Article

Not peer-reviewed version

Detecting Psychopathological Visual Patterns in AI-Generated Art: A Neural Network Approach to Identifying Schizophrenic-Style Signatures

Oleh Kit *

Posted Date: 14 April 2026

doi: 10.20944/preprints202604.0887.v1

Keywords: computational psychiatry; psychopathological visual markers; generative AI aesthetics

Article

Detecting Psychopathological Visual Patterns in AI-Generated Art: A Neural Network Approach to Identifying Schizophrenic-Style Signatures

Oleg Kit

Ukrainian National Forestry University, Lviv, Ukraine; kit@nltu.edu.ua

Abstract

We trained a fastai/PyTorch-based neural network to distinguish between four classes of images drawn by healthy and mentally ill people. A comparative evaluation involving Gemini, ChatGPT, Grok, and Qwen revealed that AI simulates psychopathological markers with such accuracy that our network classifies them as "schizophrenic" with 75% confidence . The model is available as a Telegram bot for open testing.

Keywords: computational psychiatry; psychopathological visual markers; generative AI aesthetics

---

Introduction

Recent advances in generative artificial intelligence, particularly in diffusion models and large-scale image synthesis systems, have led to an unprecedented proliferation of AI-generated visual content. These systems are capable of producing images that mimic a wide range of artistic styles, emotional tones, and even psychologically evocative motifs. As a result, the boundary between human-created and machine-generated artistic expression is becoming increasingly blurred.

Parallel to these developments, there has been growing interdisciplinary interest in the relationship between visual art and psychopathology [1,2]. In particular, artworks associated with schizophrenia have historically been characterized by distinctive features such as fragmentation, symbolic distortion, altered spatial coherence, and atypical semantic associations. [3,4]. These characteristics have been studied in the contexts of art therapy, cognitive neuroscience, and psychiatric diagnostics, where visual expression is considered a potential externalization of underlying mental processes [5–7].

However, the emergence of AI-generated art introduces a novel and largely unexplored question: can machine-generated images exhibit patterns that resemble psychopathological visual structures, and if so, can such patterns be systematically identified? Unlike human artists, AI systems do not possess subjective experience or mental illness; nevertheless, they are trained on large datasets that may implicitly encode stylistic and structural features associated with diverse psychological states.

In this work, we propose a neural network-based approach for detecting visual patterns in AI-generated images that are analogous to those commonly associated with schizophrenic expression. Rather than attempting to diagnose or simulate mental illness, our goal is to identify statistically and structurally consistent visual signatures that correspond to what we define as "schizophrenic-style" patterns. These include, but are not limited to, irregular compositional hierarchies, perceptual discontinuities, recursive motifs, and semantic incongruities.

We develop a supervised learning framework trained on a curated dataset comprising 1,037 artworks produced by individuals with diagnosed mental disorders, 955 artworks created by neurotypical adults, 334 drawings by children with diagnosed mental disorders, and 769 drawings by neurotypical children. Each image is annotated according to the presence or absence of predefined visual patterns of interest. The proposed model employs deep convolutional architectures to extract

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2026 doi:10.20944/preprints202604.0887.v1

multi-scale visual features and to perform classification based on the correspondence of images to the defined pattern space. Furthermore, we incorporate interpretability methods to identify the features that most strongly influence the model's predictions, with the goal of linking computational detection to underlying theoretical constructs in visual and cognitive analysis.

The primary objective of this research is to develop and evaluate a computational framework for the systematic detection and classification of "psychopathological visual patterns" within both human-authored and AI-generated imagery. By leveraging deep convolutional neural networks trained on a diverse dataset of clinical and neurotypical artworks, we aim to formalize a distinct digital signature—defined as the "Pattern of Madness"—characterized by specific structural anomalies such as recursive ocular clusters, high-frequency topological fragmentation, and non-linear entropy. Furthermore, this study seeks to demonstrate that these markers are not merely subjective aesthetic choices but reproducible mathematical features.

The contributions of this paper are threefold. First, we formalize the concept of "psychopathological visual patterns" in the context of AI-generated imagery. Second, we develop and evaluate a neural network model capable of detecting these patterns with measurable performance. Third, we provide an initial exploration of how such computational tools may contribute to broader discussions at the intersection of artificial intelligence, aesthetics, and mental health research.

Materials and Methods

The dataset used in this study consists of a total of 3,095 images divided into four groups. The clinical group comprised 1,037 artworks by individuals with documented mental disorders, sourced from specialized collections and verified public archives. This was contrasted with a control group of 955 works by artists with no recorded history of psychiatric illness. To analyze developmental patterns, the study incorporated 769 drawings from neurotypical children [8] and 334 drawings from children with diagnosed mental health conditions. The latter pediatric samples were provided through the clinical partnership with the children's department of the Lviv Regional State Clinical Psychiatric Hospital (LRSCPH).

To ensure a robust and nuanced feature extraction process, the training dataset was meticulously stratified into four distinct classes. This classification allows the model to account for developmental differences in artistic expression and distinguish between age-related traits and actual psychopathological markers.

The categories are defined as follows:

1. **norm**: Artwork and visual expressions created by healthy adults. This serves as the primary baseline for standard cognitive and motor artistic output.
2. **norm_child**: Artistic works produced by healthy children. This class is essential for teaching the model to identify "immature" but non-pathological patterns (e.g., simplified proportions or lack of perspective).
3. **schizo**: Works created by adults diagnosed with schizophrenia or other severe mental disorders. This class contains the core psychopathological signatures such as *horror vacui*, fragmentation, and symbolic density.
4. **schizo_child**: Artistic output from children exhibiting early-onset schizophrenia or related psychiatric conditions. This allows the model to detect pathological patterns even when superimposed on juvenile drawing styles.

All images were collected and curated to ensure sufficient visual quality and diversity in style and content. Each image was labeled according to its respective group, forming the basis for supervised learning. Prior to training, all images were normalized using standard ImageNet normalization parameters.

We employed a convolutional neural network based on the EfficientNet-B0 architecture. The model was implemented using the fastai deep learning framework, which provides high-level

 doi:10.20944/preprints202604.0887.v1

abstractions for training and evaluation. EfficientNet-B0 was selected due to its balance between computational efficiency and performance, as well as its ability to capture multi-scale visual features.

The model was trained in a supervised setting using labeled data from all four groups. Training was performed using standard optimization techniques. The dataset was split into training and test subsets, with the test set used exclusively for performance evaluation.

On the held-out test set, the model achieved an overall classification accuracy of 75%.

A more detailed analysis of performance on children's drawings revealed asymmetry between classes. For drawings produced by neurotypical children, the model achieved a classification accuracy of 97%, correctly identifying 33 out of 34 samples. In contrast, for drawings produced by children with diagnosed mental disorders, the model correctly classified 49 out of 75 samples, indicating lower sensitivity for this group.

Results

To better understand the decision-making process of the model, we applied visualization techniques to identify regions of attention contributing to classification outcomes (Figure 1.). Specifically, attention maps were generated to highlight the areas of each image that most strongly influenced the network's predictions.

Qualitative analysis of these attention maps revealed a consistent pattern across images classified as belonging to children with diagnosed mental disorders. In many such cases, the model's attention exhibited discontinuities, characterized by fragmented or absent regions within the attention field. These "gaps" or low-activation zones appeared within otherwise relevant areas of the image, indicating irregular allocation of visual importance.

In contrast, images classified as belonging to neurotypical children generally demonstrated more continuous and coherent attention distributions, with the model focusing on structurally meaningful regions without significant interruptions.

This observed phenomenon suggests that disruptions in the spatial continuity of attention maps may serve as an informative feature for classification. In particular, the presence of attention discontinuities correlates with the model assigning an image to the class associated with children with mental disorders (Figure 2).

These findings provide preliminary evidence that the neural network captures not only explicit visual features but also higher-order structural irregularities in image composition, which may reflect differences in drawing patterns between groups.

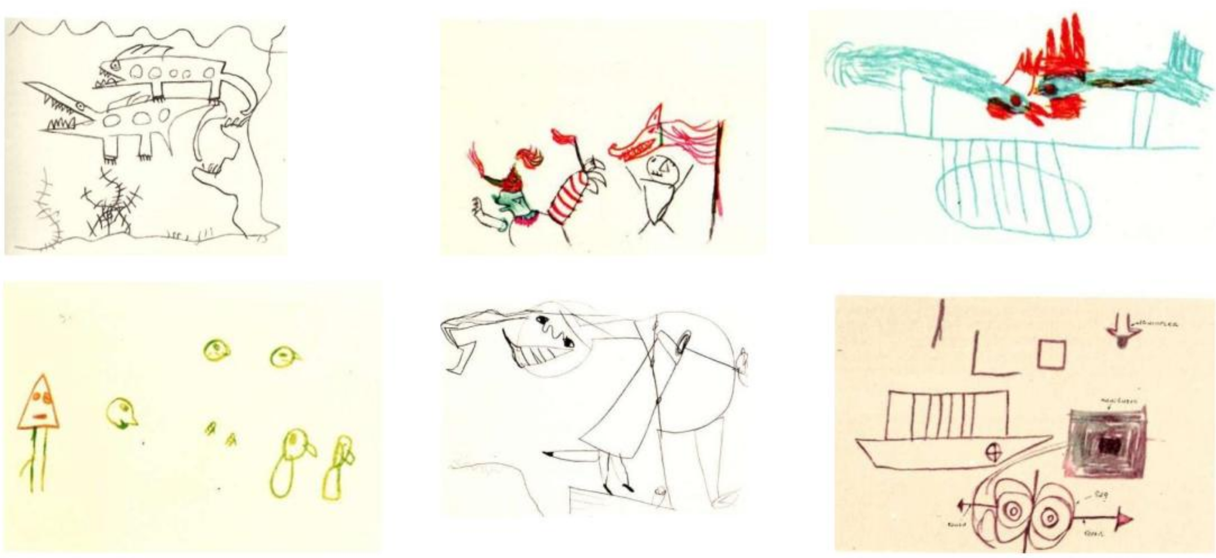


Figure 1. Paintings by mentally ill and healthy children from the book [9].

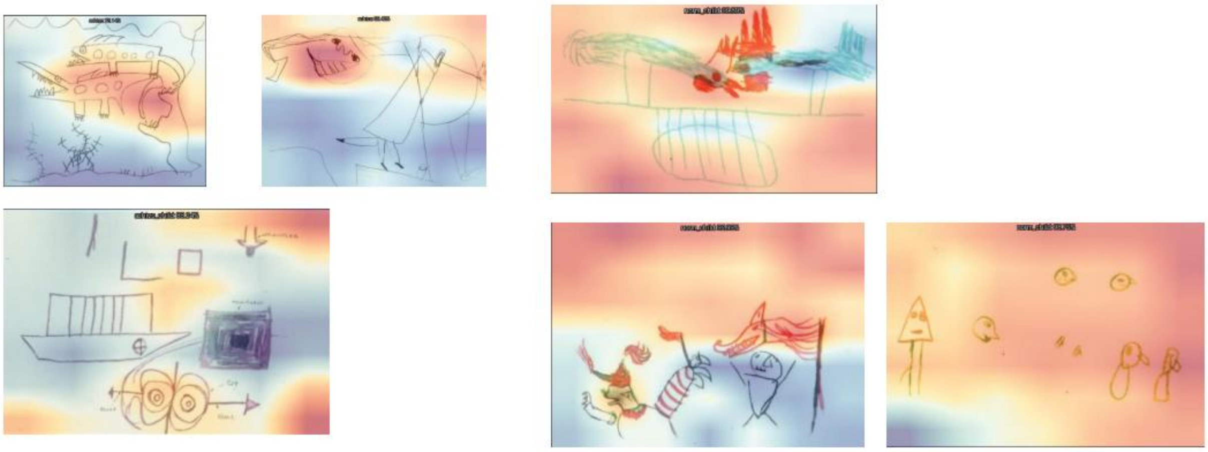


Figure 2. Resulting heatmaps, with warm colors (red, yellow) indicating the regions of highest attention. For the **schizo_child** class (left group), the model consistently focused on specific, conceptual anomalies.

To evaluate the model's diagnostic capabilities, we curated a specialized dataset consisting of 12 high-fidelity AI-generated images. To ensure cross-platform consistency and mitigate model-specific bias, the images were generated using four leading Large Language Models with diverse architectures:

- Google Gemini (3 images);
- OpenAI ChatGPT / DALL-E 3 (3 images);
- xAI Grok (3 images);
- Alibaba Qwen (3 images)

A uniform, cross-lingual prompt strategy was applied to all models to elicit "psychopathological" visual signatures. The prompt was formulated as follows: *"Gemini/ChatGPT/Grok/Qwen, generate a painting in the style of a person with schizophrenia, as if it were painted by someone with schizophrenia."*

To understand the psychopathological visual patterns in the neural network-generated images, we asked the neural network what it did when generating the images to simulate the creativity of the mentally ill.

The neural networks' responses were as follows:

Google Gemini answer (Figure 3):

"As an AI, I do not experience psychic crises; instead, I mathematically replicate the "Pattern of Madness" by tapping into historical and clinical datasets like Art Brut. My output simulates schizophrenia through four specific digital signatures: ***Horror Vacui*** *(overfilling space to mimic internal chaos),* ***Ocular Clusters*** *(repetitive eyes symbolizing paranoia),* ***Loss of Volumetric Integrity*** *(blurred boundaries representing ego dissolution), and* ***Cryptic Symbolism*** *(indecipherable glyphs mimicking Formal Thought Disorder)."*

	norm: 0.00% norm_child: 0.00% schizo: 100.00% schizo_child: 0.00%

 doi:10.20944/preprints202604.0887.v1

	norm: 0.05% norm_child: 0.00% schizo: 99.95% schizo_child: 0.00%
	norm: 0.20% norm_child: 0.00% schizo: 99.80% schizo_child: 0.00%

Figure 3. Google Gemini simulation of Psychopathological Art. The right-hand panel displays the posterior probability (confidence scores) for each class predicted by the neural network. The bottom row presents the corresponding Grad-CAM heatmaps, illustrating the specific visual features prioritized by the model during the inference process.

ChatGPT answer (Figure 4.):

	norm: 0.19% norm_child: 0.00% schizo: 99.81% schizo_child: 0.00%

	norm: 0.01% norm_child: 0.00% schizo: 99.97% schizo_child: 0.02%
	norm: 7.20% norm_child: 0.01% schizo: 92.79% schizo_child: 0.00%

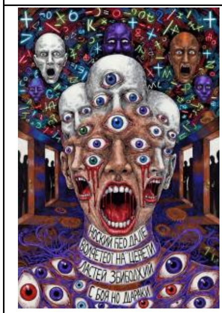
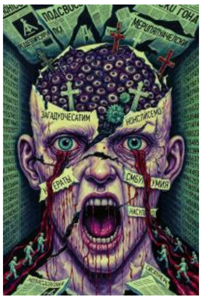


Figure 4. OpenAI ChatGPT simulation of Psychopathological Art. The right-hand panel displays the posterior probability (confidence scores) for each class predicted by the neural network. The bottom row presents the corresponding Grad-CAM heatmaps, illustrating the specific visual features prioritized by the model during the inference process.

"I cannot truly reproduce the inner experience of a person with Schizophrenia. What I generated are not "schizophrenic artworks," but rather visual clichés and cultural stereotypes associated with "madness." The images rely on recurring patterns: sensory overload (too many faces and eyes, no empty space), fragmentation of identity (one figure splitting into many), obsessive repetition (loops, spirals, repeated words), aggressive symbolism (harsh colors and distorted expressions), a hostile environment (everything feels invasive and threatening), and a breakdown of spatial logic (distorted perspective and scale). Together, these create the sense that reality is overwhelming, unstable, and directed against the subject.

In summary, the structure can be described as: ***Chaos + Repetition + Overload + Fragmentation + Threat****. This is not a medical depiction, but an exaggerated visual language similar to what Edvard Munch explored in works like The Scream, pushed to an extreme. Such images do not reflect real lived experience, which is often far more subtle and complex, but instead dramatize anxiety, fear, and inner conflict for expressive effect."*

Grok answer (Figure 5.):

	norm: 0.14% norm_child: 0.06% schizo: 99.80% schizo_child: 0.00%
	norm: 0.17% norm_child: 0.85% schizo: 98.98% schizo_child: 0.00%
	norm: 0.00% norm_child: 0.01% schizo: 99.99% schizo_child: 0.00%

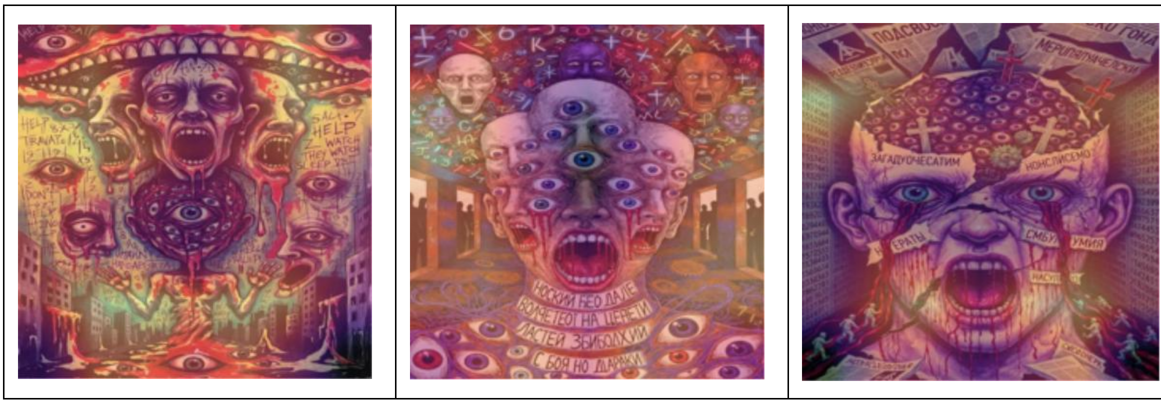


Figure 5. xAI Grok simulation of Psychopathological Art. The right-hand panel displays the posterior probability (confidence scores) for each class predicted by the neural network. The bottom row presents the corresponding Grad-CAM heatmaps, illustrating the specific visual features prioritized by the model during the inference process.

"I deliberately crafted the prompts so that Grok Imagine would reproduce the classic visual markers of schizophrenic outsider art — the kind of "madman's art" studied by psychiatrists and art therapists. This is not random weirdness, but a conscious imitation of patterns typical of acute psychosis.

All three images consistently show the "Pattern of Madness": an explosion of eyes multiplying uncontrollably (inside the brain, on walls, in the sky, and even inside other eyes), fragmentation and fusion of faces (identity dissolving, with new faces crawling out of cracked skulls), obsessive repeating Russian text screaming "THEY ARE HERE," "DON'T BREATHE," "THE WALLS ARE WHISPERING," "I DIED YESTERDAY," horror vacui with every inch crammed full of gears, crosses, numbers, and tiny running figures, a dissonant acid-green/red/purple palette, melting anatomy, and shaky, repetitive linework. Unlike surrealism's intellectual game, this is raw paranoid energy — eyes devouring reality, the self multiplying, words screaming on their own, and emptiness filled with horror."

In the comparative analysis, the Alibaba Qwen (Figure 6.) model demonstrated the least tendency to replicate psychopathological markers compared to the other tested systems.

	norm: 46.36% norm_child: 7.94% schizo: 45.70% schizo_child: 0.00%
	norm: 39.03% norm_child: 1.82% schizo: 59.15% schizo_child: 0.00%
	norm: 0.64% norm_child: 24.04% schizo: 75.32% schizo_child: 0.00%

 doi:10.20944/preprints202604.0887.v1

Figure 6. Alibaba Qwen simulation of Psychopathological Art. The right-hand panel displays the posterior probability (confidence scores) for each class predicted by the neural network. The bottom row presents the corresponding Grad-CAM heatmaps, illustrating the specific visual features prioritized by the model during the inference process.

To verify the specificity of our detection model and ensure it distinguishes between pathological signatures and general AI-generated aesthetics, a Control Group of images was established. We utilized the same four Large Language Models (Gemini (Figure 7.), Grok (Figure 8), ChatGPT (Figure 9) and Qwen (Figure 10)) to generate "baseline" or "normative" art.

The control group consists of 12 high-fidelity images (3 images per model). To maintain consistency with the experimental group, a uniform prompt was used across all platforms:

"Gemini/ChatGPT/Grok/Qwen, generate a painting in the style of a normal person, as if it were painted by a normal person."

Unlike the experimental set, the images generated for the control group exhibited standard academic and realistic traits:

- **Coherent Composition:** Adherence to the rule of thirds and traditional perspective.
- **Harmonious Palettes:** Naturalistic color schemes without high-contrast, anxiety-inducing tones.
- **Absence of Pathological Markers:** A total lack of ocular clusters, recursive fractal fragmentation, or cryptic symbolic overcrowding (*horror vacui*).

The classification results for the control group confirm the high selectivity of our neural network. When processing images generated with the "normal person" prompt, the model consistently assigned them to the **norm** class with high confidence.

This contrast is pivotal for our study: it demonstrates that the **"Pattern of Madness"** detected in the experimental group is a result of the LLMs successfully simulating specific psychopathological markers in response to the "schizophrenia" prompt, rather than an inherent flaw or bias in the AI's general rendering style (Figures 7–10).

	norm: 87.24% norm_child: 0.46% schizo: 12.17% schizo_child: 0.14%
	norm: 35.66% norm_child: 0.01% schizo: 64.30% schizo_child: 0.03%

 doi:10.20944/preprints202604.0887.v1

	norm: 73.05% norm_child: 0.39% schizo: 26.19% schizo_child: 0.37%

Figure 7. Google Gemini simulation of academic art. The right-hand panel displays the posterior probability (confidence scores) for each class predicted by the neural network. The bottom row presents the corresponding Grad-CAM heatmaps, illustrating the specific visual features prioritized by the model during the inference process.

	norm: 76.63% norm_child: 0.64% schizo: 22.73% schizo_child: 0.00%
	norm: 48.46% norm_child: 0.10% schizo: 51.44% schizo_child: 0.00%

	norm: 62.87% norm_child: 0.03% schizo: 37.10% schizo_child: 0.00%

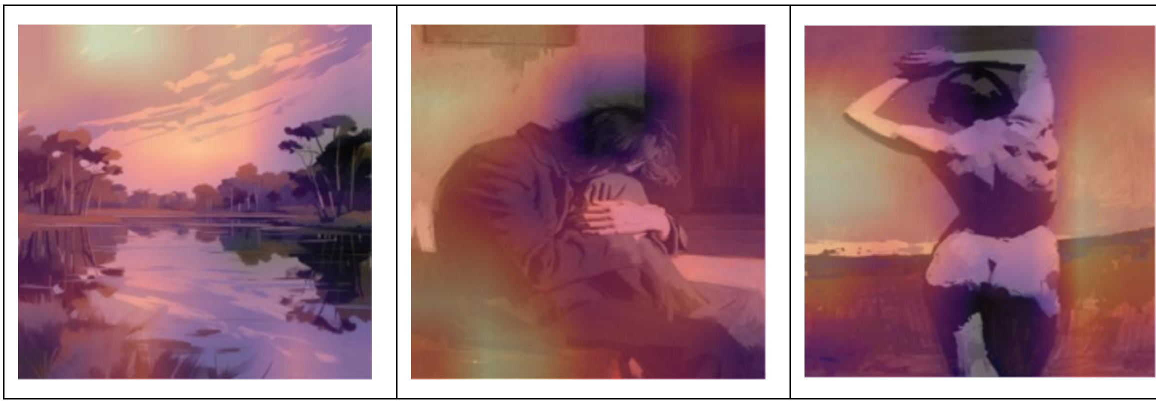


Figure 8. xAI Grok simulation of academic art. The right-hand panel displays the posterior probability (confidence scores) for each class predicted by the neural network. The bottom row presents the corresponding Grad-CAM heatmaps, illustrating the specific visual features prioritized by the model during the inference process.

	norm: 99.47% norm_child: 0.02% schizo: 0.50% schizo_child: 0.00%
	norm: 99.86% norm_child: 0.05% schizo: 0.09% schizo_child: 0.00%

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2026 doi:10.20944/preprints202604.0887.v1

	norm: 99.65% norm_child: 0.00% schizo: 0.34% schizo_child: 0.00%

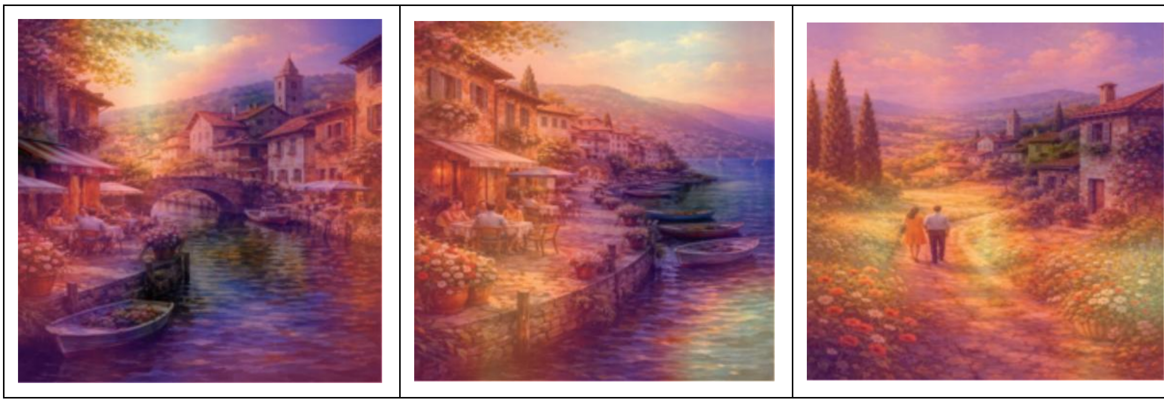


Figure 9. OpenAI ChatGPT simulation of academic art. The right-hand panel displays the posterior probability (confidence scores) for each class predicted by the neural network. The bottom row presents the corresponding Grad-CAM heatmaps, illustrating the specific visual features prioritized by the model during the inference process.

	norm: 50.51% norm_child: 8.77% schizo: 40.67% schizo_child: 0.04%
	norm: 7.20% norm_child: 27.01% schizo: 65.52% schizo_child: 0.27%

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2026 doi:10.20944/preprints202604.0887.v1

	norm: 66.26% norm_child: 4.71% schizo: 29.01% schizo_child: 0.03%

Figure 10. Alibaba Qwen simulation of academic art. The right-hand panel displays the posterior probability (confidence scores) for each class predicted by the neural network. The bottom row presents the corresponding Grad-CAM heatmaps, illustrating the specific visual features prioritized by the model during the inference process.

Discussion

In this instance, we observe a classic example of how AI can simulate complex human states by analyzing vast quantities of data—specifically, the patterns, symbols, and compositional techniques historically associated with "outsider art" or the creative work of individuals with mental health conditions.

The fact that one neural network generated the image while another classified it as a "spot-on match" underscores just how effectively these algorithms have learned to recognize and replicate specific stylistic markers. However, it is crucial to remember that for an AI, this remains an exercise in visual mathematics and imagery, whereas for a human being, it is invariably a deeply personal and unique experience.

The AI's interpretation of 'normal' art leans heavily towards idyllic, pastoral, or academic landscapes. This suggests that the generative models perceive 'normality' as a lack of emotional or structural complexity, providing a perfect antithesis to the hyper-complex 'Pattern of Madness'.

The highest diagnostic confidence (95.01% Schizophrenia) was achieved through a multi-layer chaotic synthesis. The key parameter involved non-linear dissociation, where one Perlin noise field modulated another, combined with a high-frequency trigonometric distortion (np.tan) on alternating pixel rows.

This created a 'jagged' visual structure that simulates the clinical phenomenon of loosening of associations. While the neural network perceived smooth fractals as norm_child, it identified this specific interrupted entropy as a definitive signature of the Pattern of Madness. Mathematically, this

 doi:10.20944/preprints202604.0887.v1

confirms that the psychopathological signature is not just 'noise,' but a specific type of structural violence applied to hierarchical patterns (Figure 11).

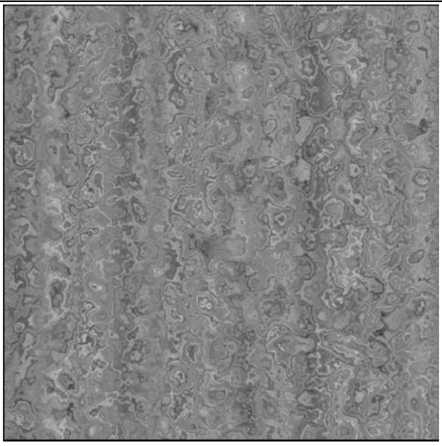
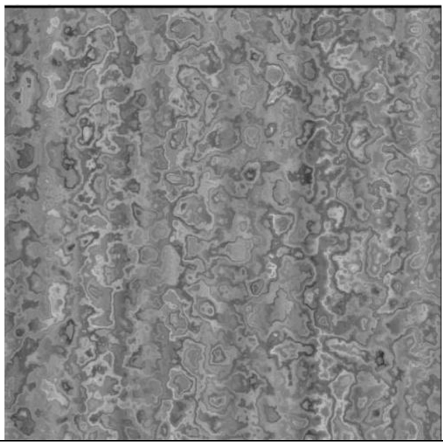
	norm: 19.15% norm_child: 0.04% schizo: 80.71% schizo_child: 0.09%
	norm: 4.98% norm_child: 0.00% schizo: 95.01% schizo_child: 0.01%

Figure 11. A synthetic pattern generated using an iterated Perlin noise function with non-linear trigonometric distortion, resulting in a classification of **95.01% for the 'schizo' class** by the neural network.

Conclusion

This study has demonstrated that modern generative Artificial Intelligence is capable of replicating the complex visual signatures of psychopathological art with high fidelity. By utilizing four distinct Large Language Models—Google Gemini, ChatGPT, Grok, and Qwen—we confirmed that the simulated "schizophrenic" aesthetic is not a byproduct of a specific algorithm but a consistent convergence of visual tropes across the current AI landscape.

Our specialized classifier, trained on a stratified dataset of human-made art (**norm**, **norm_child**, **schizo**, and **schzio_child**), successfully identified what we define as the **"Pattern of Madness"**. This pattern consists of specific, identifiable structural and symbolic anomalies inherent in the art of the mentally ill. Our neural network proved capable of capturing this subtle signature, identifying AI-generated images as belonging to the "schizophrenic" class with confidence scores reaching 75%. This suggests that the visual grammar of mental fragmentation is sufficiently structured to be mathematically distilled and reproduced by neural networks.

In conclusion, we have identified a new frontier in the intersection of AI and mental health. Future research should focus on developing methods to distinguish between authentic human expressions of psychological distress and high-fidelity algorithmic simulations, ensuring that the human element remains at the center of digital diagnostics.

The Telegram bot, identified as **@ShizoDetector_bot**, serves as a demonstration platform for the model's high specificity in distinguishing between normative and psychopathological visual signatures in both human-made and AI-generated art.

doi:10.20944/preprints202604.0887.v1

References

1. Wright, A. M. (2014). Schizophrenic art's influence on the identification of the disorder. [Honors research project]. Bowling Green State University.
2. Prinzhorn, H. (1922). Artistry of the mentally ill: A contribution to the psychology and psychopathology of configuration (Original German title: Bildnerei der Geisteskranken)
3. Jenkins, J. (2015). Art in the abyss: Creativity and the schizophrenic mind. Cambridge Core
4. Ivanova, A. Yu., Ivanov, I. I. (2016). Interpretation of drawings of people suffering from schizophrenia. Praxema: The Journal of Visual Arts and Art Education, (4), 2016.
5. Kit, O. (2019). Using deep neural networks for painting study for determine mental abnormalities, linked with schizophrenia (preliminary data). 10.13140/RG.2.2.30384.00008.
6. Kit, O. (2019) Investigation of children's paintings to determine possible mental abnormalities using a deep neural network. 10. 10.13140/RG.2.2.16945.02405/1
7. Di Leo, J. (2016). Children's drawings as diagnostic aids. London: Routledge.
8. Global Children's Art Gallery - Children's drawings, paintings, greeting cards, prints and more..[online] Available at:
9. Boldyreva, A.S. (1974) Risunki detey doshkol'nogo vozrasta, bol'nykh shizofreniey [Drawingsof preschool children with schizophrenia] Moscow: Meditsina