POLITECNICO DI TORINO
Repository ISTITUZIONALE

FAITHAI: Explainable Deep Learning with Synthetic Data for Pediatric Facial Phenotype and Autism-Related Image Classification

Original
FAITHAI: Explainable Deep Learning with Synthetic Data for Pediatric Facial Phenotype and Autism-Related Image Classification / Shah, S.T.H., Shah, S.A.H., Baqir Hussain Shah, S., Pigueiras-del-Real, J., Khaled, R., Zamir, A., Godio, S., Muhammad Jawwad Riaz, S., Ahmad Qureshi, S., Ruiz-Zafra, A., Gontard, L.C., Deriu, M.A.. - In: JOURNAL OF IMAGING. - ISSN 2313-433X. - 12:11(2026), pp. 1-20. [10.3390/jimaging12110589]

Availability:
This version is available at: 11583/3018195 since: 2026-11-20T12:51:20Z

Publisher:
MDPI

Published
DOI:10.3390/jimaging12110589

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

28 November 2026

Journal of *Imaging*

MDPI

Article

FAITHAI: Explainable Deep Learning with Synthetic Data for Pediatric Facial Phenotype and Autism-Related Image Classification

Syed Taimoor Hussain Shah [1,*], Syed Adil Hussain Shah [1,2], Syed Baqir Hussain Shah [3], Janet Pigueiras-del-Real [4], Roa'a Khaled [5], Ammara Zamir [6], Silvia Godio [1], Syed Muhammad Jawwad Riaz [7], Shahzad Ahmad Qureshi [8], Angel Ruiz-Zafra [9], Lionel C. Gontard [4,10] and Marco Agostino Deriu [1,*]

1 PolitoBIOMed Lab, Department of Mechanical and Aerospace Engineering, Politecnico di Torino, 10129 Turin, Italy; syedadilhussain.shah@gpi.it (S.A.H.S.); silvia.godio@polito.it (S.G.)
2 Department of Research and Development (R&D), GPI SpA, 38123 Trento, Italy
3 Future Information Innovation Academy, Fudan University, Shanghai 200438, China; bhssyed26@m.fudan.edu.cn
4 Department of Condensed Matter Physics, University of Cádiz, 11510 Cádiz, Spain; janet.pigueiras@gm.uca.es (J.P.-d.-R.); lionel.cervera@gm.uca.es (L.C.G.)
5 Department of Computer Engineering, University of Cádiz, 11519 Puerto Real, Spain; roaa.khaled@gm.uca.es
6 Department of Computer Science, Ca' Foscari University of Venice, 30172 Venice, Italy; ammara.zamir@unive.it
7 Department of Mathematics, COMSATS University Islamabad (CUI), Wah Campus, Wah 47000, Pakistan; jawwadriaz@ciitwah.edu.pk
8 Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences, 45650 Islamabad, Pakistan; drsaqureshi@pieas.edu.pk
9 Department of Software Engineering, University of Granada, 18071 Granada, Spain; angelr@ugr.es
10 Instituto de Investigación en Microscopía Electrónica y Materiales (IMEYMAT), University of Cadiz, 11510 Cádiz, Spain
* Correspondence: taimoor.shah@polito.it (S.T.H.S.); marco.deriu@polito.it (M.A.D.)

Abstract

Pediatric facial image classification remains challenging due to the limited representation of rare conditions and the heterogeneity of available image sources. This study presents FAITHAI, an integrated framework combining facial region-of-interest standardization, StyleGAN-3-based class-conditional image synthesis, DenseNet-169 classification, and explainable artificial intelligence. The nine-class FACED3 classifier was evaluated on 221 held-out real images, achieving an accuracy of 0.733 (95% Wilson confidence interval: 0.671–0.787), balanced accuracy of 0.645, macro F1-score of 0.635, and macro one-vs-rest ROC AUC of 0.955. Across seven random-seed runs, the mean accuracy was 0.732 ± 0.034. A complementary binary classification analysis using real images only was conducted on 525 held-out images, achieving an accuracy of 0.836 and ROC AUC of 0.904 for the seed-11 model, with a mean accuracy of 0.817 ± 0.016 across ten runs. Quantitative Grad-CAM analysis on the nine-class test set yielded a deletion AUC of 0.217 compared with 0.471 for random deletion, with lower paired deletion AUC in 216 of 221 images. Attribution maps demonstrated high stability under small brightness perturbations (median correlation: 0.956) and substantial changes following classifier-head randomization (median correlation: 0.125). These tests characterize computational attribution behavior; the clinical and anatomical relevance of highlighted regions has not been established. Overall, FAITHAI provides an integrated approach to pediatric facial image classification by combining synthetic-image-assisted learning, quantitative performance evaluation, and explainable AI. The findings demonstrate its applicability to multiclass and binary facial

Academic Editors: Shengeng Tang, Lechao Cheng and Weidong Chen

Received: 5 October 2026
Revised: 9 November 2026
Accepted: 12 November 2026
Published: 19 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

image classification while providing complementary insights into model discrimination and attribution behavior.

Keywords: pediatric facial phenotype; synthetic images; DenseNet-169; exploratory classification; StyleGAN-3; model evaluation

1. Introduction

Facial image analysis is used in both affective computing and phenotype research [1,2]. In medical image classification, the central challenge is not simply to obtain a high accuracy value but to determine whether a model learns reproducible facial patterns rather than dataset-specific artifacts. This is particularly difficult when datasets are small, heterogeneous, or collected outside controlled clinical settings [3,4].

Facial expression recognition provides a useful auxiliary benchmark because it tests whether a backbone can learn general facial representations across changes in pose, expression, and imaging conditions. Classic work established standardized emotion categories [5], and later CNN studies showed that pretrained networks can learn useful facial features [6]. Public resources such as FER [7] have therefore become common test beds for comparing architectures. In the present study, emotion recognition is used only for representation benchmarking and transfer-learning experiments; it is not treated as a proxy for disease recognition.

Pediatric facial phenotype classification presents additional challenges because of the limited availability of images representing rare conditions, substantial variation in image acquisition, and differences in facial appearance across individuals [8]. Deep learning approaches involving facial region-of-interest standardization, generative image synthesis, and explainable artificial intelligence offer complementary strategies for investigating these challenges [9]. However, classification performance must be examined using real held-out images, while visual explanations require quantitative evaluation to characterize their relationship with model predictions [10].

CNN-based methods have also been applied to facial phenotype classification and autism-related image datasets [11,12]. These models can learn hierarchical visual representations from two-dimensional and three-dimensional facial data [13,14], but image-level classification should not be equated with clinical diagnosis. This distinction is especially important for autism, for which facial differences are subtle, heterogeneous, and not a standalone diagnostic criterion.

Machine-learning studies have explored supervised facial phenotype classification and unsupervised analysis of developmental data [15]. Related work also shows the value of combining multiple data modalities when studying autism and other neurodevelopmental conditions [16–18]. These studies motivate computational phenotype research, but they also show why model outputs must be interpreted within the limits of the dataset and task definition.

Three limitations recur across the literature in question. First, rare syndromes are difficult to represent with sufficiently large and diverse image collections [19,20], and annotation quality can vary across sources [21]. Second, deep models can exploit non-clinical cues unless preprocessing and evaluation are carefully controlled. Third, post hoc explanations can make model behavior easier to inspect, but they do not by themselves establish biological validity or clinical trustworthiness [22]. Related transformer-based gastrointestinal image classification has explored alternative feature representations [23], while adversarial super-resolution has been investigated for biomedical image enhancement [24]; neither study directly validates pediatric facial phenotype classification.

To investigate these challenges, this study presents FAITHAI, an integrated framework combining facial region-of-interest standardization, StyleGAN-3-based [25] image synthesis, DenseNet-169 classification, and explainable artificial intelligence for pediatric facial image classification. Model selection was informed by an auxiliary facial expression benchmark and comparative evaluation of deep learning architectures on FACED3 images. The framework was evaluated through nine-class facial image classification using held-out real images, complementary binary classification of autism-labeled and healthy images, and repeated-seed performance assessment. Furthermore, SHAP, LIME, Grad-CAM, and saliency-based explanations were examined alongside quantitative attribution diagnostics to investigate model interpretability and explanation behavior. The study provides a comprehensive evaluation of classification performance, reproducibility, and model attribution within the FAITHAI framework.

The main contributions of this work are as follows:

- We present FAITHAI, combining standardized facial region-of-interest extraction, StyleGAN-3-based synthetic image generation, DenseNet-169 classification, and explainable AI techniques for the analysis of nine pediatric facial image categories.
- We evaluate the FACED3 classifier on 221 held-out real images, presenting both overall and real-test confusion matrices, class-wise classification metrics, discrimination curves, confidence intervals, and variability across seven random-seed runs.
- We evaluate Grad-CAM explanations across all 221 real test images using deletion-based faithfulness, brightness-perturbation stability, classifier-head randomization, and border-sensitivity controls, complementing the qualitative SHAP, LIME, and Grad-CAM analyses.
- We examine binary classification between autistic-labeled and healthy facial images through retraining on real images only, with 525 held-out test images, ten random-seed runs, validation-based model selection, and ROC and precision-recall analysis.

Section 2 describes the data, facial ROI preprocessing, synthetic image generation, DenseNet-169 training, and evaluation procedures. Section 3 presents the FACED3 classification results, the real-image binary task, and the explainability analyses. Section 4 discusses the findings and limitations; further methodological details and diagnostics are supplied in the Supplementary Materials.

2. Materials and Methods

FAITHAI combines facial region-of-interest (ROI) extraction, StyleGAN-3 class-conditional synthesis, DenseNet-169 classification, and explainability methods. The nine-class FACED3 analysis reports classifier-development behavior and evaluation on held-out real images. A complementary binary autism-label analysis uses only real images, without GAN-generated samples.

2.1. Datasets

Facial expression datasets were used for auxiliary backbone benchmarking. The nine-class facial phenotype classification used FACED3, and the expression labels were not used as syndrome labels.

2.1.1. Facial Expression Datasets

Detailed descriptions of all facial expression datasets are provided in the Supplementary File under Section S2: Datasets, Subsection A: Facial Expression Datasets.

2.1.2. FACED3 Facial Image Collection

FACED3 contains nine facial image categories, including an autistic-labeled category. These are dataset labels for image classification and should not be interpreted as clinical

diagnoses. FACED3 includes images labeled autistic, Down syndrome, Emanuel syndrome, Kabuki syndrome, Noonan syndrome, healthy, Pallister–Killian syndrome, Williams–Beuren syndrome, and X-linked hypohidrotic ectodermal dysplasia. FACED3 was compiled from existing facial image collections; related repository references and references from the literature are listed in Supplementary Section S2.2 [26–32]. The FACED3 image collection comprised 4412 labeled images across these nine classes, with class-wise counts specified in Supplementary Section S2.2. The corrected total is the sum of 4191 images in the reported real-image training partition and 221 held-out real test images; generated augmentation images are not included in this count. The reported protocol assigned real subjects to training and test partitions before StyleGAN-3 fitting, used the real training partition for image synthesis, and reserved real images for final classification evaluation. Each real FACED3 image was reported to represent a distinct subject. The limitations of the available source, partition, permission, and subject identity records are described in Section 4.6. Raw FACED3 facial images are not redistributed.

Supplementary Section S2.2 lists additional dataset repositories and publications relevant to source traceability. Image-level provenance and reuse permissions require confirmation for each selected source. FACED3 class frequencies were imbalanced. Class-conditional image synthesis was used to support a more balanced development pool; equalized sample counts do not, however, remove sources of bias that are demographic or related to acquisition source or identity. The PDIDB synthetic facial-image resource and its implementation are documented in the Wang Genomics Laboratory repository [33].

2.2. Synthetic Data Generation Using StyleGAN-3

StyleGAN-3 [25] generated class-conditional FACED3 images using the real-image training partition. The official StyleGAN-3 reference implementation is available at https://github.com/NVlabs/stylegan3 (accessed on 30 November 2023). The generated images augmented the classifier-development dataset, while the real-image test partition was reserved for evaluation. Model configuration and measurements of generator quality are detailed in Supplementary Section S3.

Under the reported split, StyleGAN-3 was trained on 4191 labeled 256 × 256 FACED3 images, with 221 real images reserved for held-out testing. Adaptive discriminator augmentation (ADA) [34] targeted a value of 0.6. FID50k_full was monitored during fitting, and image and network snapshots were recorded every 50 ticks. Supplementary Section S3 gives the generator, discriminator, and optimization settings.

Diagnostics of generator quality, including FID, LPIPS, and measurements of syndrome consistency, are reported in the Supplementary Materials. The original sampling description specified 1500 generated images per class, whereas the available image directory differs in inventory; exact reconstruction of the original development pool is consequently limited.

2.3. Synthetic Dataset Generation Strategy

Generated images were organized according to the nine FACED3 labels and used for classifier development. Class counts characterize sampling balance but do not independently establish image fidelity or subject independence. The exact membership of the original classifier-training pool cannot be fully reconstructed from the available directory.

2.4. Proposed Framework

The proposed FAITHAI framework integrates standardized facial ROI cropping, class-conditional StyleGAN-3 image synthesis, DenseNet-169 classification, and attribution-based model inspection. Real-image test subjects were held out before generator training under the reported data-partitioning protocol.

Real training images were used for generator fitting, synthetic images supported classifier training and validation, and withheld real images were used for nine-class evaluation. Training histories, the overall confusion matrix, and the real-test confusion matrix report complementary aspects of classifier performance across this workflow.

Figure 1 summarizes the stages from facial expression backbone selection and ROI processing through FACED3 synthesis, classification, and explainability. The binary autism-label classifier followed a real-image-only training and evaluation procedure.

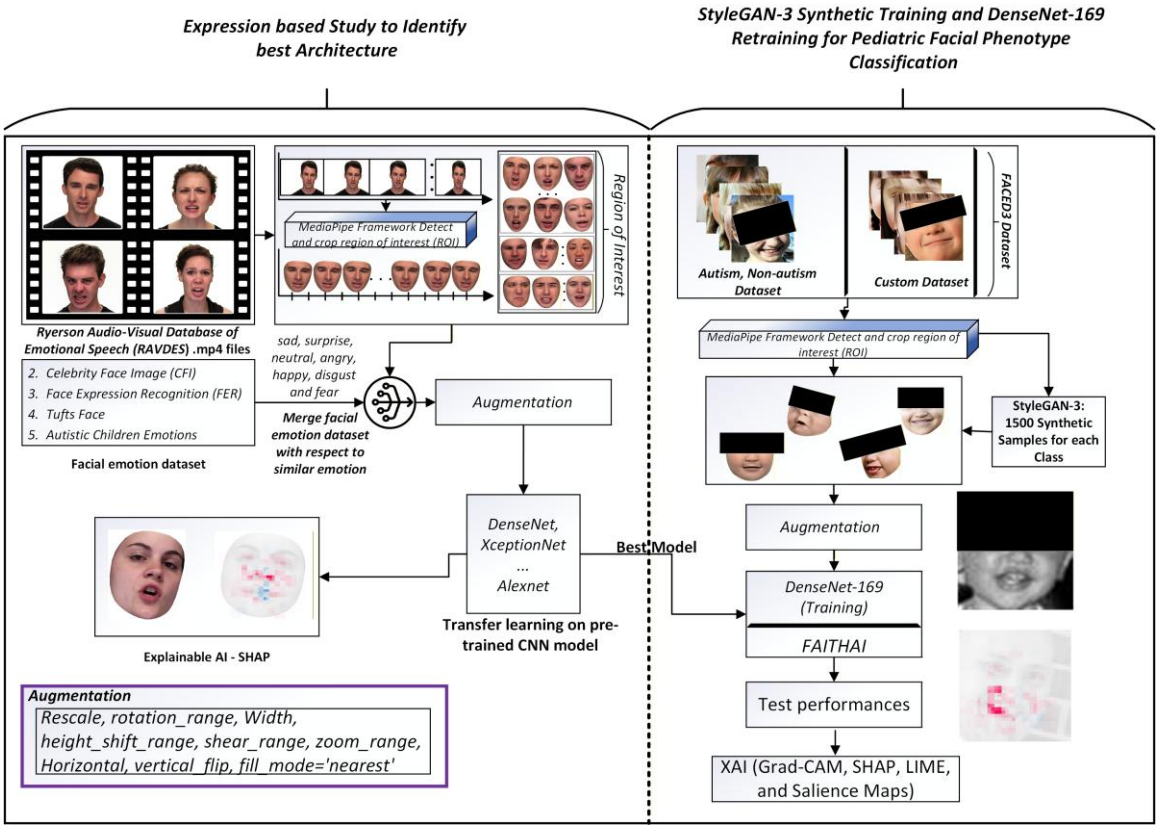

Figure 1. Conceptual FAITHAI workflow combining auxiliary backbone selection, face cropping, FACED3 synthesis, DenseNet-169 classification, and qualitative explanations, with illustrative facial images from public datasets [7,27,32,35]. Arrows indicate the direction of data processing and model selection, including the transfer of the selected backbone from facial-expression analysis to FACED3 classification. The vertical dashed line separates the two workflow stages. Black rectangular regions intentionally obscure portions of facial images to reduce recognizability.

2.5. Datasets and Task Definitions

The FACED3 analysis addresses nine predefined facial image categories. The binary task instead distinguishes the autistic-labeled and healthy categories, defining the former as positive, and uses real-image training, validation, and test data without StyleGAN-3 synthesis.

2.6. Face ROI Detection and Preprocessing

MediaPipe (version 0.10.8) [36] was used to detect the face and crop a consistent ROI before model input. This step reduces background variation and standardizes image framing across heterogeneous sources. The MediaPipe software documentation is available at https://github.com/google-ai-edge/mediapipe (accessed on 30 November 2023). It does not remove all source-specific cues, demographic differences, or acquisition artifacts, so these factors remain relevant limitations.

2.7. Data Augmentation and Dataset Balancing

Training-time image augmentation included affine rotation, translation, scaling, and shear, together with horizontal and vertical flipping. During evaluation, images were resized and normalized without stochastic augmentation.

2.8. Synthetic Sample Generation for FACED3

The StyleGAN-3 generator was fitted to the real training partition before synthetic images were used for classifier development. For comparison of candidate backbones, a real-image-only FACED3 ablation is shown in Supplementary Figure S8, using the same reported real test partition. The generated-image classifier and real-only backbone configurations differed in training protocol; their results are presented as complementary evidence of model behavior.

2.9. Model Initialization, Training Strategy, and Backbone Selection

DenseNet-169 was selected in light of the auxiliary facial expression benchmark and the FACED3 real-only backbone ablation described in Supplementary Section S4. These comparisons informed the architecture used for the primary nine-class classification task. The PyTorch (version 2.1.1) framework and torchvision model library are documented at https://pytorch.org/ and https://docs.pytorch.org/vision/stable/index.html, respectively (both accessed on 30 November 2023).

The FACED3 classifier used torchvision DenseNet-169 with ImageNet-pretrained weights. The backbone was fine-tuned, and the classification head mapped the extracted representations to the nine FACED3 labels.

The available classifier-development images were divided into stratified training (80%) and validation (20%) partitions. Checkpoint selection used the highest validation accuracy, with validation loss to resolve ties; final nine-class metrics were then obtained from the fixed 221-image real test set. Limitations in reconstructing original split and training-image membership are addressed in Section 4.6.

A further FACED3 configuration used original real training images together with StyleGAN-3-generated images to fit the nine-class DenseNet-169 classifier. Evaluation was performed on 221 held-out real FACED3 images; the training composition, learning curves, and classification results for this configuration are provided in Supplementary Section S11.

2.10. Explainable AI for Interpretability and Transparency

SHAP, LIME, Grad-CAM, saliency, and activation maps were used for qualitative model inspection; selected maps are retained as illustrative prior analyses in Section 3.4 and Supplementary Section S5. Supporting documentation and methodological resources for SHAP, LIME, Grad-CAM, saliency maps, and activation-map visualization are available at https://shap.readthedocs.io/en/latest/, https://github.com/marcotcr/lime, https://arxiv.org/abs/1610.02391, https://arxiv.org/abs/1312.6034, and https://docs.pytorch.org/docs/stable/generated/torch.nn.Module.html, respectively (all accessed on 30 November 2023). The quantitative Grad-CAM assessment used all 221 real FACED3 test images and included deletion-based faithfulness against random deletion, map stability under small brightness changes, classifier-head randomization, and a border-versus-random-region blur control. Lower deletion AUC indicates faster degradation of the predicted-class score under the specified image-removal procedure.

2.11. Final Evaluation and Reporting

Nine-class evaluation metrics included accuracy, balanced accuracy, precision, recall, specificity, F1-score, class-wise one-vs-rest ROC AUC and average precision, and cross-

entropy log loss. Micro, macro, and support-weighted averages were reported where applicable. The count confusion matrix supported calculation of class-wise decision metrics, and a Wilson interval quantified image-level accuracy uncertainty. Seven random-seed runs were evaluated on the same fixed real-image test split; a five-run subset with the highest observed test accuracies is tabulated in the Supplementary Materials. Binary classification was assessed separately using ten retrainings with validation-based checkpoint selection.

The DenseNet-169 classifier was optimized with Adam at an initial learning rate of 0.001, reduced by 1% every 10 epochs, using batch size 16, a maximum of 100 epochs, and early stopping patience of 30. Inputs were resized to 224 × 224 pixels and normalized with ImageNet statistics. Training augmentation included rotation up to 180°, shifts up to 0.3, scaling between 0.7 and 1.3, shear of 0.3°, and horizontal and vertical flips. The selected checkpoint was saved at epoch 96.

The classification head comprised dropout (0.3), a 500-unit fully connected layer with ReLU activation, a second dropout layer (0.3), and nine output units. All DenseNet-169 backbone layers were fine-tuned. The study used previously collected digital facial-image datasets. No new images were acquired using physical cameras, scanners, or other laboratory image-acquisition equipment.

3. Results

3.1. Facial Expression Benchmarking

In the auxiliary facial expression benchmark, DenseNet-169 attained an accuracy of 0.83 and balanced accuracy of 0.79 (Supplementary Table S5). The real-image FACED3 backbone comparison also favored DenseNet-169 at the evaluated 25% fine-tuning depth (Supplementary Table S6 and Figure S8). These results informed the backbone selection for the FAITHAI classifier.

3.2. FACED3 Classifier Results

The nine-class FACED3 classifier was evaluated by examining its learning behavior during training and validation, followed by performance assessment on a held-out test set of 221 real facial images. The training and validation trajectories are presented in Figure 2, while class-wise predictions and overall classification performance on the real-image test set are reported in Figure 3 and Table 1, respectively.

The training and validation curves in Figure 2 document increasing development accuracy and decreasing loss. On held-out real images, classification accuracy was 0.733; the development curves and real-test performance represent complementary stages of the FACED3 evaluation.

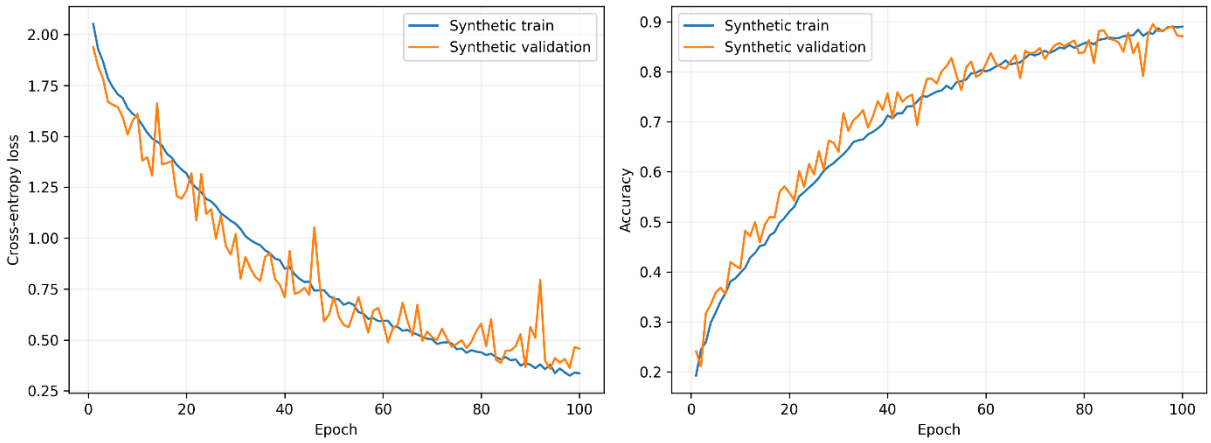

Figure 2. FACED3 classifier training and validation accuracy and loss curves.

The real-image confusion matrix (Figure 3) contains 162 correct predictions among 221 cases. The classifier correctly identified 62 of 65 autistic-labeled images and 2 of 10 Kabuki images; 22 of 68 healthy images were classified in the autistic-labeled category.

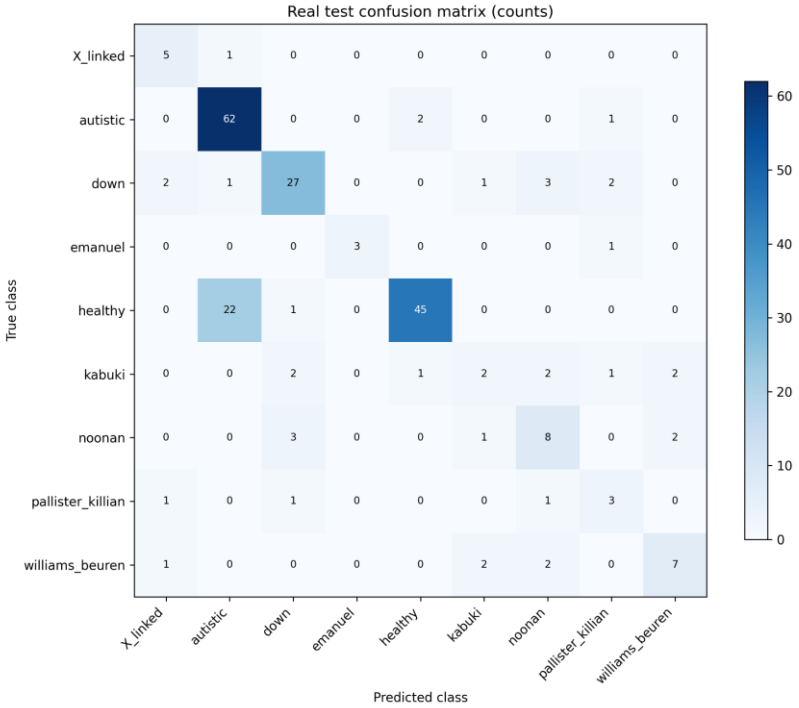

Figure 3. Nine-class FACED3 confusion matrix on 221 held-out real images (rows, true labels; columns, predicted labels).

The nine-class FACED3 classifier achieved an accuracy of 0.733 (95% Wilson confidence interval: 0.671–0.787), balanced accuracy of 0.645, macro precision of 0.650, macro F1-score of 0.635, and macro one-vs-rest specificity of 0.963 on the held-out real-image test set. Score-based evaluation yielded a log loss of 0.894, with micro, macro, and weighted one-vs-rest ROC AUC values of 0.963, 0.955, and 0.953, respectively, and corresponding average precision values of 0.803, 0.710, and 0.824, as detailed in Table 1. Macro one-vs-rest ROC AUC measures class-wise score ranking across thresholds, whereas balanced accuracy equally weights recall from final nine-class decisions. With unequal class supports and errors concentrated in particular categories, an AUC of 0.955 and balanced accuracy of 0.645 are therefore compatible.

The classifier's epoch-wise training and validation trajectories are presented in Figure 2. Class-wise discrimination was further evaluated through nine one-vs-rest ROC curves computed on the same fixed real-image test set used for the confusion matrix, as shown in Supplementary Figure S13 and Figure 3, respectively.

Across seven random-seed runs on the fixed 221-image real test split, mean ± sample SD was 0.732 ± 0.034 for accuracy, 0.654 ± 0.035 for balanced accuracy, 0.631 ± 0.035 for macro F1, and 0.961 ± 0.008 for macro AUROC. Supplementary Table S8 reports the five runs with the highest test accuracies; their selected subset mean should be distinguished from the seven-run summary.

The class-wise findings show substantial variation in recall, especially among classes with few test images. Overall classification results and the held-out real-image confusion matrix together characterize the performance of the nine-class FACED3 classifier.

Table 1. Classification metrics for the 221-image held-out real FACED3 test split.

Analysis	Accuracy	Balanced Accuracy	Precision	Recall	F1	Specificity	ROC AUC
FACED3	0.733	0.645	0.650	0.645	0.635	0.963	0.955

The FACED3 configuration trained with combined real and StyleGAN-3-generated images correctly classified 187 of 221 held-out real test images (accuracy 0.846; 95% Wilson confidence interval, 0.793–0.888). Its balanced accuracy was 0.835, macro precision 0.854, macro F1-score 0.838, and macro one-vs-rest specificity 0.978. These results pertain to the combined-training configuration; the training curves, class-wise decisions, and ROC analysis are reported in Supplementary Section S11 (Figures S15–S17 and Tables S13 and S14).

3.3. Real-Image Binary Autism-Label Classification

For the binary task, a DenseNet-169 classifier was retrained using only real images from the autistic-labeled and healthy categories, without synthetic augmentation. Evaluation used a fixed held-out set of 525 images, comprising 260 healthy and 265 autistic-labeled cases. The autistic-labeled category was the positive class; the image labels do not themselves establish clinical diagnosis.

Ten retrainings with different random seeds were evaluated on the same held-out real-image set. In each run, the checkpoint epoch was selected using minimum validation loss. The five lowest-validation-loss runs (seeds 22, 11, 66, 33, and 99) are summarized in Supplementary Table S10, and results across all ten runs are reported in Supplementary Section S9.

The detailed binary-classification figures correspond to seed 11, which ranked second by validation loss and correctly classified 439 of 525 held-out real images. Figure 4 shows the training and validation trajectories for seed 11, with minimum validation loss of 0.4329 at epoch 3. Accuracy was 0.836 (95% Wilson interval, 0.802–0.865), balanced accuracy was 0.836, autistic-label sensitivity was 0.849, healthy-label specificity was 0.823, precision was 0.830, F1-score was 0.840, and Matthews correlation coefficient was 0.672. Figure 5 shows 214 true negatives, 46 false positives, 40 false negatives, and 225 true positives; Supplementary Table S11 gives the full-precision values.

Continuous scores on the held-out real-image set yielded ROC AUC of 0.904 (95% stratified-bootstrap interval, 0.877–0.930; Figure 6) and average precision of 0.892 (Supplementary Figure S14). The ROC curve also indicates the fixed 0.5 classification threshold.

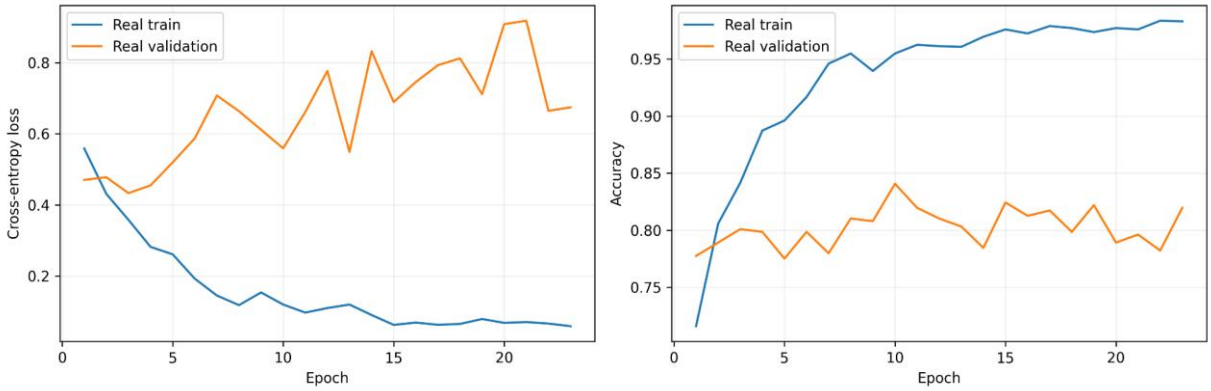

Figure 4. Training and validation accuracy and loss of the real-image binary autism-label classifier (seed 11; selected epoch 3).

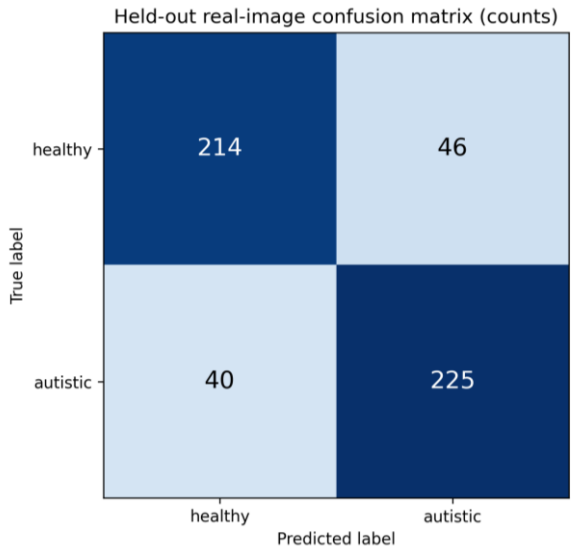

Figure 5. Binary autism-label confusion matrix on 525 held-out real images (seed 11; rows, true labels; columns, predictions; autistic-labeled cases positive).

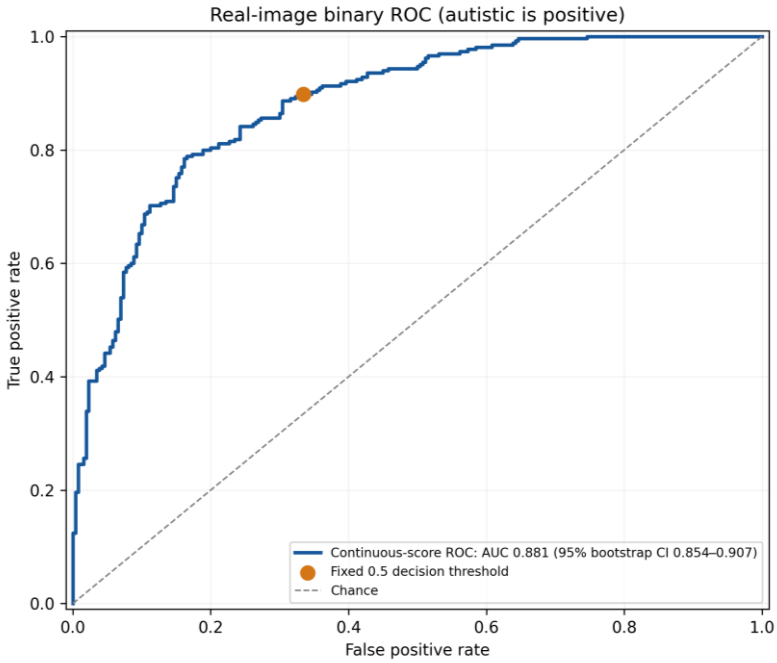

Figure 6. ROC curve for real-image binary autism-label classification (seed 11), showing the 0.5 threshold, AUROC 0.904, and 95% stratified-bootstrap interval 0.877–0.930.

3.4. Explainability Analysis

Figure 7 presents illustrative prior analyses using SHAP, LIME, Grad-CAM, and saliency maps. Quantitative Grad-CAM diagnostics were also evaluated on the nine-class FACED3 real-image test set to measure attribution faithfulness, stability, model dependence, and sensitivity to image-border perturbations. These perturbation tests quantify computational attribution responses and do not establish the clinical or anatomical relevance of the highlighted facial regions.

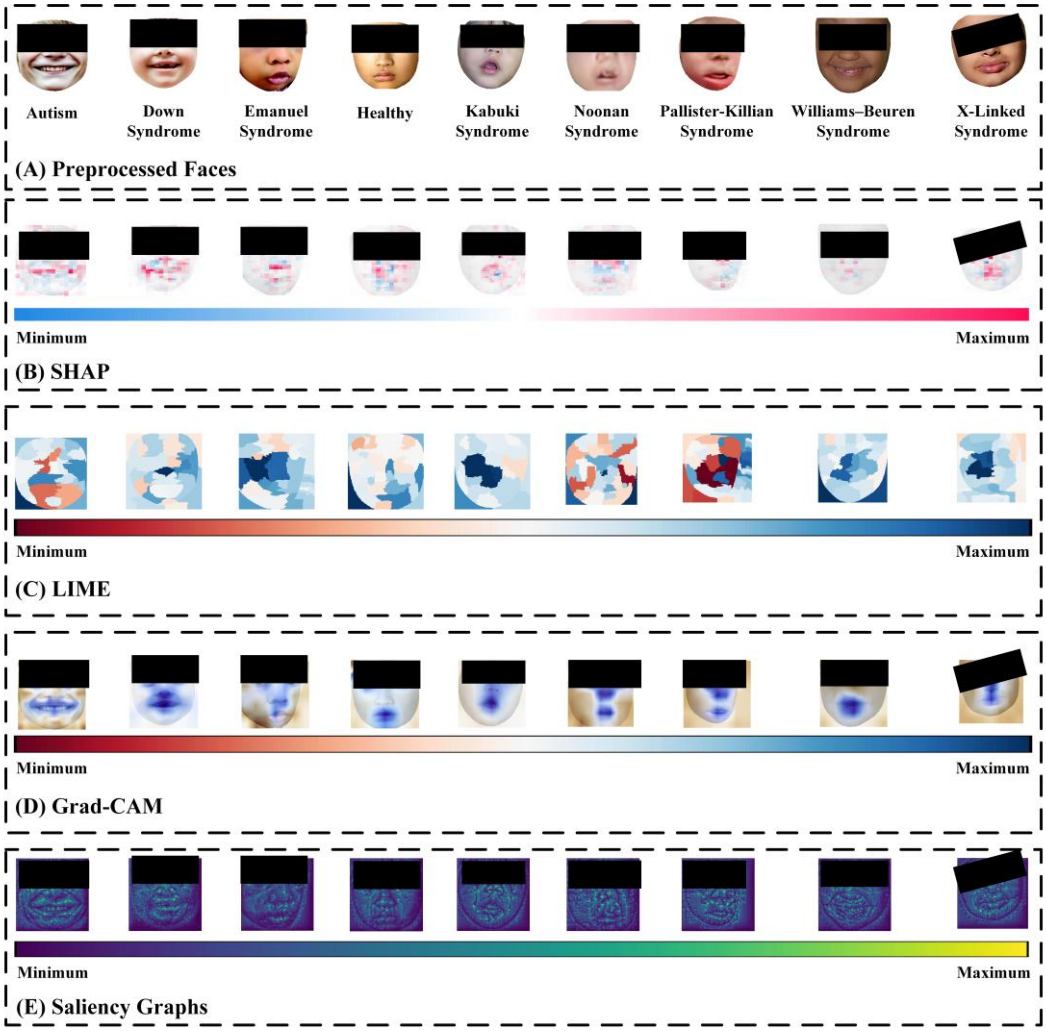

Figure 7. Illustrative prior analyses using SHAP, LIME, Grad-CAM, and saliency maps for facial image classification.

For the 221 real FACED3 test images (162 correct and 59 incorrect classifications), Grad-CAM-guided deletion produced pooled AUC 0.217, compared with 0.471 for random-region deletion, and achieved a lower paired deletion AUC for 216 images. Median map correlation was 0.956 following small brightness perturbations and 0.125 after classifier-head randomization. Border blurring caused a mean predicted-class probability change that was 0.270 smaller than blurring equally sized random regions. Table 2 summarizes these perturbation results, with the class-wise prediction outcomes provided in Supplementary Table S12.

Table 2. Quantitative Grad-CAM perturbation diagnostics on 221 held-out real FACED3 images.

Diagnostic	Aggregate Result	Interpretation and Scope
Deletion faithfulness	Grad-CAM AUC, 0.217; random AUC, 0.471; paired improvement, 216/221	Lower AUC corresponds to faster predicted-class score degradation under tested deletion.
Brightness stability	Median map correlation, 0.956	Similar attribution maps after small brightness perturbations.
Model weight dependence	Median map correlation was 0.125 after classifier-head randomization	Attribution maps changed substantially under this parameter randomization.
Border sensitivity	Mean probability effect was 0.270 smaller for border blur than for equal-area random blur	Limited perturbation control for peripheral image cues.

4. Discussion

4.1. FACED3 Classifier Performance

On 221 held-out real FACED3 images, the nine-class classifier achieved accuracy of 0.733, balanced accuracy of 0.645, and macro F1-score of 0.635. Class-wise performance varied, most notably for healthy images, 22 of which were predicted as autistic-labeled, and for the small rare-syndrome categories. The one-vs-rest ROC curves in Supplementary Figure S13 describe score ranking for each class, whereas the confusion matrices show final class assignments.

The errors between healthy-labeled and autistic-labeled images were asymmetric. Of 68 healthy-labeled test images, 45 were correctly classified (recall, 0.662), while 22 were incorrectly assigned the autistic label; by comparison, 62 of 65 autistic-labeled images were correctly classified (recall, 0.954). The concentration of healthy-to-autistic errors is consequential for this dataset-label task. Without verified subject- and source-level information, it is not possible to determine whether this pattern reflects visual overlap, source-related cues, or label noise, and it cannot support diagnostic interpretation.

The development and real-image results illustrate a gap between performance during classifier fitting and generalization to the fixed real-image test set. In addition, the reported subject-level holdout cannot be fully audited from the available original split manifest and generator input records, limiting conclusions about performance across independent subjects and data sources.

With original real and GAN-generated images jointly used for classifier training, accuracy on 221 held-out real FACED3 images was 0.846, with balanced accuracy of 0.835 and macro F1-score of 0.838 (Supplementary Section S11). The higher observed values should not be attributed to a single training component without matched comparisons using verified subject separation, identical training settings, and repeated random seeds. The synthetic-only configuration obtained 0.733 accuracy (95% Wilson interval, 0.671–0.787) on the same 221-image test set, whereas the combined-training accuracy interval was 0.793–0.888. Although the image-level intervals do not overlap, this descriptive contrast cannot substitute for a matched, paired, repeated-seed component comparison.

4.2. Binary Autism-Label Classification

Retraining on real images only yielded accuracy of 0.836 and balanced accuracy of 0.836 in the illustrated seed-11 run on 525 held-out images, with autistic-label sensitivity of 0.849 and healthy-label specificity of 0.823. Continuous scores produced ROC AUC of 0.904 (95% stratified-bootstrap interval, 0.877–0.930) and average precision of 0.892. These findings describe discrimination between dataset labels rather than clinical screening performance. The binary task uses two labels and a separate 525-image partition, whereas the nine-class FACED3 result comes from 221 test images; their accuracies of 0.836 and 0.733 are not directly comparable.

For the ten binary runs evaluated on the fixed real-image test split, the five checkpoints selected by lowest validation loss yielded mean test accuracy of 0.815 ± 0.018 and mean ROC AUC of 0.899 ± 0.015 (sample SD). Across all ten runs, these values were 0.817 ± 0.016 and 0.901 ± 0.011. The image-level uncertainty estimates do not account for possible source, identity, or demographic dependence.

4.3. Contributions and Limitations of the FAITHAI Components

The FACED3 real-image-only backbone ablation (Supplementary Figure S8) used the same reported 221-image real test partition and provides context for architecture selection. Because model initialization and training configurations differed, it does not isolate the contribution of synthetic-image training. Matched comparisons of real-only, synthetic-

only, and combined training, with and without ROI preprocessing, are needed to quantify the contribution of individual components.

On the same reported 221-image test partition, synthetic-only training yielded 0.733 accuracy, slightly below the approximately 0.74 reported for the real-only DenseNet-169 backbone at 25% fine-tuning (Supplementary Table S6). Generator quality did not map directly to classification recall: the healthy category had FID of 18.83 and syndrome-consistency of 0.88 but recall of 0.662, whereas the autistic-labeled category had FID of 21.49 and syndrome-consistency of 0.68 but recall of 0.954 (Supplementary Tables S3 and S4). Thus, the present evidence supports investigation of synthetic-image-assisted training, not a demonstrated gain from synthetic-only training over the reported real-only backbone.

The illustrative prior attribution maps provide examples of facial regions emphasized by SHAP, LIME, Grad-CAM, and saliency methods. Quantitative Grad-CAM deletion, brightness-stability, classifier-head-randomization, and border-blur diagnostics on the 221-image real test set (Table 2) add numerical characterization. Class-stratified attribution results and clinician-annotated reference regions would further clarify anatomical relevance.

4.4. Comparison with Previous Studies

Previous facial syndrome classification studies have reported high performance on tasks that are frequently binary or syndrome-specific. For example, Hong et al. [37] investigated genetic syndrome recognition using convolutional neural networks, while Huang et al. [38,39], Yang et al. [40], and Rouxel et al. [41] evaluated models targeting specific syndromes. In contrast, FACED3 incorporates nine facial image categories within a single multiclass classification task. Differences in dataset composition, class prevalence, label definitions, and evaluation protocols limit direct numerical comparisons between these studies. Their reported performance therefore provides a broader research context rather than a common benchmark for FACED3, as summarized in Table 3.

The FAITHAI classifier achieved an accuracy of 0.733 and balanced accuracy of 0.645 on 221 held-out real FACED3 images. Macro precision, recall, specificity, and F1-score were 0.650, 0.645, 0.963, and 0.635, respectively. The class-wise one-vs-rest ROC AUC ranged from 0.925 to 0.986, with a macro ROC AUC of 0.955 and macro average precision of 0.710. These metrics provide complementary assessments of classification performance, particularly in the presence of class imbalance. However, the differences between discrimination measures and final classification accuracy emphasize the importance of examining class-wise performance and confusion-matrix errors when interpreting the results.

Several previous studies have demonstrated the effectiveness of deep learning architectures for recognizing individual genetic syndromes. Islam and Shaikh [3] employed an attention-based CNN for multiple genetic syndromes, Song et al. [42] investigated facial-feature-based recognition of Turner syndrome, and Liu et al. [43] evaluated convolutional neural networks for Williams–Beuren syndrome recognition. Other approaches, including VGG-based transfer learning and deep networks incorporating ArcFace loss, have achieved strong results under their respective evaluation protocols [3,38,40–42]. The present study extends this research direction by investigating nine facial image categories within an integrated framework combining facial standardization, class-conditional image synthesis, DenseNet-169 classification, and explainable artificial intelligence. Nevertheless, the available results do not establish the individual contribution of each framework component, as matched ablation experiments were not available.

The relatively high macro ROC AUC and specificity indicate that the classifier achieved substantial one-vs-rest discrimination for several categories. However, these measures do not necessarily correspond to equally strong final multiclass decisions. A model may distinguish one category from the remaining categories in a score-based evaluation while still producing incorrect predictions when all nine classes are considered

simultaneously. Furthermore, classification errors and predictive values depend on the underlying class distribution and decision thresholds. Consequently, the observed performance should be interpreted within the FACED3 evaluation setting rather than as evidence of clinical screening or diagnostic reliability. The qualitative SHAP, LIME, Grad-CAM, and saliency analyses, together with the quantitative Grad-CAM perturbation assessments, provide additional information about model attribution behavior but do not establish clinician-validated anatomical relevance.

Taken together, the FACED3 findings demonstrate the feasibility of integrating class-conditional image synthesis, deep learning classification, and explainable AI for exploratory multiclass pediatric facial image analysis. The evaluation on held-out real images provides a more informative assessment of classification performance than results obtained exclusively from augmented or synthetic samples. Further investigation through matched component ablations, repeated training runs, verified subject-level data partitioning, and independent external datasets would help establish the generalizability of the framework. Additional evaluation across demographic and image-source subgroups, together with clinician-reviewed explanations, would support a more comprehensive assessment of its potential relevance to clinical research.

Table 3. Published facial syndrome image classification studies using differing populations, label sets, and evaluation protocols.

Study	Dataset Used	Method	Results	Explainable AI Model
Islam and Shaikh et al. [3]	Custom-built image dataset	Deep CNN with attention mechanism	Accuracy: 0.92; F1-score: 0.89	No explainable AI utilized
Liu et al. [43]	Facial photos of 340 children (104 WBS, 91 other syndromes, 145 healthy)	VGG-19 with transfer learning	Accuracy: 0.92; precision: 0.94; recall: 0.82; F1-score: 0.87; AUC: 0.90	No explainable AI utilized
Song et al. [42]	Facial images of TS patients and healthy controls	Machine learning with multiple facial features	Accuracy: 0.87	No explainable AI utilized
Hong et al. [37]	456 facial photos (228 GS, 228 healthy)	VGG-16 with transfer learning	Accuracy: 0.89; specificity: 0.91; recall: 0.86; F1-score: 0.88; AUC: 0.94	No explainable AI utilized
Huang et al. [39]	1892 facial photos (233 NS, 863 other GSs, 796 healthy)	VGG19-BN with transfer learning	Accuracy: 0.94; precision: 0.91; specificity: 0.99; F1-score: 0.78	No explainable AI utilized
Yang et al. [40]	420 facial images (127 NS, 163 healthy, 130 other syndromes)	DCNN with ArcFace loss function	Accuracy: 0.92 ± 0.01; AUC: 0.98 ± 0.0055 (NS vs. healthy); accuracy: 0.82 ± 0.0074; AUC: 0.93 ± 0.0062 (NS vs. other syndromes)	No explainable AI utilized
Huang et al. [38]	954 facial images (135 WBS, 124 other genetic disorders, 183 healthy)	VGG-19BN with transfer learning	Accuracy: 0.95; precision: 1.00; recall: 0.84; specificity: 0.93; F1-score: 0.91	No explainable AI utilized
F. Rouxel et al. [41]	Facial images of individuals with Kabuki syndrome types 1 and 2	Deep neural network (DNN)	Accuracy: 0.91; sensitivity: 0.88; specificity: 0.93	No explainable AI utilized
FAITHAI (FACED3)	221 held-out real FACED3 images	DenseNet-169; StyleGAN-3-supported classifier development	Accuracy: 0.733; balanced accuracy: 0.645; macro F1: 0.635	Illustrated in Figure 7

4.5. Comparison with Autism Previous Studies

Previous studies on autism-related facial image classification have investigated convolutional neural networks, transfer learning, and hybrid deep learning architectures, primarily for binary classification of autism-labeled and control images. Although several studies

have reported promising classification performance, differences in dataset composition, image sources, class definitions, and evaluation protocols limit direct numerical comparisons with the present approach. The FAITHAI binary classifier was evaluated using real images without synthetic augmentation, providing a complementary perspective on autism-related facial image classification. A detailed comparison of the proposed approach with previous autism-related facial image classification studies, including their datasets, methodologies, and reported performance, is provided in Supplementary Section S8.

4.6. Limitations

The FACED3 collection was reported to contain one real facial image per subject, and manual visual inspection of the available images did not identify apparent duplicates. Nevertheless, visual inspection alone cannot establish subject-level uniqueness or exclude different images of the same individual across data partitions. Although a subject-level holdout before StyleGAN-3 training was reported, the original subject-ID ledger, split manifest, GAN input list, and complete inventory of classifier development were unavailable for independent verification. Consequently, the 221-image real test set, also used for the real-only backbone ablation, cannot be considered a fully audited subject-independent evaluation. Demographic annotations and image-specific permissions for research use and publication remain incomplete. The nine-class score-based metrics and seven-seed results could not be independently recalculated without complete per-image prediction scores and run records, while the five highest-performing seeds were selected according to observed test accuracy. For the retrained binary classifier, training summaries and sample counts were available, but case-wise predictions and retraining code were not deposited. Furthermore, image-level confidence intervals do not fully account for potential subject- or source-dependent variation. Quantitative Grad-CAM results were reported at the aggregate level without complete per-image perturbation trajectories, fully specified settings, or independent clinician annotations. The limited representation of rare syndromes, absence of matched component ablations, and lack of external-site validation further constrain the generalizability and clinical interpretation of the findings.

5. Conclusions

FAITHAI combines facial ROI preprocessing, class-conditional synthesis, DenseNet-169 classification, and explainability analysis for pediatric facial images. On the 221-image real FACED3 test set, nine-class accuracy was 0.733, balanced accuracy was 0.645, macro F1-score was 0.635, and macro one-vs-rest ROC AUC was 0.955. The real-image binary classifier, retrained with ten random seeds, achieved 0.836 accuracy and 0.904 ROC AUC in the illustrated validation-selected seed-11 run on 525 images. For nine-class Grad-CAM, deletion AUC was 0.217 versus 0.471 for random deletion, with median brightness-map and classifier-head-randomization correlations of 0.956 and 0.125, respectively. These results support exploratory image-classification research, while clinical interpretation requires independent testing and carefully matched ablations.

Further work should verify the reported pre-GAN subject split and source permissions, evaluate matched real-only and synthesized-image configurations, and assess robustness across demographic groups, acquisition sources, and external clinical cohorts. The anatomical relevance of attribution maps also requires independent replication and expert evaluation.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jimaging12110589/s1. The accompanying file provides generator-quality diagnostics, auxiliary backbone and real-image ablation results, illustrative prior explainability maps, nine-class classification metrics and random-seed variability, retrained real-image binary-classification results with ROC and precision-recall analysis, and quantitative Grad-CAM diagnostics on the 221-image FACED3 real test set. A separate combined real and StyleGAN-3-generated training configuration is documented in Supplementary Section S11, including its development curves, real-image confusion matrix, ROC curves, and class-wise metrics. The Supplementary Materials additionally discuss autism-related facial-image studies [44–49], auxiliary facial-expression datasets [50–52], and craniofacial and genetic-syndrome reference resources [53–60]. The complete captions of the supplementary figures and tables are as follows: Figure S1: RAVDESS emotion categories and intensity conditions used for facial-expression benchmarking; Figure S2: Average Fourier-spectrum heatmap of FACED3 facial images; Figure S3: Feature maps from successive StyleGAN-3 generator layers; Figure S4: Generator and discriminator loss and score trajectories during StyleGAN-3 training; Figure S5: Discriminator-score trajectories for real and StyleGAN-3-generated images; Figure S6: Augmentation, hardware-resource, and timing diagnostics during StyleGAN-3 training; Figure S7: StyleGAN-3-generated synthetic FACED3 facial-image examples; Figure S8: Real-image-only FACED3 DenseNet-169 training curves and confusion matrix at 25% fine-tuning depth; Figure S9: Illustrative prior SHAP analyses of adult facial-expression examples; Figure S10: Representative child facial-expression images and corresponding SHAP attribution maps for seven emotional expressions, illustrating the spatial distribution of feature contributions in prior analyses using open-access datasets [7,35]; Figure S11: Illustrative class-wise feature activations extracted from the conv5_block32_concat layer of DenseNet-169 for nine FACED3 categories, showing differences in internal feature representations from prior analyses; Figure S12: Illustrative prior DenseNet-169 convolutional-block activation maps showing the progression from low-level visual patterns to increasingly complex feature representations across successive network layers; Figure S13: One-vs-rest ROC curves for the nine-class FACED3 classifier on 221 held-out real images; Figure S14: Precision–recall curve for the retrained binary autism-label classifier (seed 11; positive class, autistic-labeled; average precision 0.892); Figure S15: Training and validation cross-entropy loss (left) and accuracy (right) over 100 epochs for the combined real- and GAN-generated-image FACED3 training configuration; Figure S16: Count confusion matrix for the combined-training classifier on the 221 held-out real FACED3 images, with rows showing reference labels and columns showing predicted labels; Figure S17: One-vs-rest ROC curves for the nine-class FACED3 classifier trained with both original real and GAN-generated images, evaluated on 221 held-out real images, with legend AUCs rounded to two decimals; Table S1: Previously published facial-syndrome image-classification studies using heterogeneous datasets and evaluation protocols; Table S2: Previously published autism-related image-classification studies reporting dataset-label predictions rather than clinical validation; Table S3: Per-class FID, syndrome-consistency, and LPIPS diagnostics for generated FACED3 images; Table S4: Class-wise classification metrics and one-vs-rest discrimination on 221 held-out real FACED3 images; Table S5: CNN results on the auxiliary facial-expression dataset (MNet: MobileNet; DNet: DenseNet; ENet: EfficientNet; BA: balanced accuracy; P/R/F: precision/recall/F1; μ/m/w: micro/macro/weighted); Table S6: FACED3 real-image-only backbone performance at three fine-tuning depths (single runs; NR, not reported); Table S7: Illustrative prior SHAP and Grad-CAM analyses compared with literature-described facial regions; Table S8: Classification performance of the five highest-accuracy random-seed runs on the fixed 221-image FACED3 real test set, with mean and sample standard deviation (SD) calculated for the five selected runs; Table S9: Comparison of previous autism-related facial-image classification studies with the proposed FAITHAI binary classifier under different datasets and evaluation protocols; Table S10: Five binary-classification retrainings ranked by validation loss, with results on the fixed 525-image real test set; Table S11: Binary autism-label classification metrics for seed 11 on 525 held-out real images (autistic-labeled cases positive); Table S12: Class-wise prediction outcomes for the nine-class

FACED3 real-image test split (N = 221); Table S13: Image-level classification metrics for the combined real- and GAN-generated training configuration on the held-out nine-class FACED3 real-image test set; and Table S14: Class-wise real-image classification performance for the combined real- and GAN-generated training configuration, with specificity calculated using each class against all remaining classes.

Author Contributions: Conceptualization, S.T.H.S., S.A.H.S. and S.B.H.S.; data curation, S.T.H.S., S.A.H.S., S.B.H.S., J.P.-d.-R., R.K. and A.Z.; formal analysis, S.T.H.S., S.A.H.S. and S.B.H.S.; methodology, S.T.H.S., S.A.H.S. and S.B.H.S.; software, S.T.H.S., J.P.-d.-R., R.K. and A.Z.; investigation, J.P.-d.-R., R.K. and S.G.; validation, S.M.J.R., S.A.Q., A.R.-Z., L.C.G., M.A.D. and S.G.; visualization, S.A.H.S., S.B.H.S., S.M.J.R., S.A.Q., A.R.-Z. and L.C.G.; supervision, S.M.J.R., S.A.Q., A.R.-Z., L.C.G. and M.A.D.; project administration, L.C.G. and M.A.D.; resources, L.C.G., M.A.D. and S.G.; funding acquisition, M.A.D.; writing—original draft, S.T.H.S., S.A.H.S., S.B.H.S., J.P.-d.-R., R.K. and A.Z.; writing—review and editing, all authors. All authors have read and agreed to the published version of the manuscript.

Funding: The present research was carried out as part of the PARENT project. PARENT received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Innovative Training Network, Grant Agreement No. 956394 (https://parenth2020.com/; accessed on 5 October 2026).

Institutional Review Board Statement: This study used existing FACED3 images without recruiting participants or taking new facial photographs. Related datasets and repositories, described in Supplementary Section S2.2, include Kaggle (autism-labeled, healthy, and Down syndrome images), RDFace-Syn (X-linked hypohidrotic ectodermal dysplasia), and PDIDB (Kabuki, Noonan, and Williams–Beuren syndromes). For rare conditions such as Emanuel and Pallister–Killian syndromes, specialized resources, including the controlled-access GestaltMatcher Database [8], provide relevant context for facial phenotyping. Source-traceability limitations and secondary-use conditions for the FACED3 images are discussed in Supplementary Section S2.2 and the Informed Consent Statement.

Informed Consent Statement: This research did not obtain new participant consent because this study involved secondary analysis of existing FACED3 images without new participant recruitment or photography. Use of these images is subject to the consent, licensing, and access conditions of their respective sources, with relevant documentation and limitations described in Supplementary Section S2.2. Complete image-level permission records were not available for independent verification; we acknowledge this limitation in the revised manuscript. Selected facial examples were displayed with eye-region occlusion as an additional privacy precaution, which does not replace any applicable source-specific permission requirements.

Data Availability Statement: Raw FACED3 facial images are not redistributed. The submission includes aggregate nine-class metrics, seven-run summaries, the 221-image real-test confusion matrix, classifier-development curves, and nine-class ROC plots. Per-image nine-class probabilities, labels, split manifests, training logs, subject/source mappings, and permission ledgers are not deposited. For the real-image binary retraining, the 525-image confusion matrix, training and validation histories, ROC and precision-recall plots, a ten-run summary, and five validation-ranked results are provided, but case-wise prediction files and code are not deposited. Aggregate quantitative Grad-CAM measures and class-wise prediction counts are given in Table 2 and Supplementary Table S11; per-image attribution maps, perturbation trajectories, and implementation logs are not deposited.

Acknowledgments: The authors acknowledge the support of Politecnico di Torino and the University of Cádiz. Generative AI (GPT-4, GPT-5, and GPT-6, OpenAI) was used to assist with language editing and manuscript preparation. All scientific content, calculations, and citations were reviewed by the authors, who take responsibility for the manuscript.

Conflicts of Interest: Author Syed Adil Hussain Shah was employed by the company GPI SpA. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Joshi, R.; Jadeja, M. The Synergy of Clinical Psychology and Affective Computing: Advancements in Emotion Recognition and Therapy. In *Affective Computing for Social Good: Enhancing Well-Being, Empathy, and Equity*; Garg, M., Prasad, R.S., Eds.; Springer Nature: Cham, Switzerland, 2024; pp. 21–45.
2. Yu, I.-C.; Guo, J.-M.; Lin, W.-C.; Fang, J.-T. Development of Nonverbal Communication Behavior Model for Nursing Students Based on Deep Learning Facial Expression Recognition Technology. *Cogent Educ.* **2025**, *12*, 2448059. https://doi.org/10.1080/2331186X.2024.2448059.
3. Islam, T.U.; Shaikh, T.A. A Lightweight Attention-Based Deep Learning Facial Recognition System for Multiple Genetic Syndromes. *Int. J. Data Sci. Anal.* **2024**, *20*, 2871–2889. https://doi.org/10.1007/s41060-024-00658-y.
4. Martinez, B.; Valstar, M.F. Advances, Challenges, and Opportunities in Automatic Facial Expression Recognition. In *Advances in Face Detection and Facial Image Analysis*; Kawulok, M., Celebi, M.E., Smolka, B., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 63–100.
5. Ekman, P.; Friesen, W.V. Constants across Cultures in the Face and Emotion. *J. Personal. Soc. Psychol.* **1971**, *17*, 124–129. https://doi.org/10.1037/h0030377.
6. Shah, S.T.H.; Shah, S.A.H.; Qureshi, S.A.; Di Terlizzi, A.; Deriu, M.A. Automated Facial Characterization and Image Retrieval by Convolutional Neural Networks. *Front. Artif. Intell.* **2023**, *6*, 1230383. https://doi.org/10.3389/frai.2023.1230383.
7. Face Expression Recognition Dataset. Available online: https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset (accessed on 30 November 2023).
8. Hsieh, T.-C.; Bar-Haim, A.; Moosa, S.; Ehmke, N.; Gripp, K.W.; Pantel, J.T.; Danyel, M.; Mensah, M.A.; Horn, D.; Rosnev, S.; et al. GestaltMatcher Facilitates Rare Disease Matching Using Facial Phenotype Descriptors. *Nat. Genet.* **2022**, *54*, 349–357. https://doi.org/10.1038/s41588-021-01010-x.
9. Song, J.; He, M.; Zheng, X.; Zhang, Y.; Bi, C.; Feng, J.; Du, J.; Li, H.; Shen, B. Face-Based Machine Learning Diagnostics: Applications, Challenges and Opportunities. *Artif. Intell. Rev.* **2025**, *58*, 243. https://doi.org/10.1007/s10462-025-11246-2.
10. Alharbi, K.N. Explainable AI for Deep Visual Recognition: Evaluation, Methods, and Open Challenges. *Electronics* **2026**, *15*, 3222. https://doi.org/10.3390/electronics15143222.
11. Murthy, Y.S.; Bansal, R.; Chodisetty, R.S.C.M.; Chakravorty, C.; Sai, K.P. Transforming Minds AI and Machine Learning Applications in Cognitive Neuroscience. In *Transforming Neuropsychology and Cognitive Psychology with AI and Machine Learning*; IGI Global Scientific Publishing: New York, NY, USA, 2025; pp. 107–128.
12. Alam, M.S.; Tasneem, Z.; Khan, S.A.; Rashid, M.M. Effect of Different Modalities of Facial Images on ASD Diagnosis Using Deep Learning-Based Neural Network. *J. Adv. Res. Appl. Sci. Eng. Technol.* **2023**, *32*, 59–74. https://doi.org/10.37934/araset.32.3.5974.
13. Wang, J.; Wang, J.; Wang, S.; Zhang, Y. Deep Learning in Pediatric Neuroimaging. *Displays* **2023**, *80*, 102583. https://doi.org/10.1016/j.displa.2023.102583.
14. Huang, R.; Suttie, M.; Noble, J.A. An Automated CNN-Based 3D Anatomical Landmark Detection Method to Facilitate Surface-Based 3D Facial Shape Analysis. In *Proceedings of the Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures*; Greenspan, H., Tanno, R., Erdt, M., Arbel, T., Baumgartner, C., Dalca, A., Sudre, C.H., Wells, W.M., Drechsler, K., Linguraru, M.G., et al, Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 163–171.
15. Yuan, M.; Goovaerts, S.; Vanneste, M.; Matthews, H.; Hoskens, H.; Richmond, S.; Klein, O.D.; Spritz, R.A.; Hallgrimsson, B.; Walsh, S.; et al. Mapping Genes for Human Face Shape: Exploration of Univariate Phenotyping Strategies. *Comput. Biol.* **2024**, *20*, e1012617.
16. Rahman, M.M.; Usman, O.L.; Muniyandi, R.C.; Sahran, S.; Mohamed, S.; Razak, R.A. A Review of Machine Learning Methods of Feature Selection and Classification for Autism Spectrum Disorder. *Brain Sci.* **2020**, *10*, 949. https://doi.org/10.3390/brainsci10120949.
17. Parlett-Pelleriti, C.M.; Stevens, E.; Dixon, D.; Linstead, E.J. Applications of Unsupervised Machine Learning in Autism Spectrum Disorder Research: A Review. *Rev. J. Autism Dev. Disord.* **2023**, *10*, 406–421. https://doi.org/10.1007/s40489-021-00299-y.
18. Roy, A.; Rivas, P.; Tayba, M.N. Unsupervised Machine Learning Methods for Diagnosing Autism Spectrum Disorder Using Multimodal Data: A Survey. In *Proceedings of the 2022 International Conference on Computational Science and Computational Intelligence (CSCI)*; IEEE: New York, NY, USA, 2022; pp. 41–52.

19. Rezaee, K. You Look at the Face of an Angel: An Innovative Hybrid Deep Learning Approach for Detecting Down Syndrome in Children's Faces Through Facial Analysis. *J. AI Data Min.* **2024**, *12*, 287–303. https://doi.org/10.22044/jadm.2024.14817.2580.
20. Khodatars, M.; Shoeibi, A.; Sadeghi, D.; Ghaasemi, N.; Jafari, M.; Moridian, P.; Khadem, A.; Alizadehsani, R.; Zare, A.; Kong, Y.; et al. Deep Learning for Neuroimaging-Based Diagnosis and Rehabilitation of Autism Spectrum Disorder: A Review. *Comput. Biol. Med.* **2021**, *139*, 104949. https://doi.org/10.1016/j.compbiomed.2021.104949.
21. Klie, J.-C.; de Castilho, R.E.; Gurevych, I. Analyzing Dataset Annotation Quality Management in the Wild. *arXiv* **2024**, arXiv:2307.08153. https://doi.org/10.48550/arXiv.2307.08153.
22. Nasarian, E.; Alizadehsani, R.; Acharya, U.R.; Tsui, K.-L. Designing Interpretable ML System to Enhance Trust in Healthcare: A Systematic Review to Proposed Responsible Clinician-AI-Collaboration Framework. *Inf. Fusion* **2024**, *108*, 102412. https://doi.org/10.1016/j.inffus.2024.102412.
23. Başaran, E.; Tümen, V.; Çelik, G.; Toğaçar, M.; Karaoğlan, K.M.; Akın, E. Advanced Gastrointestinal Image Classification Based on Vision Transformer, Hypercolumn and Biological Optimization. *Sci. Rep.* **2026**, *16*, 37609. https://doi.org/10.1038/s41598-026-66459-4.
24. Güngür, Z.; Ayaz, İ.; Tümen, V. Biomedical Image Super-Resolution Using SRGAN: Enhancing Diagnostic Accuracy. *Bitlis Eren Üniversitesi Fen. Bilimleri Dergisi* **2025**, *14*, 198–212. https://doi.org/10.17798/bitlisfen.1565824.
25. Zhu, T.; Chen, J.; Zhu, R.; Gupta, G. StyleGAN3: Generative Networks for Improving the Equivariance of Translation and Rotation. *arXiv* **2023**, arXiv:2307.03898.
26. Feng, G.; Long, Y.; Ali, H.; Lou, E.; Butt, F.; Liu, Q.; Wang, Y.; Hu, P. RDFace: A Benchmark Dataset for Rare Disease Facial Image Analysis under Extreme Data Scarcity and Phenotype-Aware Synthetic Generation. *arXiv* **2026**, arXiv:2604.03454.
27. Autism_Image_Data. Available online: https://www.kaggle.com/datasets/cihan063/autism-image-data (accessed on 30 November 2023).
28. Down Syndrome Dataset. Available online: https://www.kaggle.com/datasets/mamunhasan2cs/down-syndrome-dataset (accessed on 28 January 2026).
29. Wang Genomics Laboratory Phenotype Disease Image Database (PDIDB). Available online: https://pdidb.wglab.org/ (accessed on 12 December 2025).
30. RDFace-Syn. Available online: https://www.kaggle.com/datasets/ganlinfenggg/rdface-synthetic-data (accessed on 29 October 2026).
31. Feng, G. RDFace (Kkathyf/RDFace): Official GitHub Repository. 2026. Available online: https://github.com/Kkathyf/RDFace (accessed on 29 October 2026).
32. Phenotype Disease Image Database | PDIDB. Available online: https://pdidb-dev.wglab.org/ (accessed on 29 October 2026).
33. Wang Genomics Laboratory. PDIDB: Phenotype Diseased Image Synthesis Database. GitHub Repository. Available online: https://github.com/WGLab/PDIDB (accessed on 11 November 2026).
34. Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; Aila, T. Training Generative Adversarial Networks with Limited Data. *arXiv* **2020**, arXiv:2006.06676.
35. Autistic Children Emotions—Dr. Fatma M. Talaat. Available online: https://www.kaggle.com/datasets/fatmamtalaat/autistic-children-emotions-dr-fatma-m-talaat (accessed on 4 December 2023).
36. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.-L.; Yong, M.G.; Lee, J.; et al. MediaPipe: A Framework for Building Perception Pipelines. *arXiv* **2019**, arXiv:1906.08172.
37. Hong, D.; Zheng, Y.-Y.; Xin, Y.; Sun, L.; Yang, H.; Lin, M.-Y.; Liu, C.; Li, B.-N.; Zhang, Z.-W.; Zhuang, J.; et al. Genetic Syndromes Screening by Facial Recognition Technology: VGG-16 Screening Model Construction and Evaluation. *Orphanet J. Rare Dis.* **2021**, *16*, 344. https://doi.org/10.1186/s13023-021-01979-y.
38. Huang, P.; Huang, J.; Huang, Y.; Yang, M.; Kong, R.; Sun, H.; Han, J.; Guo, H.; Wang, S. Optimization and Evaluation of Facial Recognition Models for Williams-Beuren Syndrome. *Eur. J. Pediatr.* **2024**, *183*, 3797–3808. https://doi.org/10.1007/s00431-024-05646-9.
39. Huang, Y.; Sun, H.; Chen, Q.; Shen, J.; Han, J.; Shan, S.; Wang, S. Computer-Based Facial Recognition as an Assisting Diagnostic Tool to Identify Children with Noonan Syndrome. *BMC Pediatr.* **2024**, *24*, 361. https://doi.org/10.1186/s12887-024-04827-7.
40. Yang, H.; Hu, X.-R.; Sun, L.; Hong, D.; Zheng, Y.-Y.; Xin, Y.; Liu, H.; Lin, M.-Y.; Wen, L.; Liang, D.-P.; et al. Automated Facial Recognition for Noonan Syndrome Using Novel Deep Convolutional Neural Network with Additive Angular Margin Loss. *Front. Genet.* **2021**, *12*, 669841. https://doi.org/10.3389/fgene.2021.669841.
41. Rouxel, F.; Yauy, K.; Boursier, G.; Gatinois, V.; Barat-Houari, M.; Sanchez, E.; Lacombe, D.; Arpin, S.; Giuliano, F.; Haye, D.; et al. Using Deep-Neural-Network-Driven Facial Recognition to Identify Distinct Kabuki Syndrome 1 and 2 Gestalt. *Eur. J. Hum. Genet.* **2022**, *30*, 682–686. https://doi.org/10.1038/s41431-021-00994-8.

42. Song, W.; Lei, Y.; Chen, S.; Pan, Z.; Yang, J.-J.; Pan, H.; Du, X.; Cai, W.; Wang, Q. Multiple Facial Image Features-Based Recognition for the Automatic Diagnosis of Turner Syndrome. *Comput. Ind.* **2018**, *100*, 85–95. https://doi.org/10.1016/j.compind.2018.03.021.
43. Liu, H.; Mo, Z.-H.; Yang, H.; Zhang, Z.-F.; Hong, D.; Wen, L.; Lin, M.-Y.; Zheng, Y.-Y.; Zhang, Z.-W.; Xu, X.-W.; et al. Automatic Facial Recognition of Williams-Beuren Syndrome Based on Deep Convolutional Neural Networks. *Front. Pediatr.* **2021**, *9*, 648255. https://doi.org/10.3389/fped.2021.648255.
44. Khan, B.; Bhatti, S.M.; Akram, A. Autism Spectrum Disorder Detection in Children Via Deep Learning Models Based on Facial Images. *Bull. Bus. Econ. (BBE)* **2024**, *13*, 557–572. https://doi.org/10.61506/01.00241.
45. Fahaad Almufareh, M.; Tehsin, S.; Humayun, M.; Kausar, S. Facial Classification for Autism Spectrum Disorder. *J. Disabil. Res.* **2024**, *3*, 20240025. https://doi.org/10.57197/JDR-2024-0025.
46. Anjum, J.; Hia, N.A.; Waziha, A.; Kalpoma, K.A. Deep Learning-Based Feature Extraction from Children's Facial Images for Autism Spectrum Disorder Detection. In *Proceedings of the Proceedings of the Cognitive Models and Artificial Intelligence Conference*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 155–159.
47. Melinda, M.; Andriyani, N.A.C.; Nurdin, Y.; Khariyunnisa, V.; Yulita, Y.; Enriko, I.K.A. Deep Learning Performance Analysis for Facial Expression Based Autism Spectrum Disorder Identification. *Radioelectron. Comput. Syst.* **2024**, *2024*, 30–40. https://doi.org/10.32620/reks.2024.2.03.
48. Lu, A.; Perkowski, M. Deep Learning Approach for Screening Autism Spectrum Disorder in Children with Facial Images and Analysis of Ethnoracial Factors in Model Development and Application. *Brain Sci.* **2021**, *11*, 1446. https://doi.org/10.3390/brainsci11111446.
49. Sharma, A.; Tanwar, P. Autism Spectrum Disorder Prediction System Using Machine Learning and Deep Learning. *Int. J. Appl. Syst. Stud.* **2024**, *11*, 159–173.
50. Celebrity Face Image Dataset Available online: https://www.kaggle.com/datasets/vishesh1412/celebrity-face-image-dataset (accessed on 1 December 2023).
51. Panetta, K.; Wan, Q.; Agaian, S.; Rajeev, S.; Kamath, S.; Rajendran, R.; Rao, S.P.; Kaszowska, A.; Taylor, H.A.; Samani, A.; et al. A Comprehensive Database for Benchmarking Imaging Systems. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 509–520. https://doi.org/10.1109/TPAMI.2018.2884458.
52. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. https://doi.org/10.1371/journal.pone.0196391.
53. Quatrosi, G.; Genovese, D.; Galliano, G.; Zoppé, H.; Amodio, E.; Bonnet-Brilhault, F.; Tripi, G. Cranio-Facial Characteristics in Autism Spectrum Disorder: A Scoping Review. *J. Clin. Med.* **2024**, *13*, 729. https://doi.org/10.3390/jcm13030729.
54. CDC Chromosomal Abnormalities: Trisomy 21 (Down Syndrome). Available online: https://archive.cdc.gov/www_cdc_gov/ncbddd/birthdefects/surveillancemanual/quick-reference-handbook/trisomy-21-down-syndrome.html (accessed on 3 March 2026).
55. Orphanet: Emanuel Syndrome. Available online: https://www.orpha.net/en/disease/detail/96170?utm_source=chatgpt.com (accessed on 3 March 2026).
56. Kabuki Syndrome: MedlinePlus Genetics. Available online: https://medlineplus.gov/genetics/condition/kabuki-syndrome/ (accessed on 3 March 2026).
57. FAQs • Facial Features • Noonan Syndrome Association. Noonan Syndrome Association. Available online: https://www.noonansyndrome.org.uk/faqs/facial-features/ (accessed on 11 November 2026).
58. Orphanet: Pallister-Killian Syndrome. Available online: https://www.orpha.net/en/disease/detail/884?utm_source=chatgpt.com (accessed on 3 March 2026).
59. Orphanet: Williams Syndrome. Available online: https://www.orpha.net/en/disease/detail/904?utm_source=chatgpt.com (accessed on 3 March 2026).
60. Hypohidrotic Ectodermal Dysplasia: MedlinePlus Genetics. Available online: https://medlineplus.gov/genetics/condition/hypohidrotic-ectodermal-dysplasia/ (accessed on 3 March 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.