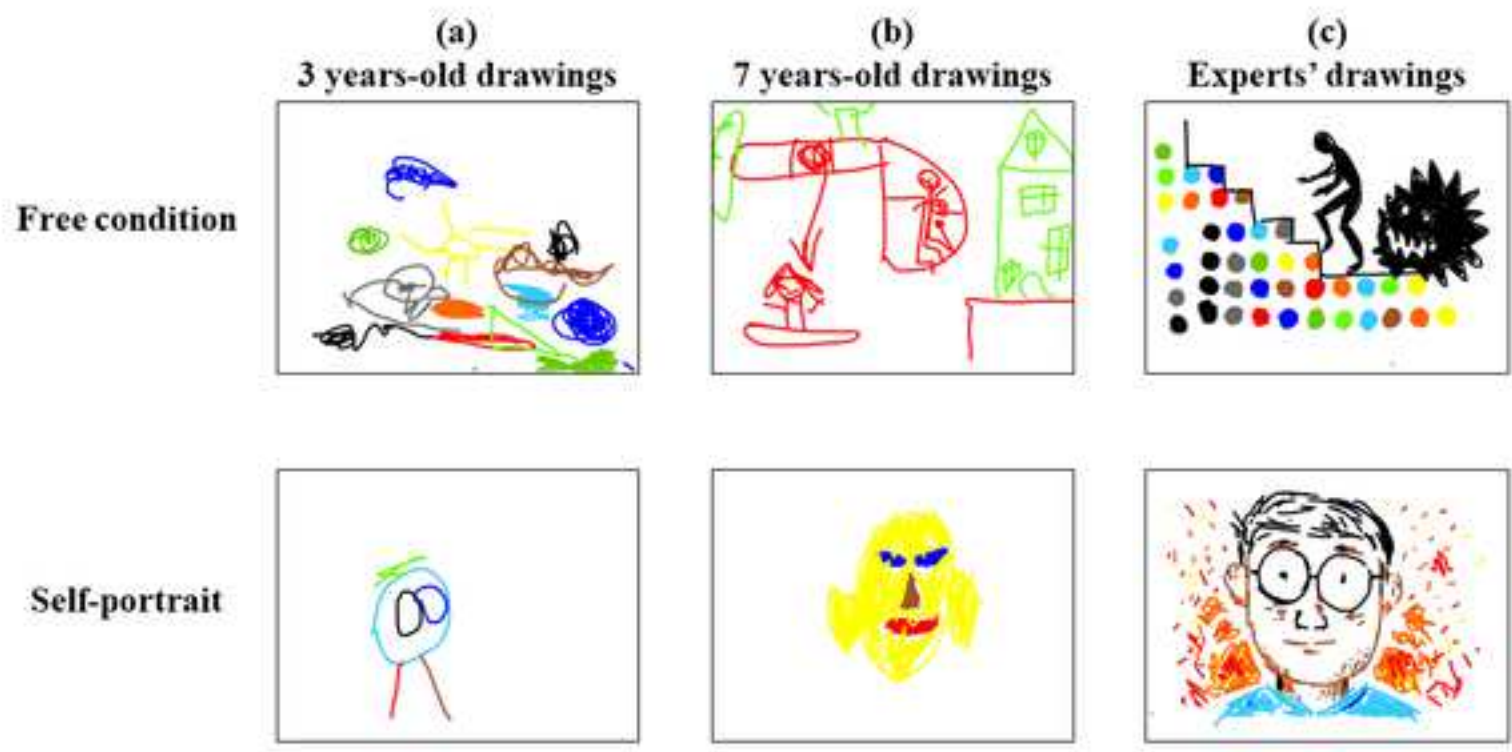
611 Figure 5. Visual exploration of Grad-CAM (Gradient-weighted Class Activation Mapping)
612 heatmaps for drawings from the validation set of different models, showcasing how the model
613 focuses on specific areas of the drawings to make its classifications. Grad-CAM is a technique
614 used to highlight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

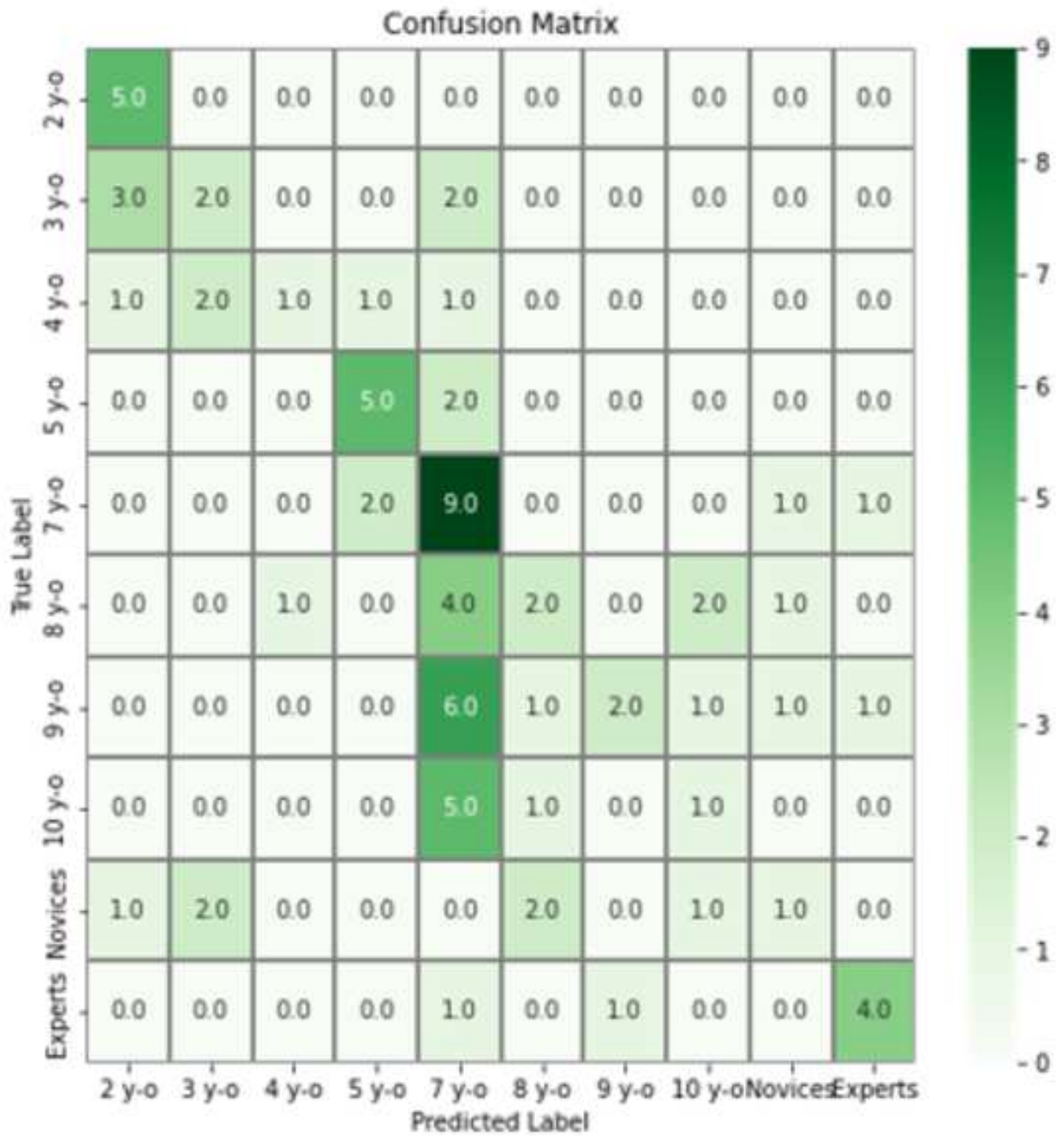
622 highlight discernible features that justify its classification. This suggests that while the model's
623 prediction was correct, the rationale behind its focus is unclear, raising questions about the
624 interpretability of the model's decision-making process. The heatmaps for correctly classified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

635

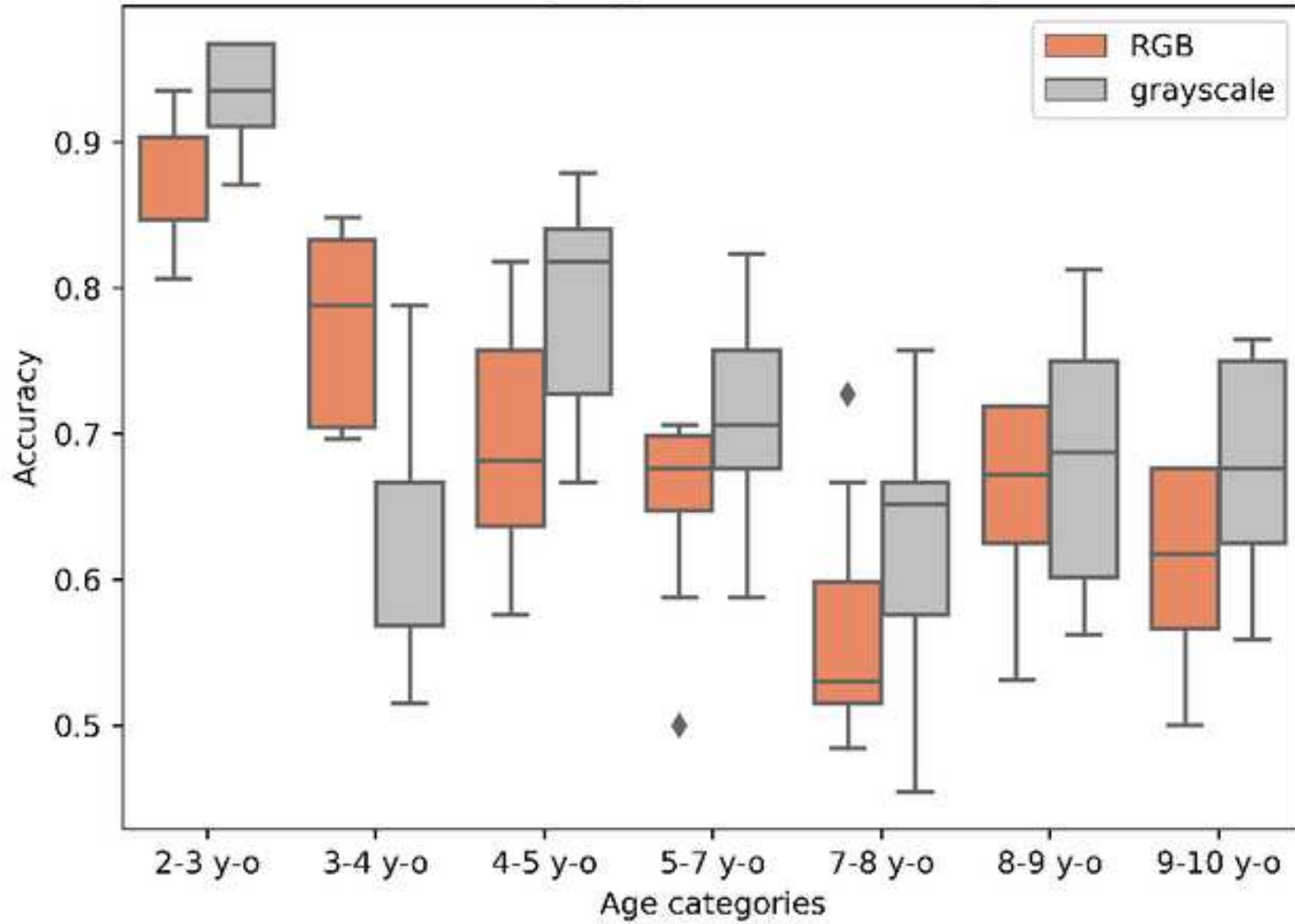
636 Figure 6. Examples of Captum recognition, where black pixels represent elements deemed
637 most important in the models' age classification. (a) and (b) showcase instances where faces
638 are accurately recognised. (c) illustrates an example where the body, but not the face, is
639 identified. (d)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

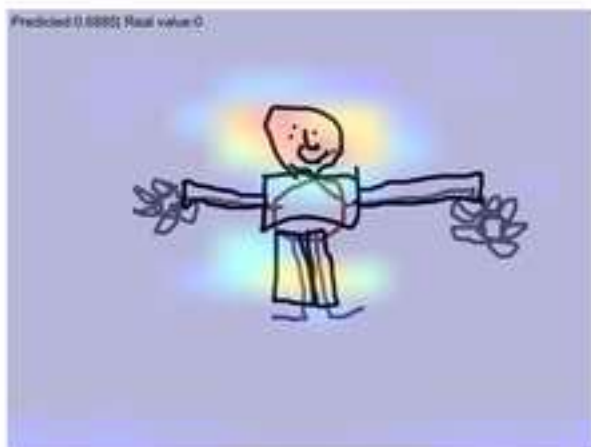
645





Accuracy of the models of children drawings against adults drawings





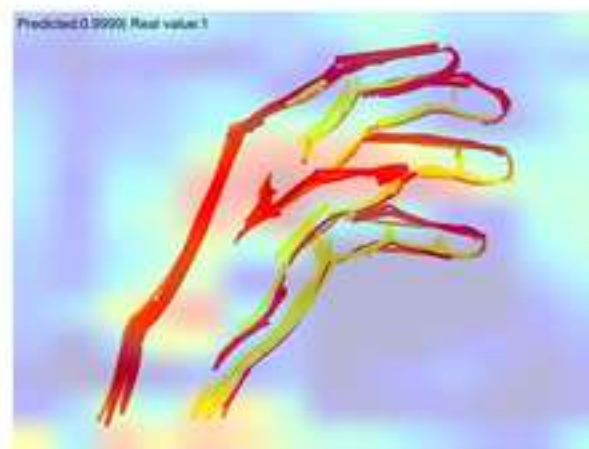
(a)



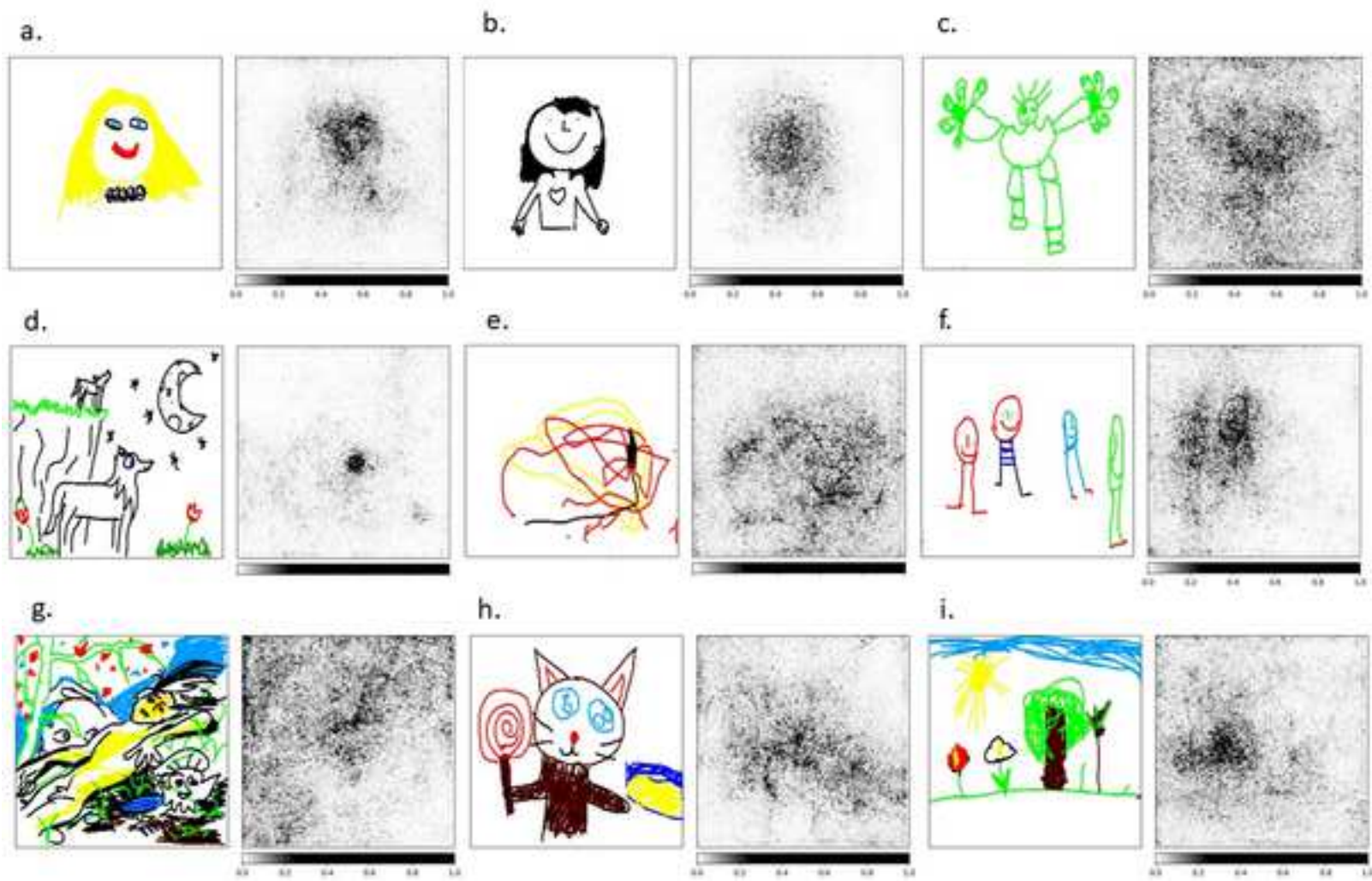
(b)



(c)



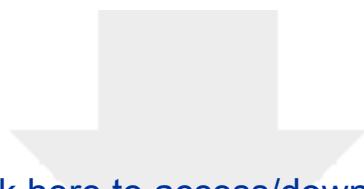
(d)



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:



Click here to access/download
e-Component
Supplementary information.docx

