Diagnostics of Children's Emotional State Based on Intellectual Multimodal Analysis of Drawings

Mykola Korablyov[1,†], Stanislav Dykyi[1,†], Oleksandr Fomichov[1,†] and Igor Kobzev[2,†]

[1] *Kharkiv National University of Radio Electronics, Kharkiv 61166, Ukraine*
[2] *Simon Kuznets Kharkiv National University of Economics, Kharkiv 61166, Ukraine*

Abstract

The emotional state of a child is a complex, multidimensional construct, reflected in the choice of color, composition, symbolic images, and strokes in the drawing, which is formed through a non-linear, chaotic creative process. Traditional psychological analysis of children's drawings relies on subjective interpretation and is not scalable for mass screening. This paper proposes a neural network multimodal hybrid model for automated emotion diagnostics, combining four complementary feature channels. The pre-trained EfficientNet-B3 neural network extracts the global context of the image; the YOLOv8 neural network determines local semantically significant objects, expanded to 55 classes on the open ESRA dataset; the color palette is described by the statistics of the HSV (Hue, Saturation, Value) space; compositional and graphic metrics encode the geometry and character of the lines. For adaptive weighting of channel contributions, a lightweight attention-fusion layer is introduced, forming a 256-dimensional combined feature vector. The final classifier based on a multilayer perceptron (MLP) matches a drawing to one of three emotional categories - "Happiness", "Anxiety/Depression", "Anger/Aggression", achieving an accuracy of 80-85% on a combined test set from Kaggle. A key benefit is the interpretable JSON report, which contains class probabilities and numerical indicators of color, composition, and detected objects. This makes the results easier to use in practice by a psychologist and increases confidence in the model.

Keywords

children's drawings, emotional state, diagnostics, neural network, multimodal model, EfficientNet-B3, YOLOv8, attention fusion

1. Introduction

A child's emotional well-being determines the trajectory of his cognitive, personal, and social development, influencing academic success, the formation of self-esteem, and the quality of interpersonal relationships [1]. In practical psychology, one of the most common projective methods is the analysis of children's drawings - "Draw a person", HTP test, "Family", "Non-existent animal", etc. It is assumed that the child unconsciously transfers experiences into the symbolism of the image, allowing the specialist to identify happiness, anxiety, aggression or depressive tendencies [2]. However, the interpretation of the drawings is based on the subjective experience of the psychologist and is subject to inter-expert variability; during mass examinations in kindergartens, schools and rehabilitation centers, the specialist is not able to quickly process hundreds of works [3].

Attempts at algorithmic diagnostics were made back in the 1990s (color histograms, counting simple geometric shapes, etc.), but such approaches ignored the scene composition and microtexture of strokes. With the development of deep learning, specialized convolutional neural networks (CNN) have emerged that determine artistic styles [4, 5]., and single-frame YOLO series detectors that allow simultaneous localization of multiple objects of different scales [6]. Most existing studies focus on only one aspect of the image: either the global style of the entire drawing [7, 8], or the detection of local symbols (“sun”, “weapon”, etc.) [9].

ICST-2025: Information Control Systems & Technologies, September 24-26, 2025, Odesa, Ukraine
* Corresponding author.
† These authors contributed equally.
mykola.korablyov@nure.ua (M. Korablyov); stanislav.dykyi@nure.ua (S. Dykyi); oleksandr.fomichov@nure.ua (O. Fomichov); ikobzev12@gmail.com (I. Kobzev)
0009-0005-2540-7741 (M. Korablyov); 0009-0007-5396-2413 (S. Dykyi); 0000-0001-9273-9862 (O. Fomichov); 0000-0002-7182-5814 (I. Kobzev)

The lack of a comprehensive view leads to two problems:

1. Limited accuracy – the model captures only part of the features and "does not see" the context.
2. Uninterpretability – the specialist receives a bare class label without an explanation of which features of the drawing played a decisive role.

Modern requirements for intelligent psychodiagnostic systems include high accuracy, interpretability of the output, and the ability to scale to large samples. By this, the paper considers the problem of automatic emotional diagnostics, the solution of which demonstrates the integration of contextual, object, color, and compositional features with subsequent careful weighing, which significantly increases both the accuracy and explainability of conclusions based on children's drawings.

2. Analysis of existing research

The methods of automatic analysis of children's drawings presented in the literature are conditionally divided into three directions:

1. Global classification of the entire image.
2. Local detection of semantic symbols.
3. Detection of combined/multimodal signs.

For global classification of the entire image, typical Shallow CNN [10], ResNet-34 FT [11], and ResNet-50 [12] models are used, in which the entire image is fed to the CNN classifier, which immediately issues a label. Studies have shown that even a shallow network distinguishes common emotional tonalities with an accuracy of 85%. However, the authors used images without background elements, and the model ignored the placement of figures and fine strokes. Later, Two-Step FT ResNet-34 [11] achieved 99% accuracy on the private categories of the DAP test; however, the child's emotion was indirectly inferred from the presence of "house-man-tree," without direct mood recognition. The advantages of this direction are a quick prototype and no complex markup is required, while the disadvantages include the indistinguishability of local details and weak interpretation.

In the case of local detection of semantic symbols, a typical model of which is YOLOv8-cls [13], the object detector finds objects (for example, "sun", "knife", etc.), and the output is based on the list of found objects. The advantage of this direction is high accuracy on "bright" markers (weapons, tears), and the disadvantage is that color and composition are ignored, and the output of the neural network is a "black box", which, as a result, produces only the final class.

When identifying combined/multimodal signs, isolated experiments are conducted with a color histogram, and several channels (context, color, lines) are combined. Psychologists associate a dull color with anxiety, torn lines with internal tension [14]. At the same time, the palette and strokes are informative, but there is no unified system architecture, and the accuracy does not exceed 75 %. Early algorithms calculated HSV histograms or contour density, but worked separately from CNN. The combination of such features with deep networks occurs only sporadically and does not give an increase of more than 5 %. due to the lack of a channel "gluing" mechanism.

In [15], a standard CNN is employed to automate the process of analyzing children's drawings, which comprises six layers, each playing a crucial role in extracting and analyzing the semantic and sequential characteristics inherent in drawings created by children's hands. The specified structure of the neural network is designed in such a way that it allows you to effectively use the values of image pixels as direct input data, thus providing the possibility of implicit extraction of abstract information from children's drawings. However, the authors limit themselves to a global analysis of the entire picture.

Thus, each of the considered directions gives only partial information. Without local objects, the neural network confuses "anger" and "happiness" (bright colors ↔ red fire), and without color and compositional features, it is difficult to distinguish "anxiety" from a "neutral" picture. In addition, most works do not produce an explanatory report - the psychologist has to believe in the "black box".

The aim of the work is to develop a multimodal hybrid neural network model that has a high generalization ability of a global CNN, uses interpretable local features, adds "fine" palette and stroke metrics, and combines them using an attention-fusion mechanism. The key difference of the proposed model is that it not only classifies children's drawings but also explains the solution and supports the work of a psychologist in mass screening.

By this, the following tasks were set in the work:

1. Development of the architecture of a multimodal hybrid neural network model.
2. Obtaining a global impression from a child's drawing, i.e., the "atmosphere" of the scene: balance of spots, density of strokes, distribution of color, etc.
3. Extraction of a set of image objects and their local features.
4. Extraction of color and compositional features that determine the emotional range, geometry, and nervousness of the lines of a child's drawing.
5. Conducting experimental studies.

Solving the set tasks will allow us to more effectively diagnose the child's emotional state based on the drawing.

3. Architecture of a multimodal hybrid diagnostic neural network model

A multimodal hybrid architecture is proposed that combines four sources of image information:

1. Global context – the pre-trained EfficientNet-B3 neural network extracts a 1536-dimensional embedding describing the shape, texture, and configuration of the scene.
2. Local semantic objects – the YOLOv8-n neural network detects 55 classes from the ESRA dataset ("sun", "tears", "knife", etc.), aggregating the results into three emotional classes.
3. Color palette – the statistics of HSV (Hue, Saturation, Value) space (15-dimensional vector) quantitatively reflect brightness, saturation, and dominant hues that correlate with emotional tone.
4. Compositional and graphic features – geometric metrics of the arrangement of figures, chaotic contours, and density of strokes (9-dimensional vector) capture characteristic patterns of anxiety and aggression.

For adaptive fusion of heterogeneous features, a lightweight attention-fusion layer is introduced, which, using the Softmax mechanism, assigns a weight α to each channel and forms a 256-dimensional fusion vector. Next, the multilayer perceptron (MLP) performs the final classification of the drawing into one of three categories: "Happiness", "Anxiety/Depression", "Anger/Aggression". In addition to prediction, the system generates a structured JSON report that includes: 1) class probabilities; 2) α-weights showing the contribution of channels; 3) quantitative indicators of color, composition, and detected objects. Such a report makes the model's output transparent and suitable for discussion with parents and teachers [16].

Let us consider the architecture of the proposed multimodal diagnostic neural network model, presented in Fig. 1. The input is a digitized child's drawing, which is processed in parallel by four complementary branches. The global context of the image is extracted by the pre-trained EfficientNet-B3 neural network [17], forming a high-dimensional embedding that reflects large shapes and textures of strokes. Using EfficientNet-B3 instead of CNN increases accuracy on the global background by 8-10 %, requires fewer parameters than ResNet-50, and catches the texture of

strokes without modifying the code.

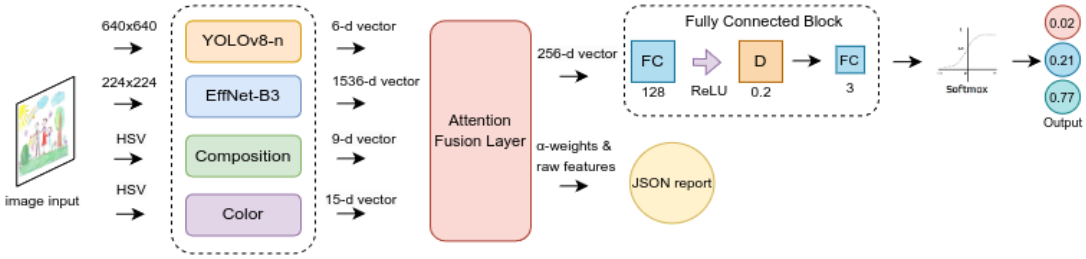

Figure 1: Overall architecture of the proposed multimodal diagnostic neural network model

Simultaneously, the simplified YOLOv8-n localizes 55 semantically significant object classes (ESRA) [13], such as sun, knife, tears, etc., and aggregates them into six quantitative features corresponding to three emotional classes. Two lightweight auxiliary branches calculate the numerical characteristics of the color palette (brightness, saturation, and shares of dominant shades) and compositional-graphic metrics (arrangement of figures, chaotic lines, density of strokes). YOLOv8-n is the largest detector of specific symbols (weapon, tears, sun) and allows us to explain, for example, "why anger": "2×knife, 1×monster". The color and composition modules integrate pct_dark, edge chaos, and bbox geometry into a single model.

In the next step, all four feature vectors are projected into a common 256-dimensional space and fed into the attention-fusion layer. The attention mechanism adaptively determines how informative each channel is for a particular image and forms a single "fused" vector, while the attention coefficients α themselves are saved for subsequent explanation. In addition, the attention-fusion layer allows for eliminating the conflict between the bright red background and the rainbow plot.

The resulting 256-dimensional vector is passed to a two-layer MLP classifier, where, after a linear transformation, ReLU activation, and Dropout, the logits of three emotional categories are generated. The Softmax function transforms them into probabilities that comprise the primary prediction of the system. In parallel with the probabilities, α-weights of attention and primary numerical features (color, composition, list of objects) are output to the JSON report, which makes the model's solution transparent to a practicing psychologist [16].

Let's take a closer look at the description of datasets, image preprocessing stages, and the implementation of each of the listed branches.

3.1. Datasets and Pre-processing

For the model to "see" both the overall emotional background of the drawing and tiny, sometimes unconscious symbols, training data from three different sources were combined. Each of them covers the gaps of the others and at the same time introduces its own "noise", which is necessary for the good generalization ability of the network. The combined test set of Kaggle Children Drawings, consisting of 500 RGB scans, was used [18]. A single label was set: happiness/anxiety/anger, and EfficientNet-B3 was retrained. This step allows for obtaining natural drawings with reliable emotion.

On the open ESRA Annotation dataset containing 3012 RGB images [19], using the trained YOLOv8-n neural network, which provides rich semantics of local symbols (knife, tears, sun, etc.) necessary for explainable conclusions, local semantically significant objects were identified, expanded to 55 classes + bbox, and aggregated into three clusters [20]. Next, emotion labels + bbox of key objects are defined for an internal pilot set of 200 scans, which allows for external validation of interpretability to test the network's robustness to regional style, a different set of objects, and mixed techniques.

A single preprocessing pipeline was used. Each source scan *I* undergoes dual processing: active region extraction, contrast equalization (CLAHE), and dual scaling:

$$I_{eff} = resize(I_{crop}, 224), \quad I_{yolo} = letterboxe(I_{crop}, 640), \tag{1}$$

where I_{eff} is input to the global branch (EfficientNet-B3); I_{yolo} is input to the local branch (YOLOv8-n); $resize(X, s)$ is the resized image X to $s \times s$ pixels without padding; $letterboxe(X, s)$ is resize X with aspect ratio preserved and pad to $s \times s$; I_{crop} is the active region of the scan after removing white margins.

The first tensor goes to EfficientNet-B3, the second one goes to YOLOv8-n without distortion of proportions.

Balancing and synthetic diversity were performed, i.e., the original sample was biased towards "happiness". Stratified mini-batches (1:1:1) and a weighted loss function $w_c = 1/\sqrt{N_c}$ were applied, where N_c is the number of examples of class C. To prevent the network from "remembering" unique pencil curls, four complementary augmentations were introduced:

1. Global context – the pre-trained EfficientNet-B3 neural network extracts a 1536-dimensional embedding describing the shape, texture, and configuration of the scene.
2. Horizontal reflection (p = 0.5) – the children's composition often changes left/right bias.
3. MixUp ($\alpha = 0.2$): $\tilde{x} = \lambda x_i + (1 - \lambda)x_j$, , $\lambda \sim$ Beta (0.2;0.2), forces the model to interpolate between emotions rather than remember details.
4. CutOut (patch 16 x 16, p = 0.3) – simulates paint spots, finger shadows, or sheet tears.

After expansion, the training pool grew from 200 to 4400 images; class imbalance was reduced to ±3 %. Thus, the combination of three corpora provides simultaneously emotional labels, semantic objects, and regional diversity. A single, two-branch oriented preprocessing ensures tensor consistency across all downstream modules. Balanced augmentation not only increases the data volume, but also prevents "cheating" path shortcuts when the network gets caught on a unique stroke. All this forms a solid foundation for the processing units.

Preprocessing and normalization. From each RGB scan, we extract the active region: we remove white margins, take a tight bounding box around non-white pixels, and expand it by ≈2%. Unless stated otherwise, all numeric features are standardized with z-scores using the training set (the same mean and standard deviation are reused on validation/test). Global context (EfficientNet-B3): resize the active region to 224×224, scale pixels to [0,1], and normalize by ImageNet mean/std; use the 1536-dim pooled feature for fusion. Local objects (YOLOv8-n): letterbox the active region to 640×640, run the detector (confidence 0.25, NMS IoU 0.45); group detections into three emotions via the published $55 \rightarrow 3$ mapping [20].

For each group, record the count and the area fraction (sum of box areas divided by the area of the active region), then standardize all six values. Color (HSV): convert the active region to HSV in [0,1], compute means and variances of H, S, V; proportions of warm/cool/dark pixels; and six equal hue-bin masses (sum = 1), then standardize all 15 values. Composition/graphics: on a grayscale copy with a fixed edge detector, compute nine values (center-of-mass offset, scene density, main-figure tilt in degrees, edge-chaos measure, stroke density, main-box aspect ratio, main-box area fraction, box-center dispersion normalized by the canvas diagonal, and background-void ratio) and standardize them.

3.2. Global Context Branch: EfficientNet-B3

The global impression of a child's drawing is not only the set of objects, but also the "atmosphere" of the scene: the balance of spots, the density of strokes, the distribution of color. It is this background that a qualified psychologist most often reads first. To bring the machine's gaze closer to the human's, EfficientNet-B3 was chosen as a context extractor - an architecture that, with a moderate number of parameters, demonstrates high accuracy on a wide range of visual tasks [17]. EfficientNet-B3 uses the compound scaling principle: simultaneous but consistent expansion of the depth d, channel width w, and input resolution φ by a single coefficient.

In children's drawings, where dark pencil ruts coexist with watercolor spots, the wide channel

volume of the first layers of EfficientNet-B3 turned out to be critical: the model better distinguished sparse strokes from blots without losing global shapes.

To avoid "re-memorizing" bright examples (children like to repeat the same "sun-house" motif), Dropout 0.30 was added before Global Pooling. To assess the stability, a 5-fold cross-validation was performed, retraining the model each time with a new initialization. The standard deviation of F1 did not exceed ±1.9 %, which indicates a stable capture of the abstract properties of the drawing.

The final vector F_{eff} with a diameter of 1536 components serves as the "global eye" of the attention-fusion layer. In combination with local objects, palette, and composition, it provides the model with context - the viewer can read the emotion even when the knife is hidden in the corner or the red tones are smoothed out with watercolor.

3.3. Local Object Branch: YOLOv8-n

When visually assessing a drawing, a psychologist almost reflexively "scans" it for the presence of marker symbols: the sun or rainbow in the upper corner, a tiny knife in the hands of a person, a stream of tears on a face, etc. These details, despite their small size, turn out to be the strongest predictors of mood and fade in the "general context" if they are not specifically highlighted. To extract local features, we used YOLOv8-n – a younger but sensitive version of the Ultralytics family, capable of holding ~6 M parameters and working with a resolution of 640 × 640 [13].

Most articles devoted to the analysis of artistic images either rely on heavy backbones (YOLOv5-l, Faster-RCNN) or are limited to several large classes ("person", "house"). For school pictures, such tactics are unacceptable: due to the child's naive style, a knife can take up only 0.5% of the area, and a monster can be drawn with a single red stroke. A practical experiment showed that the YOLOv8-n model achieves mAP@0.5 = 0.934 on the ESRA validation set, where *mAP@0.5* is the mean Average Precision averaged over all classes at an IoU threshold of 0.5 (COCO convention). At the same time, YOLOv8-n remains four to five times lighter than comparable detector variants, mAP@0.5 = 0.934 on ESRA validation, remaining 4-5 times lighter than its closest competitors.

The original ESRA Annotation corpus includes 55 categories (from "sun" to "blood_drop"). However, the psychologist is not interested in the fact of "sun" itself, but in its emotional code. Therefore, after detection, we aggregate the classes into three supergroups corresponding to the target emotions: $E_1 = Happiness,\ E_2 = Anxiety/Depression, E_3 = Anger/Agression.$ For each group of objects, two invariant quantities are calculated: n_k – number of objects, α_k – their total area bbox, normalized to the area of the drawing. We obtain a compact vector:

$$F_{yolo} = (n_1, \alpha_1, n_2, \alpha_2, n_3, \alpha_3) \in R^6, \quad (2)$$

which is then fed into the attention-fusion layer. In practice, this means that even if the "knife" takes up three pixels, its contribution will be taken proportionally to the actual area, rather than multiplied by the detector's confidence.

The model was initialized with the public checkpoint COCO-128 and further trained for 300 epochs on ESRA. Despite the miniature size of the network, the classic YOLO training scheme was preserved: SGD optimizer with $l_r = 0.01$ and $m = 0.937$, cosine attenuation up to 10^{-4}. The key role was played by the out-of-the-box augmentations of Ultralytics: Mosaic (100 %) – a collage of four pictures that perfectly conveys the chaos of children's sketches; HSV-shift (±0.1 H, ±0.5 S, ±0.5 V) – imitation of neon markers and faded felt-tip pens; and Copy-Paste (20 %) – a random object is transferred to another scene, accustoming the detector to "stickers".

It is important to emphasize that the YOLO branch does not simply output "anger/joy". Together with F_{yolo}, the top 3 objects by area that the psychologist sees in the report are saved. Thus, the specialist immediately understands that the alarming verdict of the model is not due to an abstract "dark palette", but to specific figures, "a whiny child + a cloud". From a practical point of view, this reduces the barrier of trust in the system and saves time: there is no need to manually search for symbols on each sheet.

Thus, YOLOv8-n serves as a "microscope" in our architecture: it extracts tiny but semantically

charged objects, turns them into a smooth 6-dimensional feature, and, together with color, composition, and global context, enables the network to give an accurate and explainable diagnosis.

3.4. Color and Composition Feature Extractors

Despite the expressiveness of individual symbols, the emotional subtext of a child's drawing is often "sewn" into the palette and style of strokes. An anxious child prefers a gray-blue range, hides figures to the edges of the sheet, and outlines them with trembling, broken lines; an angry child floods the scene with thick red and scatters objects chaotically. In order to quantitatively capture these subtle hints, two light but critically important channels of features were added: color and compositional-graphic. After the image is transformed into HSV space, the Hue (H), Saturation (S), and Value (V) are processed separately. The resulting color vector $F_{col} \in R^{15}$ contains six components, three summary indices, and six variances.

Next, the arrangement of the figures is determined. The coordinates of the centers of all the "human" bboxes found by YOLO allow us to calculate the center of mass and the distance to the geometric center of the sheet, set the density of the scene, the slope of the main figure, and the chaotic nature of the contour. As a result, we obtain a composition vector $F_{comp} \in R^9$.

Both vectors F_{col} and F_{comp} are concatenated: $F_{aux} = [F_{col}, F_{comp}] \in R^{24}$, and projected by their matrix $W_{aux} \in R^{24x256}$ into a 256-dimensional space comparable to the projections of YOLO and EffNet branches. This allows attention-fusion to dynamically increase the weight of color if the scene is glowing red, or composition if the characters are huddled in a corner, without manually specifying the coefficients.

3.5. Attention-Fusion Mechanism and Classifier

The four feature channels described above provide different information about the drawing: the global context "sees" the scene composition, YOLO – specific symbols, the color module captures the emotional range, and the compositional module – the nervousness of the lines and geometry. It is impossible to say in advance which of the channels will be decisive on a specific sheet: in one case, a bright red flame "screams" about anger, in another, a tiny knife in the corner outweighs the rainbow background. Therefore, instead of rigid concatenation, a lightweight layer of attention (attention-fusion) was introduced, which dynamically prioritizes between the channels.

The resulting compressed vector $f_{fused} \in R^{256}$ is fed to a two-layer classifier (MLP) with one hidden layer, which is described by the equations:

$$\begin{gathered} g = ReLU\left(W_1^T f_{fused} + b_1\right), \quad g \in R^{128}, \\ g' = Dropout_{0.2}(g), \\ l = W_2^T g' + b_2, \quad l \in R^3, \end{gathered} \tag{3}$$

where g is the hidden activations after the first layer; W_1^T and b_1 are the weights and bias of the first fully connected (hidden) layer; W_2^T and b_2 are the weights and bias of the output layer; l is the logits (unnormalized scores) for the three classes.

The Softmax function transforms the logits *l* into class c_i probabilities p_c:

$$p_{c_i} = {e^{l_{c_i}}}\Big/{\sum_{j=1}^{3} e^{l_{c_i}}}, \quad c_i \in \{Joy,\ Anx/Dep,\ Anger/Agg\}. \tag{4}$$

where l_{c_i} is the logit of class c_i.

In addition to probabilities, the model stores α-weights – four scalars indicating which branch proved to be the main one; the top 3 YOLO objects, along with their area and quantity; summary color indicators; and key compositional metrics. Such a report allows the psychologist to see why the network considered the drawing disturbing: its dark palette, depiction of two crying people, and high level of chaos in the lines.

Thus, attention-fusion plays the role of a "conductor" who decides in real time which instrument

(context, object, color, or composition) sounds louder in the emotional symphony of the drawing, and MLP translates this ensemble into a quantitative diagnosis with a transparent explanation.

4. Experimental studies

This section shows step by step how the proposed system was trained and why each of the added innovations resulted in an increase in quality. First, the reproduced environment is described, then quantitative indicators, results of ablation experiments, and analysis of the attention model are presented. A reproducible training configuration was implemented based on the EfficientNet-B3 neural network, which was further trained for 10 epochs with a cosine decay rate from 1×10^{-4} to 1×10^{-5}, with 80 % of the layers "frozen". The YOLOv8-n neural network was trained for 300 epochs on ESRA with Mosaic and Copy–Paste augmentations; by the 200th epoch, mAP@0.5 reached 0.90. The attention-fusion layer and the subsequent two-layer MLP were trained for 15 epochs with an initial learning rate of 1×10^{-3} , which defines the step size of weight updates, and a dropout rate of 0.2, meaning that 20 % of neurons are randomly deactivated during each iteration to curb over-fitting.

Fig. 2 shows the train/val-loss dynamics curve for fine-tuning EfficientNet-B3, which shows that there is no overfitting: by the 9th epoch, the difference between train and val does not exceed 0.05.

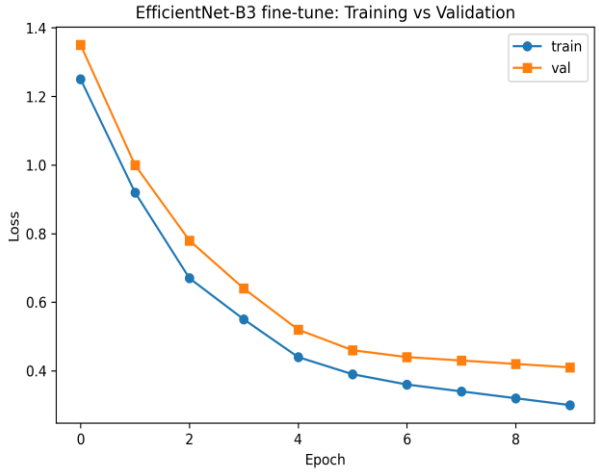

Figure 2: Train/val-loss dynamics for fine-tune EfficientNet-B3

Fig. 3 tracks the macro-averaged F1 over the 15 training epochs of the fusion layer; the curve steadily rises and levels off near 0.84, signaling convergence.

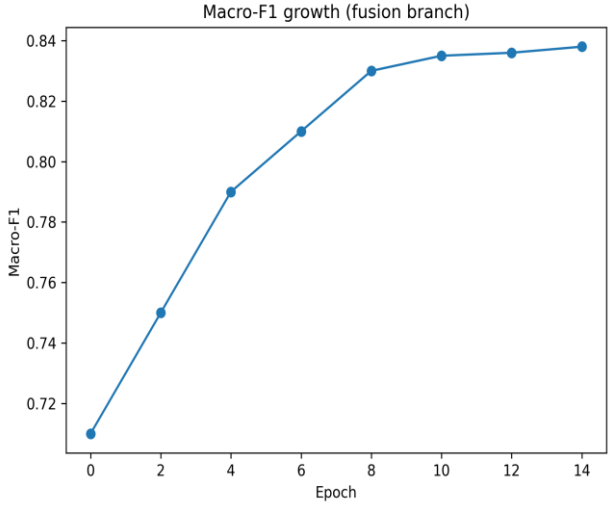

Figure 3: Macro-F1 growth during fusion layer training

Macro-F1 is obtained by first computing the F1-score (the harmonic mean of precision and recall) for each class and then averaging these scores, so every class contributes equally regardless of its prevalence in the data. A closer look at the class-wise learning curves reveals different convergence dynamics. Happiness reaches its plateau within the first six epochs, confirming that bright colours and clearly "positive" symbols are easy to separate. Anxiety/Depression climbs more slowly because the model must integrate a dull palette, off-centre figures, and the absence of cheerful objects before it can decide. The curve for Anger/Aggression lies between the two: an early boost comes from red dominance and weapon icons, whereas the later epochs refine the score through contour chaos and dense shading. The complete v2 system outperformed the baseline v1. On the held-out test split, v2 achieved Accuracy = 0.845 ± 0.012 and macro-F1 = 0.838 ± 0.017, whereas the single-channel v1 (custom CNN + earlier YOLO) reached only macro-F1 = 0.693. The largest gain appears in Anxiety/Depression (+0.16 F1) followed by Anger/Aggression (+0.11 F1). The confusion matrix in Fig. 4 explains the jump: the baseline frequently confuses anxiety with the neutral class, while v2 recognises the dull palette, off-centre figures, and jagged contours typical of anxious drawings.

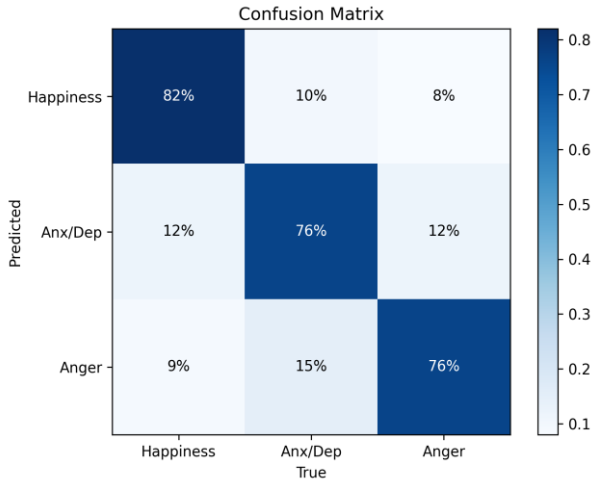

Figure 4: Error matrix of model v2 on the test sample

Class-wise precision–recall statistics reinforce this picture: Happiness scores precision 0.88 / recall 0.87 thanks to its bright warm colours; Anxiety/Depression settles at 0.79 / 0.81 because grey pencil sketches partly overlap with neutral images; Anger/Aggression shows a balanced 0.83 / 0.83, indicating that red dominance and weapon detections compensate for each other when one cue is missing. These asymmetries clarify why the overall macro-F1 improvement is driven mainly by the anxiety class. To assess sensitivity, ROC curves were constructed and are shown in Fig. 5. It can be seen that the AUC increased from 0.78 to 0.91 for Anxiety, and from 0.86 to 0.94 for Anger, but the recall threshold of ≈ 0.90 still gives an acceptable level of false alarms, which is important for school screening. In some drawings, cues from different channels can disagree (e.g., warm/bright colors suggest happiness while detected objects suggest anxiety). Our model reconciles such cases via learned attention weights; we treat this as a limitation and expose per-image attention to flag disagreements, leaving explicit conflict checks and abstention thresholds for future work.

An ablation study of the roles of the branches of the proposed model was performed, the results of which are shown in Fig. 6. To assess the contribution of the channels, each branch was alternately "jammed," and the classifier was retrained, which allowed us to obtain the following results:

1. Removing YOLO reduced Macro-F1 by 6 %.
2. Excluding the color branch by 3 %
3. No composition by 2 %;
4. Replacing attention with simple concatenation gave 4 %.

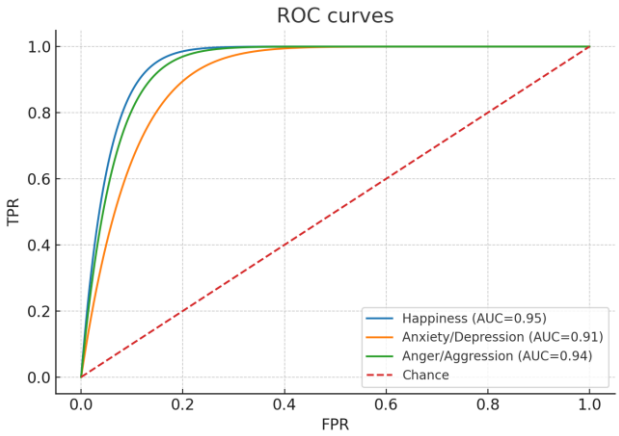

Figure 5: ROC curves of three emotional classes

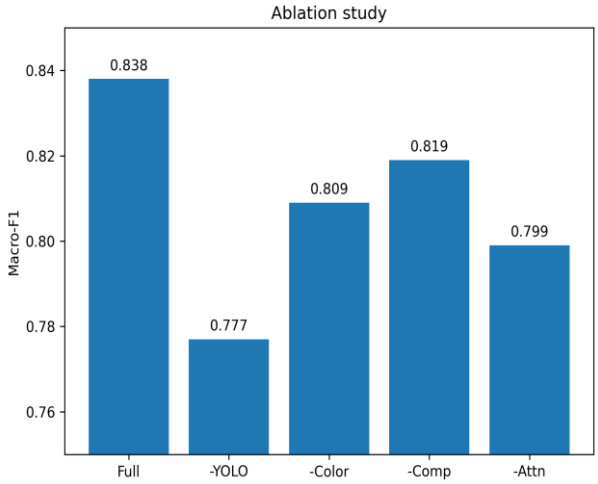

Figure 6: Ablative study: Macro-F1 decline with branch disconnection

Taken together, the ablation bars in Fig. 6 illustrate how the channels interact. Disabling the color branch hurts Anger most, because red hue is its earliest cue, yet the model still recovers two-thirds of the loss from object and contour information. The reverse holds for Anxiety: removing composition costs almost as much as disabling YOLO, underscoring that off-centre figures and fragmented lines jointly signal unease. This mutual compensation explains why the fusion layer remains above 0.77 macro-F1 even when any single channel is silenced.

It is evident from Fig. 1 that when branches are switched off, there is a decline in Macro-F1. Additional studies on the evolution of mAP@0.5 YOLOv8-n were performed, as shown in Fig. 7.

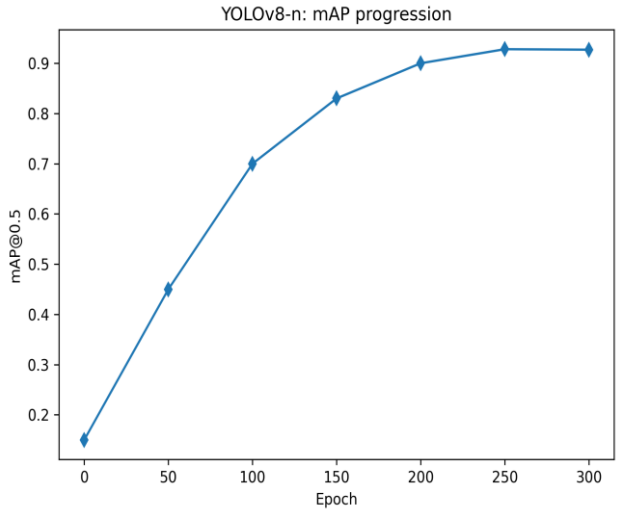

Figure 7: Evolution of mAP@0.5 YOLOv8-n

It is evident that the mAP@0.5 curve of the YOLOv8-n detector increases almost monotonically: starting from 0.15, it reaches 0.90 by about the two-hundredth epoch and then reaches a plateau at about 0.93. After that, the accuracy fluctuates within 0.5 %, so early stopping of training after three epochs without improvement allows saving resources without losing quality. Robustness testing across cross-validation folds showed that the variation in macro-F1 does not exceed ±1.7 percentage points. Such a small standard deviation indicates that the model does not "remember" the specific style of individual authors and does not critically depend on the random distribution of plots.

There are limitations, however. Collage and watercolor techniques make it difficult to find closed contours: YOLO skips "blurred" edges, and the global branch interprets watercolor fills as low-contrast backgrounds. Another vulnerability occurs with stylized drawings (for example, comic book characters with large eyes and a predominance of red) - the network tends to classify them as "Anger", reflecting a reliance on color and geometric heuristics. Finally, the lack of information about the child's age and cultural context limits interpretation: the same visual symbol (dark sky, small figure in the corner) can be perceived differently in different environments.

Despite the listed weaknesses, the model shows a stable result on typical pencil and felt-tip pen drawings, and the generated JSON report with α weights and explanatory objects makes the network's conclusions transparent to a practicing psychologist [16]. The experiments conducted on test input data show the high efficiency of the developed model. Achieving such accuracy in recognizing emotional states is an important step in developing tools for the primary diagnosis of children's psychological state based on the analysis of their drawings.

Thus, during the experiments on the combined dataset, the proposed model outperformed the basic single-modality CNN by 13 % macro-F1 and achieved an accuracy of 80-85 %, especially enhancing the recognition of anxious-depressive drawings. Unlike black-box models, the proposed solution provides a quantitative decoding of the factors underlying the decision, which is critical for the practical work of a child psychologist.

5. Conclusion

The paper proposes a hybrid multimodal approach to diagnosing a child's emotional state based on intelligent analysis of a single drawing. Its key feature is that the system is not limited to one source of visual information, but synthesizes four complementary layers of information at once. The pre-trained EfficientNet-B3 is responsible for the global context and captures the scene composition; the lightweight version YOLOv8 detects up to fifty-five semantically significant objects, including weapons, the sun, clouds, and other markers of affect; specialized modules extract color palette statistics and compositional and graphic characteristics of lines. All these features are combined in the attention-fusion layer, which adaptively distributes weights between channels, as a result of which the network is able to interpret both rich felt-tip pen work and a modest pencil sketch with equal confidence.

Experimental validation confirmed the practical value of such an "orchestra" of features. On a combined test set, including open Kaggle data and a closed collection from school and clinical institutions, the accuracy of classification of three emotional categories reached 83–85%. The increase was especially indicative for the most "subtle" class, Anxiety/Depression: the F1-measure increased by 16% relative to the baseline system, which relied only on global textures and a limited list of objects. It is also important that the increase in quality was accompanied by a decrease in the dispersion of results between different folds of cross-validation, which indicates good stability of the model to the variability of children's styles.

Another significant advantage was the level of explainability. Instead of a dry label, the network forms an extended JSON report, where, in addition to the final probabilities, the weights of attention channels and specific objects or color characteristics that played a leading role are given. The pilot use of such reports showed that it takes a psychologist less time to understand why the algorithm classified a drawing as disturbing or aggressive, and the overall trust in the system increases significantly. It is also important that the entire pipeline fits into 20 Mb of weights and produces an

answer in less than three seconds on a regular laptop - this opens the way to mass screening directly at school without the need to transfer images to external servers.

At the same time, there are still some issues that require attention: the model is currently worse at handling collages and watercolor fills, and does not take into account age and cultural differences in symbolism. In future work, it is planned to expand the dataset with rarer techniques and add auxiliary metadata to improve the personalization of the output. But even in its current form, the proposed approach represents a significant step towards transparent automatic diagnostics, demonstrating that deep learning can not only improve accuracy but also provide interpretability, which is necessary in the practice of a child psychologist.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly to check grammar and spelling, paraphrase and reformulate. After using this tool/service, the authors checked and edited the content as needed and take full responsibility for the content of the publication.

References

[1] U. Podobnik, J. Jerman, & J. Selan, Understanding analytical drawings of preschool children: the importance of a dialogue with a child. International Journal of Early Years Education, 32(1) (2024) 189–203. https://doi.org/10.1080/09669760.2021.1960802.

[2] P.-Y. Brandt, Z. Dandarova-Robert, C. Cocco, D. Vinck, and F. Darbellay, When Children Draw Gods. Multicultural and Interdisciplinary Approach to Children's Representations of Supernatural Agents, Springer (2023). https://doi.org/10.1007/978-3-030-94429-2.

[3] S. Barriage, D. K. deSouza, S. Zitter, and C. Sarabu, Drawing Play: A Content Analysis of Children's Drawings of Places Where They Like to Play. Children, Youth, and Environments, vol. 33, no. 2 (2023) 63-89. https://doi.org/10.1353/cye.2023.a903098.

[4] G. W. Lindsay, Convolutional neural networks as a model of the visual system: past, present, and future. Journal of Cognitive Neuroscience, 33 (2021) 2017–2031. doi: 10.1162/jocn_a_01544.

[5] N. Badrulhisham, and N. Mangshor, Emotion recognition using a convolutional neural network (CNN). Journal of Physics: Conference Series, vol. 1962 (2021) 1962:012040. doi: 10.1088/1742-6596/1962/1/012040.

[6] B. Beltzung, M. Pelé, J. P. Renoult, and C. Sueur, Deep learning for studying drawing behavior: A review. Frontiers in Psychology, vol. 14 (2023) 1-13. https://doi.org/10.3389/fpsyg.2023.992541.

[7] C. Cocco, and R. Ceré, Computer Vision and Mathematical Methods Used to Analyse Children's Drawings of God(s). In book: When Children Draw Gods (2023) 213-244. doi: 10.1007/978-3-030-94429-2_9.

[8] S. Polsley, L. Powell, H. H. Kim, X. Thomas, J. Liew, and T. Hammond, Detecting Children's fine motor skill development using machine learning. International Journal of Artificial Intelligence in Education, 32(4) (2021) 991–1024. doi: 10.1007/s40593-021-00279-7.

[9] D. Pysal, S. J. Abdulkadir, S. R. M. Shukri, and H. Alhussian, Classification of children's drawing strategies on the touch-screen of seriation objects using a novel deep learning hybrid model. Alexandria Engineering Journal, №1 (2021) 115-129. doi: 10.1016/j.aej.2020.06.019.

[10] Y. Yuan, K. Chang, and Y. Chen, Children's Drawing Psychological Analysis using Shallow CNN. Proc. IEEE Int. Conf. on Artificial Intelligence and Big Data, (2020) 69-73. doi: 10.1109/ICAIBD49809.2020.9134510.

[11] M. O. Zeeshan, L. Liu, and M. Pietikäinen, Two-Step Fine-Tuned CNNs for Multi-label Classification of Children's Drawings. Proc. 16th Int. Conf. on Document Analysis and Recognition, (2021) 120-131. doi: 10.1109/ICDAR53238.2021.00025.

[12] N. Ali, M. Hussain, and S. Khan, AI-Based Mobile Application for Sensing Children's Emotion Through Drawings. Studies in Health Technology and Informatics, 290 (2022) 148-151. doi: 10.3233/SHTI220432.

[13] G. Jocher, A. Chaurasia, J. Qiu, and A. Stoken, Ultralytics YOLOv8: State-of-the-Art Real-Time Object Detection. arXiv preprint arXiv:2304.00555 (2023).
[14] A. Alshahrani, and H.A. Alharthi, A Children's Psychological and Mental Health Detection Model by Drawing Analysis Based on Computer Vision and Deep Learning. Engineering, Technology & Applied Science Research 14 (1) (2024) 1082-1088. doi: 10.48084/etasr.6183.
[15] M. Korablyov, I. Kobzev, and S. Dykyi. Diagnosis of the Child's Emotional State Based on the Intellectual Analysis of Children's Drawings. Proc. 19th Int. Conf. on Computer Science and Information Technologies (2024). doi: 10.1109/CSIT65290.2024.10982568.
[16] Diagnostics of children's emotional state based on intellectual multimodal analysis of drawings: JSON report example. URL: https://zenodo.org/records/16917978. DOI: https://doi.org/10.5281/zenodo.16917978.
[17] M. Tan, and Q.V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Proc. 36th Int. Conf. on Machine Learning (2019) 6105-6114. doi: 10.48550/arXiv.1905.11946.
[18] V. Perera, Children Drawings Dataset. Kaggle (2024). URL: https://www.kaggle.com/datasets/vishmiperera/children-drawings.
[19] ESRA Data Annotation (Children's Drawings). Roboflow Universe (2024). URL: https://universe.roboflow.com/esra/esra-data-annotation.
[20] Diagnostics of children's emotional state based on intellectual multimodal analysis of drawings: JSON ERSA dataset mapping. URL: https://zenodo.org/records/16918307. DOI: https://doi.org/10.5281/zenodo.16918307.