



Predicted-gender information in ASD facial-image classification: multitask performance, shuffled-label controls, and representation analysis

Verma A¹, Greenberg C²

Submitted: June 10, 2026, Revised: version 1, August 4, 2026, version 2, September 22, 2026, version 3, September 27, 2026
 Accepted: October 5, 2026

Abstract

Autism spectrum disorder (ASD) facial-image classifiers may benefit from an auxiliary gender-prediction task, but an improvement alone does not establish that meaningful gender information is responsible. This study tested that distinction using a public ASD-labeled facial-image dataset and ImageNet-pretrained EfficientNet-B0 networks. An exploratory comparison found no significant advantage from training separate predicted-male and predicted-female classifiers rather than one general classifier (accuracy: 79.07% versus 82.56%; $p = 0.3075$). A subsequent comparison across 40 matched random seeds on a cleaned 168-image test set found that adding a predicted-gender auxiliary task increased mean accuracy from 77.3% to 79.4% and specificity from 59.0% to 63.5%, while sensitivity remained nearly unchanged. Crucially, an otherwise matched model trained with shuffled gender labels also improved performance over baseline and did not differ significantly from the genuine-label model on any evaluated metric. Thus, meaningful predicted-gender supervision was not shown to explain the observed gain. A separate exploratory analysis removed five-dimensional attribute-associated subspaces from the learned image representations. Predicted-gender subspace removal reduced ASD classification AUC more than removal of comparison-attribute subspaces on average in a pooled analysis, but this association did not establish the cause of the multitask improvement. Together, these findings distinguish the benefit of adding a training task from evidence that its intended information produces that benefit, and distinguish both from information encoded in the learned representation. Results are limited to this dataset and do not establish clinical diagnostic validity.

Keywords

Autism Spectrum Disorder, Facial image classification, Predicted gender, Multitask learning, Gender-specific modeling, EfficientNet-B0, Representation subspace ablation, Computer vision, Transfer learning, Convolutional Neural Network

¹Corresponding author: Aaryan Verma, William Mason High School, 6100 S Mason Montgomery Rd, Mason, OH 45040, USA. aav221009@gmail.com

²Mentor: Clayton Greenberg, Inspirit AI, San Diego, California, USA, and Graduate School of Computer Science, Saarland University, Campus, E1 3, 66123 Saarbrücken, Germany. clayton.greenberg@lsv.uni-saarland.de

1. Introduction

Autism Spectrum Disorder (ASD) is a developmental condition that affects many aspects of communication, behavior, and social interaction (1). Autism is best identified early to help guide support and intervention for children. However, autism is complex, and no single visual feature can diagnose it (1). Recent machine learning studies have explored whether ASD can be classified using facial images (2,3). This study uses facial images as a supervised computer vision classification problem. The model's task is to classify each image according to its dataset label: "autism" or "typical."

ASD is diagnosed more often in males than in females (4). Because of this, models trained on facial images may learn different patterns for predicted-male and predicted-female images. If visual patterns associated with ASD differ by gender, then a gender-specific modeling approach might perform better than one general model. Male and female faces may differ in shape, facial structure, hair patterns, presentation, and other visual features. A model trained on faces may accidentally learn gender-related visual patterns instead of ASD-related patterns. One possible approach is to train separate predicted-male and predicted-female models, although dividing the dataset reduces the number of training images available to each model. Another approach is to use predicted gender as an auxiliary task within one shared model, allowing the full training set to remain available.

Evaluating both approaches against their respective baselines helps assess whether predicted-gender information benefits classification through separate subgroup models or through shared multitask learning. It also

distinguishes improvements in overall classification performance from changes in the balance between false-positive and false-negative classifications.

Across two experimental stages, this study evaluated separate predicted-gender-specific modeling and shared gender-aware multitask learning using different baseline comparisons. Originally, the public dataset used did not include verified gender labels, hence image-derived predicted gender labels were created using an off-the-shelf classification model. Because these labels were not verified demographic labels, the analyses concern predicted-gender information as an image-derived modeling variable rather than as biological sex. The final dataset was balanced across autism/typical class and predicted gender groups.

This study was performed in two experimental stages. The first stage compared a baseline model with separate predicted-gender-specific models. The second stage tested a shared gender-aware multitask model against a newly trained matched baseline using a locked repeated-seed protocol. The shared model and separate gender-specific system were not directly compared within the same locked evaluation. Subsequent to the shared multitask comparison, a shuffled predicted-gender control was used to test whether any observed benefit originated specifically from predicted-gender information or could also result from the presence of an auxiliary learning task with the same architecture and training procedure. A quantitative Grad-CAM analysis was also designed to evaluate whether the baseline and gender-aware models differed in the proportion of attribution falling within the detected facial

region versus in the surrounding image content. An exploratory representation subspace ablation analysis was performed to determine whether removal of a predicted-gender-decodable subspace produced a greater reduction in ASD AUC than the removal of subspaces associated with measured comparison attributes. These analyses were intended to distinguish performance effects associated with multitask training from separate representation-level associations involving predicted-gender information.

1.1 Background

ASD is typically identified through developmental and behavioral evaluation (1). Clinical descriptions of ASD focus on differences in social communication and interaction, alongside repetitive behavior (1). Current standardized tools such as the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) use behavioral observation as one component of ASD evaluation (5). Therefore, facial image classification should not be used as a direct replacement for current methods of diagnosis.

Even with these limitations, facial image classification has been explored as a possible machine learning approach for ASD classification. Rahman et al. used a public dataset of autistic and non-autistic facial images for testing pretrained CNN models including MobileNet, Xception, EfficientNetB0, EfficientNetB1, and EfficientNetB2 (2). Transfer learning is also commonly used in ASD image studies because pretrained models can apply visual patterns learned from larger image datasets to smaller ASD image datasets. Ahmad et al. compared pretrained CNN models

such as ResNet34, ResNet50, AlexNet, MobileNetV2, VGG16, and VGG19 (3).

Some studies use facial attributes instead of raw facial images. Li et al. used facial expressions, arousal, valence, and action units for ASD detection (6). Their work is useful because it shows that facial-related features may contain information useful for ASD classification. Published ASD facial image datasets exist, although with many limitations. Some studies use public datasets that do not have clinically verified diagnoses, demographic data, or image consistency (2,3,7).

Image-based ASD models are hence technically feasible, but they must be interpreted carefully because of these limitations. Such models may learn dataset-specific patterns, including lighting, pose, background, or camera angle instead of ASD features. Prior facial image ASD studies mainly tested different CNN architectures and models. This study builds on that work by comparing separate predicted-gender models with a shared gender-aware multitask approach. Because this study used predicted-gender labels and a public facial image dataset, it should be viewed as a computer vision classification study and not a concrete clinical diagnosis system.

2. Methods

2.1 Dataset

A vision-based facial image dataset for binary classification of children with autism and children without autism was used in this study (7). The original dataset contained a total of 2936 images, split into training, validation, and test splits. The training set consisted of 2536 images, with 1268 autism images and 1268

typical images. The validation set consisted of 100 images, with 50 autism images and 50 typical images. Finally, the test set consisted of 300 images, with 150 autism images and 150 typical images. This original dataset was balanced by class within each split. Although it contained autism and typical labels for each image, it did not include gender labels. Predicted-gender annotations were therefore generated using an off-the-shelf age and gender prediction workflow based on the cited method (8). The original off-the-shelf model produced both gender and age predictions, but this study used only the predicted gender labels for grouping (8). Each facial image within the dataset was assigned a predicted gender label. These labels were used for grouping purposes and not as diagnostic labels.

After the gender labels were assigned to each image, the dataset was organized into four crossed groups: autism/predicted-male, autism/predicted-female, typical/predicted-male, and typical/predicted-female. The annotated training manifest was balanced across ASD status and predicted gender by groupwise replacement resampling to 815 rows per crossed group using random state 42. This produced 3,260 training rows, including 1,630 autism and 1,630 typical rows, representing 1,773 unique image paths. Of the 2,536 source training paths, 763 were not selected, and 1,487 rows repeated a selected image path. The validation and test manifests were downsampled without replacement to 13 and 43 images per crossed group, respectively, producing balanced sets of

52 and 172 unique images. These balanced splits were used for the original baseline and separate gender-specific comparison and for validation-only development of the shared gender-aware model. The cleaned 168-image test set was used for the primary locked comparison between the matched baseline and shared gender-aware model. Exact and near-duplicate procedures for screening were performed on training, validation, and test splits. Exact duplicates were assessed using SHA-256 file hashes, and no exact duplicates were found across the dataset splits. Near-duplicate candidates were identified using 64-bit average hash, difference hash, and DCT-based perceptual hash comparisons. An image pair was flagged when the Hamming distance was 4 or lower for at least one of the three methods. Perceptual-hash screening identified 26 candidate image pairs, and manual review identified four train-test near-duplicate pairs. These were subsequently removed from the primary analysis. After removal of the four manually confirmed train-test near-duplicate images, the cleaned primary test set contained 168 images: 84 autism and 84 typical, with 85 predicted-male and 83 predicted-female images. The four crossed groups contained 43 autism/predicted-male, 41 autism/predicted-female, 42 typical/predicted-male, and 42 typical/predicted-female images. The uncleaned balanced 172-image test set was retained for secondary analysis. Subject-level identity overlap could not be verified independently due to subject identifiers being unavailable. The composition of the original and balanced dataset splits is summarized in Table 1.

Table 1. Composition of original and balanced dataset splits

Original and balanced training, validation, and test split sizes. The balanced training split included repeated image-path rows introduced through groupwise replacement resampling, while validation and test rows remained unique.

Dataset Version	Split	Total rows	Unique image paths	Autism rows	Typical rows	Repeated image-path rows
Original	Training	2536	2536	1268	1268	0
Original	Validation	100	100	50	50	0
Original	Test	300	300	150	150	0
Balanced	Training	3260	1773	1630	1630	1487
Balanced	Validation	52	52	26	26	0
Balanced	Test	172	172	86	86	0

2.2 Workflow

Across two experimental stages, this study evaluated separate predicted-gender-specific models and a shared gender-aware multitask model using distinct baseline comparisons. The separate gender-specific system divided the training data into subgroups by predicted gender. The shared gender-aware model included predicted gender as an auxiliary task without dividing the dataset. The separate gender-specific system was evaluated in the original experiment, while the shared gender-aware model was evaluated later against a separately trained matched baseline under the locked repeated-seed protocol. The shared data-preparation workflow and the two experimental stages are summarized in Figure 1.

The models in both experimental stages used the final balanced facial image dataset. Each facial image in the downloaded dataset was labeled “autism” or “typical.” Each image also had a predicted gender label which was used for subgroup assignment in the separate system and as an auxiliary target in the shared multitask model.

EfficientNet-B0 was used as the base architecture for the baseline, separate predicted-gender-specific, and shared gender-aware model systems, with ImageNet-pretrained weights loaded during model initialization. EfficientNet is an image classification CNN architecture, which balances model depth and width along with image resolution efficiently (9). The original classifier layer was removed and replaced with a task-specific fully connected output layer that produced two class scores. This allowed for the model to use visual features learned from large-scale image pretraining while adapting the final classifier to the specific autism and typical classification task.

For the baseline model, all balanced training data was used. This model learned an autism/typical classifier, which served as a comparison point for the other systems. The gender-specific models were predicted-male-specific and predicted-female-specific models, trained only on their specific gender’s images. During testing, predicted male images were classified by the predicted-male-only model and predicted female images were classified by the

predicted-female-only model. These outputs were combined into one gender-specific result.

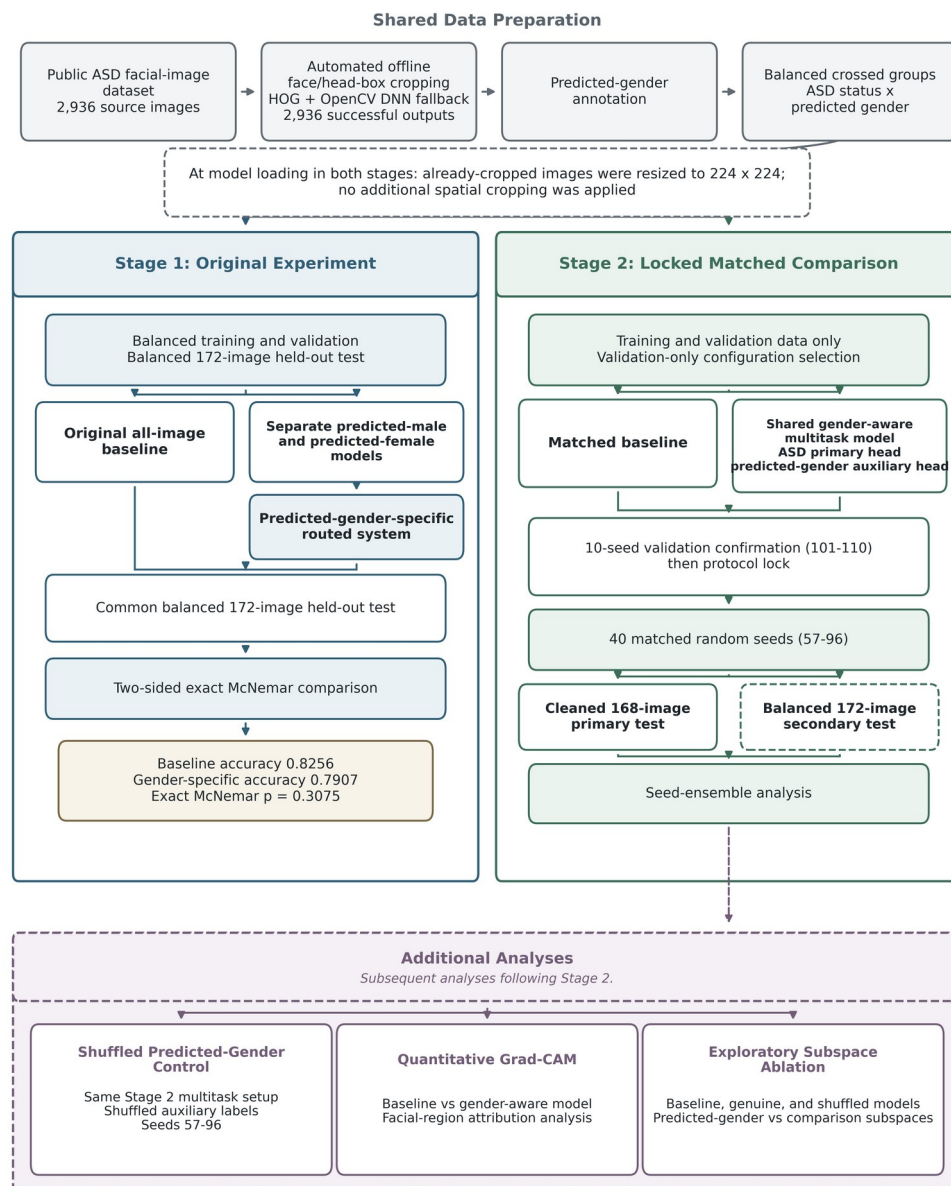


Figure 1. Shared Preparation and Two-Stage Study Workflow. The public ASD facial-image dataset underwent automated offline face/head-box cropping and predicted-gender annotation before balancing across ASD status and predicted gender. Stage 1 compared an all-image baseline with a routed system comprising separate predicted-male and predicted-female models on the balanced 172-image held-out test set. Stage 2 used validation-only selection and confirmation before locking a matched comparison between a newly trained baseline and a shared gender-aware multitask model across 40 random seeds; the cleaned 168-image test set was used for the primary analysis and the uncleaned balanced 172-image test set was kept for secondary analysis. The gender-specific system and shared gender-aware model belonged to different experimental stages and were not directly compared within one locked evaluation. Additional analyses following Stage 2 included the shuffled-label multitask control, quantitative Grad-CAM analysis, and exploratory representation subspace ablation.

Before both experimental stages, the 2936 source images were processed offline using an automated face/head-box cropping workflow. The cropped image tree was used for the original experiment and the later shared gender-aware experiment. The cropper used HOG-based face detection, with an OpenCV DNN face detector and image-scale and CLAHE retries when needed. The retained processing log recorded the crop coordinates and detection method for every image, and all 2936 images were processed successfully without being routed to the configured manual review folder.

Before training, all images were resized to 224 by 224 pixels. The training transform included resizing, random horizontal flipping, small rotation, color jitter, affine translation, tensor conversion, and ImageNet normalization (10). Models were trained using a balanced training split. The validation split was used to monitor training and select an optimal checkpoint. The test split was held out for final evaluation of the model. The same general training procedure was used for the baseline and gender-specific models. The EfficientNet-B0 models were trained using class-weighted cross-entropy loss, with class weights calculated as total samples divided by two times the class count. Optimization used AdamW with an initial learning rate of 3×10^{-4} and weight decay of 1×10^{-4} (11). The batch size was 32, and models were trained for up to 20 epochs with an early stopping patience of 5. Training used an ASD-label WeightedRandomSampler with replacement and a ReduceLROnPlateau learning-rate scheduler with a factor of 0.5 and patience of 2. The best checkpoint was selected based on validation AUC and validation accuracy was used as a fallback.

In the original baseline and separate gender-specific comparison, model performance was determined by accuracy, precision, sensitivity, specificity, F1-score, and AUC on the common balanced held-out test set. Accuracy was calculated as the number of correct predictions divided by the total number of test set images. AUC was calculated to evaluate how well each model separated autism and typical images across classification thresholds. Due to the baseline and gender-specific systems being evaluated on the same test set, McNemar's test was used to compare predictions and determine whether the baseline model significantly differed from the gender-specific model (12).

2.3 Shared gender-aware multitask model

The shared gender-aware multitask model used one EfficientNet-B0 network trained on the full balanced training set. The model contained a shared EfficientNet-B0 feature extractor, a primary autism/typical classification head, and an auxiliary predicted-gender classification head. The shared feature extractor produced 1280 features for each image. These features were passed into two separate linear output heads, with one 1280-to-2 head for autism classification and one 1280-to-2 head for predicted-gender classification. No additional dropout or activation layers were added to the output heads. Autism classification was the primary task, and predicted gender was used only as an auxiliary training signal (13). Images were not divided into separate predicted-male and predicted-female models.

The matched baseline used the same EfficientNet-B0 feature extractor and autism classification head but did not contain the auxiliary predicted-gender head. The combined multitask objective was defined as

$L_{total} = L_{ASD} + 0.2 L_{gender}$ (eq1) where L_{ASD} and L_{gender} represent the class-weighted cross-entropy losses for ASD classification and predicted-gender classification, respectively. For the selected Stage 2 configuration, the predicted-gender auxiliary loss was weighted by 0.2.

For the locked matched-baseline and gender-aware comparison, the selected training settings were a learning rate of 3×10^{-4} , a weight decay of 1×10^{-4} , and standard mixup augmentation. Both the ASD and predicted-gender objectives used class-weighted cross-entropy. Mixup was applied during each training batch with $\alpha = 0.2$ (14). The mixed images were used for both tasks, and the loss for each task was calculated from the two original target labels using the same mixup proportion. The matched baseline and gender-aware models used the same training and validation splits, and the same preprocessing and augmentation. During Stage 2 training, both models used a WeightedRandomSampler with replacement, with sampling weights based on the ASD class labels in addition to the pre-balanced training metadata. Both models used AdamW, a batch size of 32, a maximum of 20 epochs, an early stopping patience of 5, and neither model used a learning-rate scheduler. The best checkpoint was selected primarily using validation AUC. If validation AUC could not be calculated, validation F1-score was used, followed by validation accuracy as the final fallback.

During evaluation, only the autism/typical output head was used to produce the primary classification result, and the predicted-gender auxiliary head did not directly determine the final ASD classification. This architecture

allowed predicted-gender supervision to be incorporated without dividing the training images into smaller subgroup datasets. For each run, the random seed was applied to Python, NumPy, PyTorch, and CUDA random-number generation. The data loaders used zero worker processes. Fully deterministic CUDA execution was not separately enforced.

2.4 Validation model selection and threshold

The candidate gender-aware configurations were evaluated using only the training and validation sets. The test set was not used for hyperparameter selection, checkpoint selection, model comparison, or threshold selection. During the validation confirmation stage, the selected gender-aware configuration was compared with a matched baseline across 10 random seeds, numbered 101 through 110. A prespecified advancement rule required the gender-aware model's sensitivity not to decrease by more than 0.02 relative to the baseline. The selected model passed this rule and advanced to the locked final evaluation.

Classification thresholds were selected using validation predictions only. The candidate thresholds consisted of all unique validation-set ASD probabilities, along with 0, 0.5, and 1. Sensitivity was required to remain at or above 0.90, and among the thresholds that satisfied the requirement, the threshold with the highest specificity was chosen. Ties were resolved using higher precision, followed by higher F1-score and then the higher threshold. If no candidate threshold satisfied the sensitivity requirement, the fallback rule maximized sensitivity, followed by specificity, precision, and F1-score. A complete tie retained the lower threshold. This procedure was applied separately to the baseline and gender-aware models. These

thresholds were fixed before test evaluation and were not changed after the test results were viewed.

2.5 Locked repeated seed and ensemble evaluation

The Stage 2 model configuration, seed range, evaluation endpoints, and threshold-selection rule were fixed before the Stage 2 test evaluation. For each trained run, the numeric threshold was selected using that run's validation predictions before evaluating its test predictions. However, the cleaned 168-image test set was a subset of the 172-image test set previously evaluated in Stage 1 and therefore was not an entirely new test sample. The final run consisted of 40 matched baseline runs and 40 matched gender-aware runs across random seeds 57 through 96. Each baseline run was paired with the matching gender-aware run using the same seed. The primary analysis was performed on the 168-image cleaned test set, while the uncleaned balanced 172-image test set was reserved for secondary analysis. The same trained checkpoints were used for both test sets, along with the same validation-selected thresholds. Hyperparameters and thresholds were not changed after test evaluation, and the model was not reselected after viewing the test results.

The ASD-class softmax probabilities were collected from each of the 40 models, and they were averaged within each model family. This averaging was performed before thresholding. Images from the test set were aligned using their exact file paths.

For the ensemble configuration, the baseline threshold was the mean of 40 baseline validation-selected thresholds. Similarly, the

gender-aware threshold was the mean of 40 gender-aware validation-selected thresholds. These thresholds were not modified using ensemble predictions. This locked protocol separated the final test evaluation from model development which allowed the baseline and gender-aware systems to be compared across repeated seeds under the same evaluation procedure.

2.6 Performance evaluation and statistical analysis

ASD images were treated as the positive class in this study. The reported metrics are accuracy, which captures the overall proportion of correctly classified images. Since accuracy can obscure class-specific behavior, sensitivity and specificity are also reported, the proportions of ASD and typical images correctly identified. Precision, the proportion of ASD predictions that were correct, is combined with sensitivity into the F1-score, their harmonic mean, to summarize the performance on the positive class. Additionally, beyond these metrics dependent on threshold, the AUC reflects the model's ability to rank ASD images above typical ones across all decision thresholds, and the Brier score, which is the mean squared error between predicted ASD probabilities and true labels, measured probabilistic prediction error.

The accuracy, sensitivity, specificity, precision, and F1-score used the validation-selected thresholds. AUC and Brier scores were calculated from the ASD-class probabilities before thresholding. For the final matched seed analysis, the metrics were separately computed for each seed and then summarized across the 40 runs using the mean and sample standard deviation. Seed-level results were compared using two-sided paired t tests across matched

seeds. Mean paired differences and their 95% confidence intervals were calculated from the 40 seed-level differences using the *t* distribution with 39 degrees of freedom. The seed-level *p* values were not adjusted for multiple comparisons. As a *post hoc* sensitivity analysis, Holm adjustment was applied across the seven paired metric comparisons. This adjustment was not part of the prespecified locked protocol. As a secondary same test analysis, the resulting baseline and gender-aware ensemble classifications were compared using a two-sided exact McNemar's test. Because the original baseline and separate predicted-gender-specific systems were evaluated on the same test images, their paired classification outcomes were compared using a two-sided exact McNemar's test. The exact conditional form of McNemar's test was used as the conservative option. The cited methodological comparison recommends mid-*p* and asymptotic variants as generally preferable to the exact conditional test (12). For the ensemble comparison reported here, the mid-*p* variant would have produced a smaller *p*-value than the exact test, hence use of the exact test does not overstate the observed difference. The predicted-male and predicted-female groups within the baseline subgroup analysis contained different images. Accuracy, sensitivity, specificity, and precision were compared using Fisher's exact tests, while the subgroup AUCs were compared using a stratified permutation test with 10,000 permutations.

2.7 Shuffled predicted-gender auxiliary control

An additional shuffled-label multitask control was evaluated to test whether the performance differences observed with the shared gender-aware model depended specifically on predicted-gender information. The shuffled

model used the same EfficientNet-B0 multitask architecture, ASD classification head, auxiliary classification head, auxiliary-loss weight of 0.2, training procedure, and model seeds as the genuine predicted-gender multitask model. The only difference was that the auxiliary training labels were replaced with a fixed shuffled version of the predicted-gender labels.

The shuffled mapping was generated while preserving ASD class and the consistency of repeated image paths, so repeated occurrences of the same image retained the same shuffled auxiliary label. The training manifest preserved the same number of unique images and rows as the genuine-label training set and remained balanced across ASD status and shuffled predicted-gender groups.

The shuffled multitask model was trained across the same 40 model seeds used in the Stage 2 comparison. Performance was evaluated on the cleaned 168-image test set using the same ASD classification metrics. Genuine and shuffled multitask models were compared across matched seeds, and Holm-adjusted results were also calculated across the seven reported performance metrics. This control was performed after the cleaned test set had already been evaluated in Stage 2.

2.8 Grad-CAM analysis

Quantitative non-image Grad-CAM analysis (15) was also performed to determine whether the baseline and genuine predicted-gender multitask models differed in their relative attribution to the detected facial region versus surrounding image content. Using the retained face bounding boxes, the proportion of total Grad-CAM attribution falling within the detected facial region was calculated for

matched model and image evaluation. Attribution maps of zero were treated as unusable. Because the analysis was designed around 40 matched seed pairs per image, the primary comparison required at least 36 usable pairs per image so that each image-level estimate was based on nearly the full repeated-seed evaluation.

2.9 Representation subspace ablation

To further investigate the role of predicted-gender information within the learned representations, an exploratory representation subspace ablation analysis was performed. Related work has used projection-based removal of decodable information from learned representations to investigate how encoded attributes relate to model behavior (16,17). This analysis used the existing 1280-dimensional feature representations from the matched baseline, genuine predicted-gender multitask, and shuffled-label multitask models across the same 40 optimization seeds, producing 120 arm-by-seed analysis units.

For each model, a five-dimensional attribute-associated subspace was estimated using training data only. Separate subspaces were constructed for predicted gender and four comparison attributes. These four attributes were predicted age, image brightness, image sharpness, and six-dimensional crop-box geometry. Partial least squares (18) followed by QR orthogonalization was used to construct each attribute basis. The corresponding subspace was then removed from the learned representation, and ASD classification AUC was recalculated on the cleaned 168-image test set. The reduction in ASD AUC after removal was used to quantify the effect of ablating each attribute-associated subspace.

Several controls were used to determine whether the effects could be explained by removing a five-dimensional component of the representation. Uniform random orthonormal subspaces were used to compare each observed attribute ablation with removal of unrelated subspaces of the same dimensionality. An attribute-associated subspace was considered effective when its observed self-decoding reduction was strictly greater than the 95th percentile of the corresponding Tier 1 random-subspace null. Energy-matched random subspaces were additionally used to account for differences in the amount of representation energy removed, and a variance-rank-matched principal component control (19) was used as a separate comparison with high-variance representation directions.

The primary analysis tested whether removal of the predicted-gender associated subspace produced greater ASD AUC reduction than removal of the eligible comparison-attribute subspaces. The statistical model included an indicator distinguishing predicted gender from the comparison attributes and fixed effects for arm-by-seed blocks. Standard errors were clustered by the 40 optimization seeds, with inference based on the two-sided Student *t* distribution with 39 degrees of freedom and 95% confidence intervals.

This analysis was treated as exploratory since the cleaned 168-image test set had already been evaluated. The findings were interpreted within the context of the selected predicted-gender decodable representation subspace and this specific evaluation setting.

3. Results

3.1 Baseline performance by predicted-gender subgroup

The baseline EfficientNet-B0 model was evaluated separately on predicted-male and predicted-female test images. For predicted-male images, the baseline model achieved an 87.21% accuracy and an AUC of 0.9102, compared with an accuracy of 77.91% and an AUC of 0.8599 for predicted-female images. Although the baseline model's performance was numerically higher for the predicted-male images, the differences in accuracy ($p = 0.1588$), sensitivity ($p = 0.5202$), specificity ($p = 0.2985$), precision ($p = 0.3069$) and AUC ($p = 0.3092$) were not statistically significant.

3.2 Original comparison of the baseline and gender-specific systems

The initial experiment compared the baseline EfficientNet-B0 model to a gender-specific system with predicted-male and predicted-

female models. The test set was balanced by class (autism/typical) and predicted gender, containing 172 images in total. The accuracy of the baseline model on the held-out test set was 82.56%, as compared with the accuracy of the gender-specific system being 79.07%. The gender-specific system achieved a higher AUC than the baseline (0.9158 versus 0.8864) and higher sensitivity (93.02% versus 87.21%), but lower specificity (65.12% versus 77.91%) and precision (72.73% versus 79.79%). A two-sided exact McNemar's test was used to compare the paired predictions from the two systems and did not show a statistically significant difference ($p = 0.3075$). Because the original Stage 1 primary comparison used one trained run per system, it was treated as exploratory and did not assess variability across random initialization. Performance metrics for the original baseline and separate predicted-gender system are shown in Table 2.

Table 2. Original baseline and predicted-gender-specific system performance

Performance of the original baseline EfficientNet-B0 model and predicted-gender-specific system on the same balanced 172-image test set. The paired difference was not significant by a two-sided exact McNemar's test, $p = 0.3075$.

System	Accuracy	Precision	Sensitivity	Specificity	F1-score	AUC
Baseline	0.8256	0.7979	0.8721	0.7791	0.8333	0.8864
predicted-gender-specific	0.7907	0.7273	0.9302	0.6512	0.8163	0.9158

3.3 Duplicate and near-duplicate screening

Exact duplicate screening found no files that were identical across the training, validation, and test splits. Perceptual-hash screening found 26 candidate image pairs. Manual review identified four train-test near-duplicate pairs. The four corresponding test images were removed from the primary analysis, which resulted in a cleaned test set of 168 images. The

uncleaned balanced 172-image test set was retained for secondary analysis.

3.4 Validation selection of the shared gender-aware model

A shared gender-aware multitask model was evaluated as an alternative to the initial experiment's separate gender-specific system. The new model used one EfficientNet-B0

network trained on the full balanced training set, with autism classification being the primary task and predicted gender as an auxiliary task. Validation-only screening and tuning selected a multitask EfficientNet-B0 model with an auxiliary-loss weight of 0.2, a learning rate of 3×10^{-4} , a weight decay of 1×10^{-4} , and standard mixup augmentation.

The selected model was compared with a matched baseline during a separate validation confirmation stage. The gender-aware model achieved a mean sensitivity of 0.938, as compared to 0.927 for the baseline, while both models resulted in the same mean specificity of 0.592 at the constrained operating point. Therefore, the gender-aware model passed the prespecified sensitivity-drop rule and was

advanced to the locked final evaluation. No test set results were used during model selection or confirmation.

3.5 Locked cleaned test comparison

Following validation-only model selection, the final gender-aware model configuration, matched-baseline configuration, evaluation metrics, threshold-selection rule, and random seeds were fixed before the Stage 2 test evaluation; as described in Section 2.5, the test images had previously been evaluated in Stage 1. The final comparison consisted of 40 matched baseline runs and 40 gender-aware runs across seeds 57 through 96, and the cleaned test set with 168 images was used as the primary evaluation set.

Table 3. Matched baseline and gender-aware model performance across 40 seeds

Performance of the matched baseline and gender-aware models across 40 paired random seeds on the cleaned 168-image test set. Mean paired differences are the gender-aware value minus the matched-baseline value, calculated from the unrounded seed-level values, so they may differ by up to 0.001 from the difference of the rounded column means. Confidence intervals were calculated from the 40 paired differences using the t distribution with 39 degrees of freedom. For the Brier score, lower values indicate lower probabilistic prediction error, so a negative difference favors the gender-aware model. Holm-adjusted p-values were calculated as a *post hoc* sensitivity analysis.

Metric	Matched baseline, mean $\pm$ SD	Gender-aware, mean $\pm$ SD	Mean paired difference (95% CI)	Raw p-value	Holm-adjusted p-value
Accuracy	0.773 $\pm$ 0.036	0.794 $\pm$ 0.031	+0.022 (0.007, 0.036)	0.0044	0.0311
Precision	0.702 $\pm$ 0.038	0.726 $\pm$ 0.038	+0.024 (0.008, 0.040)	0.0049	0.0311
Sensitivity	0.956 $\pm$ 0.023	0.954 $\pm$ 0.029	−0.002 (−0.013, 0.010)	0.7535	0.7535
Specificity	0.590 $\pm$ 0.075	0.635 $\pm$ 0.074	+0.045 (0.015, 0.076)	0.0048	0.0311
F1-score	0.808 $\pm$ 0.025	0.823 $\pm$ 0.022	+0.015 (0.004, 0.025)	0.0071	0.0311
AUC	0.902 $\pm$ 0.019	0.909 $\pm$ 0.018	+0.007 (−0.002, 0.015)	0.1190	0.3570
Brier score	0.131 $\pm$ 0.017	0.125 $\pm$ 0.014	−0.005 (−0.012, 0.002)	0.1338	0.3570

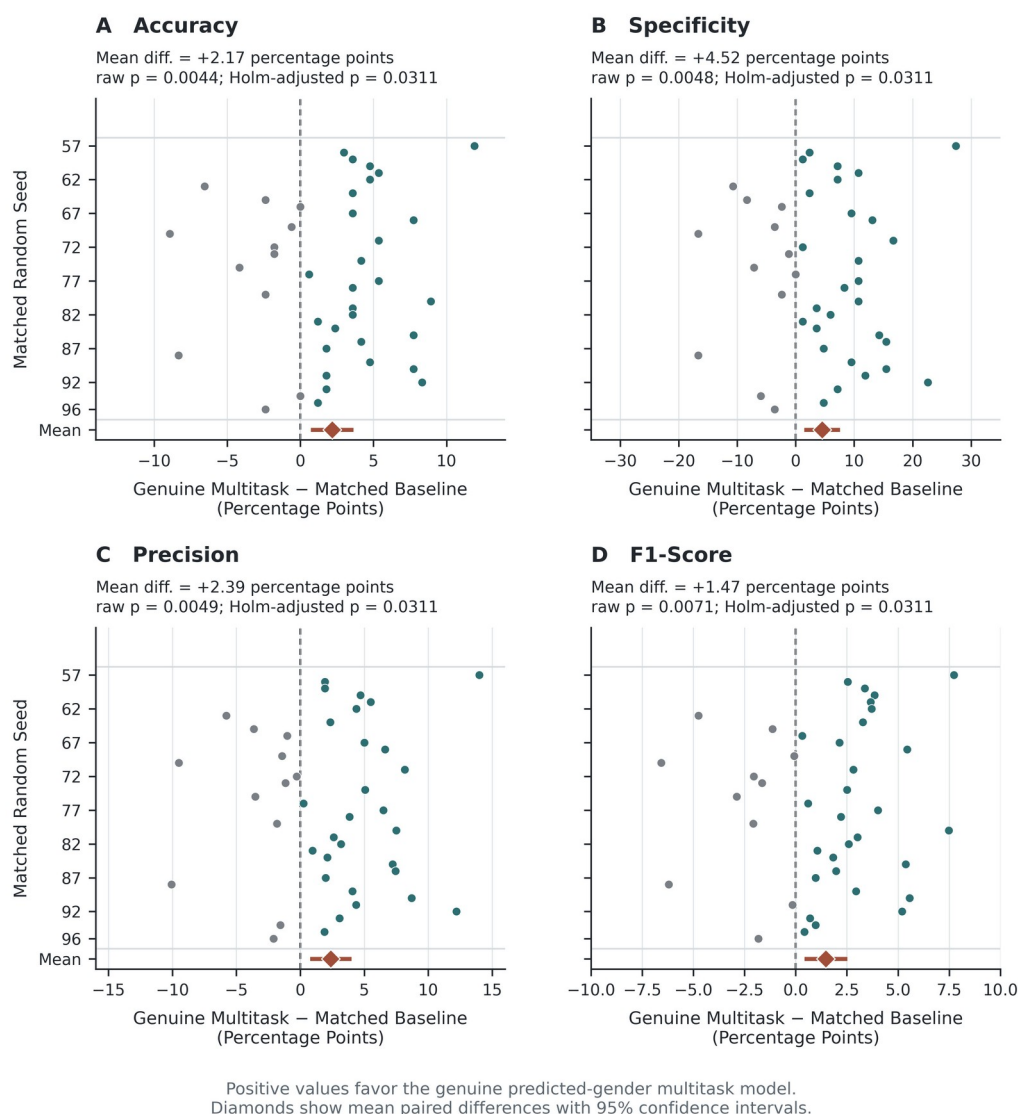


Figure 2. Paired Seed-Level Metric Differences on the Cleaned Primary Test Set. Differences between the gender-aware multitask model and matched baseline are shown for 40 matched random seeds on the cleaned 168-image primary test set. Each point represents one matched seed pair. Diamonds show mean paired differences with 95% confidence intervals calculated from the 40 paired differences. Accuracy, specificity, precision, and F1-score had raw paired p-values below 0.05 and remained statistically significant in a *post hoc* Holm adjustment across the seven evaluated outcomes. Exact p-values are reported within each panel.

The Stage 1 baseline accuracy of 0.8256 and the trained model, the uncleaned balanced 172-Stage 2 matched-baseline mean accuracy of image test set, and the default 0.50/argmax 0.773 should not be interpreted as directly operating point, whereas Stage 2 used 40 comparable replications. Stage 1 used one random seeds, validation-selected sensitivity-

constrained thresholds, mixup augmentation, and the cleaned 168-image primary test set. The observed difference therefore cannot be attributed to a single procedural change.

Across the 40 matched seed pairs, the gender-aware model achieved a higher mean accuracy of 0.794 ± 0.031 , compared with 0.773 ± 0.036 for the matched baseline ($p = 0.0044$). Mean specificity increased from 0.590 ± 0.075 for the baseline to 0.635 ± 0.074 for the gender-aware model ($p = 0.0048$). Mean precision increased from 0.702 ± 0.038 to 0.726 ± 0.038 ($p = 0.0049$), and mean F1-score increased from 0.808 ± 0.025 to 0.823 ± 0.022 ($p = 0.0071$). Mean sensitivity remained nearly unchanged, at 0.956 ± 0.023 for the baseline and 0.954 ± 0.029 for the gender-aware model ($p = 0.7535$).

The gender-aware model also achieved a numerically higher mean AUC of 0.909 ± 0.018 , compared with 0.902 ± 0.019 for the baseline ($p = 0.1190$). Its mean Brier score was numerically lower, at 0.125 ± 0.014 compared with 0.131 ± 0.017 for the baseline ($p = 0.1338$). Neither difference was statistically significant.

As a *post hoc* sensitivity analysis across the seven paired metric comparisons, a Holm adjustment was applied to the raw p-values. The differences in accuracy, specificity, precision, and F1-score remained statistically significant, with adjusted p-values of ~ 0.031 . The differences in sensitivity, AUC, and Brier score remained nonsignificant. Because this adjustment was not prespecified, it was interpreted as a robustness check rather than part of the primary locked analysis. Mean paired differences with 95% confidence intervals for all seven metrics are reported in Table 3. The gender-aware model achieved higher accuracy

in 28 of the 40 matched seed pairs, higher specificity in 28, higher precision in 28, and a higher F1-score in 29. The 40-seed matched comparison is summarized in Table 3, while paired seed-level differences in accuracy, specificity, precision, and F1-score are shown in Figure 2.

3.6 Ensemble and secondary analyses

The gender-aware ensemble demonstrated higher accuracy, precision, sensitivity, specificity, F1-score, and AUC, as well as a lower Brier score, than the baseline ensemble. On the cleaned test set, the baseline ensemble achieved an accuracy of 0.762, sensitivity of 0.964, specificity of 0.560, F1-score of 0.802, and AUC of 0.931. The gender-aware ensemble achieved an accuracy of 0.798, sensitivity of 0.976, specificity of 0.619, F1-score of 0.828, and AUC of 0.935. The gender-aware ensemble correctly classified six images that were misclassified by the baseline ensemble, with the baseline ensemble not uniquely correcting any images that were misclassified by the gender-aware model. A two-sided exact McNemar's test was used to compare the paired ensemble predictions, which showed a statistically significant paired difference ($p = 0.03125$).

The uncleaned balanced 172-image test set was used for secondary analysis, using the same trained checkpoints and validation-selected thresholds. The gender-aware model achieved a mean accuracy of 0.799 ± 0.031 , compared with 0.778 ± 0.035 for the matched baseline. Mean specificity was 0.643 ± 0.073 compared with 0.599 ± 0.074 , mean precision was 0.730 ± 0.038 compared with 0.707 ± 0.038 , and mean F1-score was 0.827 ± 0.022 compared with 0.812 ± 0.025 for the gender-aware and baseline models respectively. Mean sensitivity remained

similar, at 0.955 ± 0.028 for the gender-aware model and 0.957 ± 0.022 for the baseline. Mean AUC was 0.913 ± 0.018 compared with 0.906 ± 0.018 , while mean Brier score was 0.123 ± 0.014 compared with 0.128 ± 0.017 . Because the cleaned test set excluded the four images with overlap concerns, the 168-image results remained the primary findings. Results from the secondary analysis using the uncleaned balanced 172-image test set are summarized in Table 5.

Table 4. Baseline and gender-aware ensemble performance

Performance of the matched baseline and gender-aware ensembles on the cleaned 168-image test set. The paired difference was significant by a two-sided exact McNemar's test, $p = 0.03125$.

Metric	Matched baseline ensemble	Gender-aware ensemble
Accuracy	0.762	0.798
Precision	0.686	0.719
Sensitivity	0.964	0.976
Specificity	0.560	0.619
F1-score	0.802	0.828
AUC	0.931	0.935
Brier score	0.108	0.102

Table 5. Secondary analysis on the uncleaned balanced 172-image test set across 40 matched seeds using the same checkpoints and validation-selected thresholds from the primary analysis.

Metric	Matched baseline, mean $\pm$ SD	Gender-aware model, mean $\pm$ SD
Accuracy	0.778 ± 0.035	0.799 ± 0.031
Precision	0.707 ± 0.038	0.730 ± 0.038
Sensitivity	0.957 ± 0.022	0.955 ± 0.028
Specificity	0.599 ± 0.074	0.643 ± 0.073
F1-score	0.812 ± 0.025	0.827 ± 0.022
AUC	0.906 ± 0.018	0.913 ± 0.018
Brier score	0.128 ± 0.017	0.123 ± 0.014

3.7 Shuffled predicted-gender auxiliary control

The shuffled predicted-gender multitask model was evaluated across the same 40 matched seeds used for the Stage 2 baseline and genuine predicted-gender models. The shuffled model achieved a mean accuracy of 0.802 ± 0.032 , compared with 0.794 ± 0.031 for the genuine model. Mean specificity was 0.655 ± 0.074 for the shuffled model and 0.635 ± 0.074 for the

genuine model, while mean sensitivity remained similar for both models at 0.949 ± 0.027 and 0.954 ± 0.029 .

The genuine-label and shuffled-label models did not differ significantly on any of the seven evaluated metrics, either before or after Holm adjustment. The unadjusted p-values were 0.3010 for accuracy and 0.7478 for AUC.

The shuffled model was also compared with the matched baseline. Mean accuracy, precision, specificity, and F1-score were significantly higher for the shuffled model, with raw p-values of 0.00027, 0.00033, 0.00038, and 0.00041, respectively. These differences remained statistically significant after Holm adjustment. Differences in sensitivity, AUC, and Brier score were not statistically significant. The three-arm results are summarized in Table 6.

Table 6. Baseline, genuine predicted-gender, and shuffled-label multitask performance across 40 paired random seeds on the cleaned 168-image test set.

Metric	Baseline mean $\pm$ SD	Genuine mean $\pm$ SD	Shuffled mean $\pm$ SD	Genuine vs shuffled, p (Holm p)	Shuffled vs Baseline, p (Holm p)
Accuracy	0.773 $\pm$ 0.036	0.794 $\pm$ 0.031	0.802 $\pm$ 0.032	0.3010 (1.0000)	0.00027 (0.00189)
Precision	0.702 $\pm$ 0.038	0.726 $\pm$ 0.038	0.736 $\pm$ 0.039	0.2390 (1.0000)	0.00033 (0.00200)
Sensitivity	0.956 $\pm$ 0.023	0.954 $\pm$ 0.029	0.949 $\pm$ 0.027	0.3843 (1.0000)	0.2592 (0.5184)
Specificity	0.590 $\pm$ 0.075	0.635 $\pm$ 0.074	0.655 $\pm$ 0.074	0.2215 (1.0000)	0.00038 (0.00200)
F1-score	0.808 $\pm$ 0.025	0.823 $\pm$ 0.022	0.828 $\pm$ 0.022	0.3542 (1.0000)	0.00041 (0.00200)
AUC	0.902 $\pm$ 0.019	0.909 $\pm$ 0.018	0.908 $\pm$ 0.014	0.7478 (1.0000)	0.1644 (0.4932)
Brier score	0.131 $\pm$ 0.017	0.125 $\pm$ 0.014	0.128 $\pm$ 0.014	0.4830 (1.0000)	0.4299 (0.5184)

3.8 Grad-CAM analysis

The Grad-CAM analysis produced all 13,440 planned attribution measurements across the baseline and genuine predicted-gender models. Of these, 133 maps had zero total attribution and therefore could not be used to calculate facial-region attribution proportions, including 46 from the baseline models and 87 from the genuine predicted-gender models. Nine of the 168 test images consequently had fewer than the required 36 usable matched seed pairs, with a minimum of 19 matched pairs for one image.

Because these images did not meet the minimum paired-coverage requirement described in the Methods section, the planned primary paired attribution comparison across the full test set was not performed. Therefore, model-level mean or median facial-region attribution proportions, a paired effect estimate, and a confidence interval are not reported.

3.9 Representation subspace ablation

The exploratory representation subspace ablation was completed across all 120 arm-by-seed analysis units. Removal of the predicted-gender-decodable subspace produced greater ASD AUC reduction than removal of the eligible subspaces associated with predicted age, brightness, sharpness, and crop-box geometry. In the primary clustered regression, the predicted-gender coefficient was +0.0126 (95% CI, 0.0075 to 0.0177; $p = 0.000011$).

The mean ASD AUC reduction following predicted-gender subspace removal was 0.363 ± 0.064 for the baseline models, 0.419 ± 0.054 for the genuine predicted-gender models, and 0.443 ± 0.064 for the shuffled-label models. The results for the four comparison attributes are shown in Table 7.

Table 7. ASD AUC reduction after attribute-associated subspace ablation

Removal of five-dimensional subspaces associated with predicted gender and the four eligible comparison attributes across depicted model families.

Ablated attribute	Baseline mean $\pm$ SD	Genuine mean $\pm$ SD	Shuffled mean $\pm$ SD
Predicted gender	0.3625 $\pm$ 0.0639	0.4188 $\pm$ 0.0539	0.4431 $\pm$ 0.0642
Predicted age	0.3232 $\pm$ 0.0591	0.3829 $\pm$ 0.0451	0.4002 $\pm$ 0.0651
Brightness	0.3520 $\pm$ 0.0671	0.4060 $\pm$ 0.0465	0.4299 $\pm$ 0.0641
Sharpness	0.3725 $\pm$ 0.0596	0.4238 $\pm$ 0.0452	0.4410 $\pm$ 0.0567
Crop-box geometry	0.3655 $\pm$ 0.0570	0.4106 $\pm$ 0.0483	0.4382 $\pm$ 0.0667

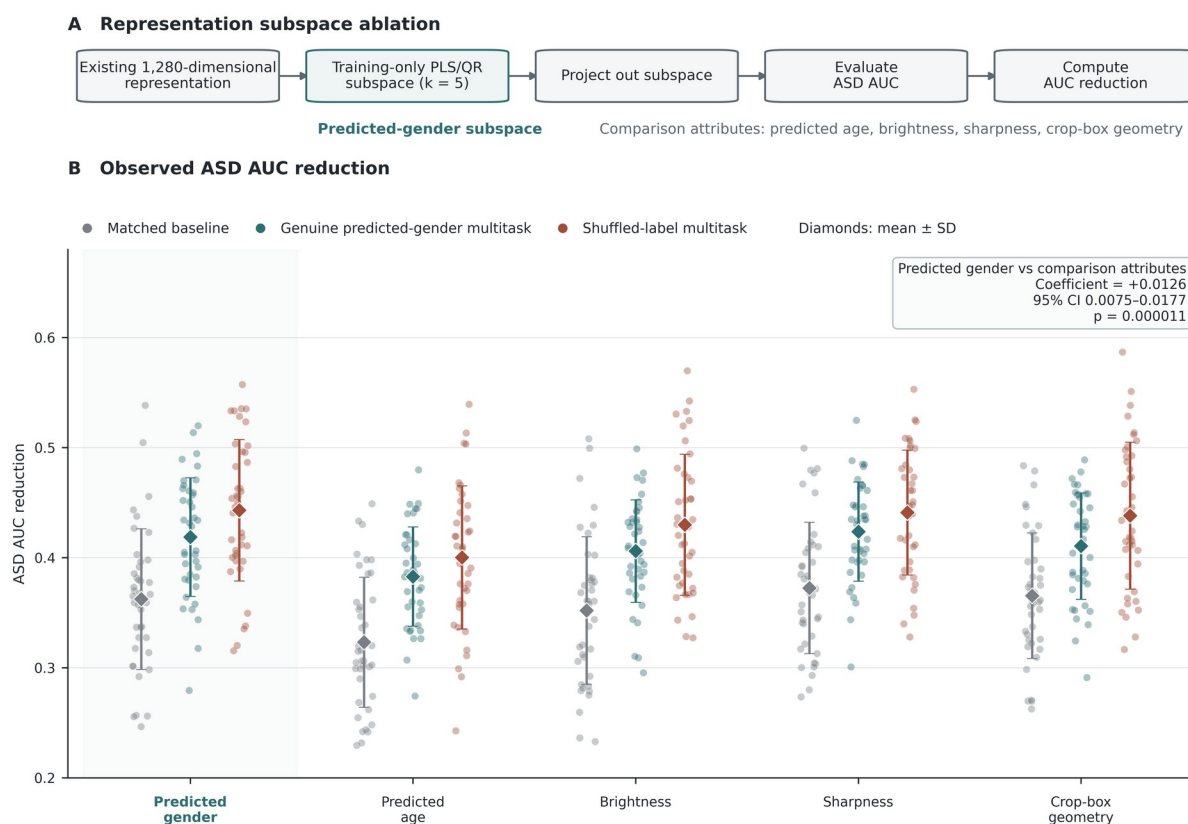


Figure 3. Representation subspace ablation analysis. The exploratory representation subspace ablation analysis is shown for the matched baseline, genuine predicted-gender multitask, and shuffled-label multitask models across 40 random seeds. Five-dimensional subspaces associated with predicted gender, predicted age, brightness, sharpness, and crop-box geometry were removed from the existing representations before ASD AUC was recalculated. Each point represents one arm-by-seed analysis unit, and diamonds show mean AUC reduction $\pm$ SD. Removal of the predicted-gender-decodable subspace produced greater AUC reduction than removal of the eligible comparison-attribute subspaces in the primary clustered regression.

The predicted-gender ablation was also supported by the two retained generic-subspace comparisons. The energy-matched random-subspace and variance-rank-matched PCA controls supported the predicted-gender ablation result, and all four comparison-attribute subspaces met the effectiveness threshold for inclusion in the primary comparison. The representation subspace ablation procedure and the observed arm-by-seed ASD AUC reductions are shown in Figure 3.

4. Discussion

4.1 Multitask performance gains without a demonstrated gender-specific advantage

Two methods of incorporating predicted gender into ASD facial image-classification were evaluated across two experimental stages. The original experiment compared separate gender-specific models with a baseline, while the later locked experiment compared the shared gender-aware multitask model with a newly trained matched baseline. In the exploratory single-run comparison, the separate gender-specific system resulted in a lower accuracy than the baseline, and the paired difference was not statistically significant by McNemar's test. This result did not provide clear evidence of a benefit from splitting the dataset by predicted gender, although the original Stage 1 primary comparison did not assess variability across training runs.

The shared gender-aware multitask model used the full training dataset. Predicted gender was included as an auxiliary task, without being used to divide the data, as was done for the initial gender-specific system. The final shared gender-aware multitask model improved accuracy, specificity, precision, and F1-score,

while keeping the sensitivity nearly unchanged from the baseline. The most important improvement was the gender-aware model's increased specificity, reducing the number of false-positive ASD classifications, without meaningfully increasing the amount of false negatives. These mean differences were estimated across 40 matched seed pairs and also were observed in the ensemble analysis, suggesting that the improvement was not dependent on a single random initialization. However, the shuffled-label control showed that the genuine predicted-gender multitask model did not significantly outperform a multitask model trained with shuffled auxiliary labels on any of the seven evaluated ASD classification metrics. The shuffled-label model also improved accuracy, precision, specificity, and F1 score relative to the matched baseline. These findings did not provide evidence that the Stage 2 performance improvement was specifically attributable to the meaningful content of the predicted-gender labels.

The exploratory representation subspace ablation addressed a separate question about the learned representations. Removal of the predicted-gender-decodable subspace produced a greater reduction in ASD AUC than removal of the eligible subspaces associated with predicted age, brightness, sharpness, and crop-box geometry. The retained random-subspace and variance-rank-matched controls provided additional support for this ablation result within the exploratory analysis. Overall, these results distinguish the observed multitask performance improvement from the representation-level importance of predicted-gender-related information within the dataset.

These results, however, cannot be interpreted as evidence of biological facial differences associated with ASD or gender, and rather should be seen as a difference in model design within a specific dataset. Prior ASD facial image studies mainly compared different CNN architectures or used facial attributes for classification (2,3,6). This study extends that work by examining both how predicted-gender supervision is incorporated into the model and whether the meaningful content of those labels uniquely explains the observed multitask performance differences. Direct numerical comparison with these earlier studies is limited since datasets, data splits, preprocessing, and evaluation procedures were different. These results should be interpreted as a comparison of model designs within this dataset rather than as evidence that one approach would perform better across other ASD facial image datasets.

4.2 Gender-specific model interpretations

One limitation of the separate predicted-gender modeling approach was the reduced number of training images available to each model because the dataset was divided by predicted gender. Each subgroup model therefore learned from a smaller and less unique sample than the baseline, which may have increased instability. Additionally, the less diverse sample may have made the gender-specific models more vulnerable to overfitting and subgroup-specific noise. The gender-specific system had a higher sensitivity than the baseline, with substantially lower specificity and precision. The system may have classified more images as ASD rather than producing a consistently better separation between ASD and typical images. This led to additional true-positive classifications, accompanied by more false-positive classifications. The higher AUC did not

translate into higher accuracy under the default decision rule, using the 0.50/argmax operating point. The baseline showed numerically better performance for predicted-male images than predicted-female images, however, the subgroup difference was not statistically significant. There was no strong evidence that separate models were necessary in the first place. Importantly, gender labels were model-predicted, which may have assigned some images to the wrong subgroup and weakened the case for using separate models.

4.3 Gender-aware model interpretations

Unlike the separate gender-specific models, the shared gender-aware model retained the full training dataset. ASD classification remained the primary task, and predicted gender was included as an auxiliary task. The shared EfficientNet-B0 feature extractor could learn from all available images while also representing variation associated with predicted gender.

The gender-aware model improved specificity and precision at the validation-selected operating point, with little change in sensitivity. Whether this advantage persists at matched sensitivity remains unresolved. However, the shuffled-label control did not provide evidence that these performance improvements were unique to meaningful predicted-gender supervision. The genuine predicted-gender model did not significantly outperform the shuffled-label multitask model on any of the seven evaluated ASD classification metrics, while the shuffled-label model also improved accuracy, precision, specificity, and F1-score relative to the matched baseline. This suggests that the Stage 2 performance advantage may reflect properties of the multitask training setup

more generally, while the comparison did not establish that the advantage depended specifically on the content of the predicted-gender labels.

The results therefore did not establish that predicted-gender supervision provided biologically meaningful ASD information. The auxiliary task may instead have acted as a form of regularization or altered how the model represented features of this dataset. The exact mechanism responsible for improvement was not directly tested. The shuffled-control findings further support caution in assigning the Stage 2 improvement to predicted-gender information itself.

The exploratory representation subspace ablation provided a separate result concerning the learned representation. Removing the predicted-gender-decodable subspace produced a greater reduction in ASD AUC than removing the eligible subspaces associated with predicted age, brightness, sharpness, and crop-box geometry. The retained random-subspace and variance-rank-matched controls supported this effect within the exploratory analysis. These findings show that removal of the predicted-gender-decodable subspace was associated with greater ASD AUC reduction than removal of the eligible comparison subspaces within this exploratory analysis.

The lack of a genuine-over-shuffled performance advantage and the stronger ablation effect of the predicted-gender-decodable subspace are not necessarily contradictory. The ablation result is consistent with an association between predicted-gender-decodable representation structure and ASD discrimination, but it does not establish that this

structure caused the Stage 2 performance improvement. The findings therefore suggest that the shared model's behavior reflects broader changes in its learned representation, with predicted-gender-related information associated with ASD discrimination within the learned representation in this dataset.

4.4 Threshold selection and operating point tradeoffs

Classification performance depended partly on the chosen decision threshold. Thresholds were selected using validation data only, with a prespecified rule that required a sensitivity of at least 0.90 and then selecting the threshold with the highest specificity. This process prevented the final test threshold from being chosen after seeing test results. The gender-aware model achieved higher specificity under the same validation-based threshold-selection rule as the baseline, ensuring a consistent evaluation procedure without establishing superiority at matched sensitivity. The selected threshold objective was appropriate for this study's evaluation goal and it was not established as a clinically optimal cutoff, as a different threshold rule could produce a different sensitivity and specificity balance. The Stage 2 improvement was most evident in threshold-dependent performance measures. AUC and Brier scores did not differ significantly between the models.

Because each model received its own validation-selected cutoff, comparing specificity at matched sensitivity would help determine whether the multitask models provide a better sensitivity-specificity tradeoff or whether the baseline could achieve similar specificity at a different cutoff. This is particularly relevant across the high-sensitivity range used in this study because the AUC

differences were not statistically significant. Future work should therefore compare specificity at matched sensitivity across this high-sensitivity range.

4.5 AUC, probabilistic prediction error, and ensemble interpretations

The AUC for the gender-aware model was numerically higher than the baseline, although the difference was not statistically significant. The gender-aware model also achieved a numerically lower Brier score, yet that was also not statistically significant. There was no statistically significant evidence of improved threshold-independent discrimination or lower probabilistic prediction error. Calibration was not evaluated separately using reliability curves, calibration slope, or calibration intercept. The strongest evidence, therefore, for improvement was in the accuracy, specificity, precision, and F1-score. These results were threshold-dependent, selected at a specific operating point.

The shuffled-label control showed a similar pattern for the threshold-independent measures. AUC and Brier score did not differ significantly between the genuine and shuffled multitask models, and neither measure differed significantly between the shuffled model and baseline. This further suggests that the clearest performance changes associated with multitask training occurred in classification behavior at the selected operating point.

Combining predictions through ensemble analysis reduced the possibility of dependence on one random initialization and showed the same general advantage for the gender-aware model. The gender-aware ensemble uniquely corrected six images misclassified by the baseline ensemble, as compared to the baseline

ensemble being unable to uniquely correct any images misclassified by the gender-aware ensemble, showing a statistically significant paired difference through McNemar's test. Although ensemble analysis strengthens the evidence that the finding was not caused by one seed, it still originated from the same test dataset. Ensembling does not replace independent external validation.

The quantitative Grad-CAM analysis was limited by the number of usable matched seed pairs available for several test images, and the planned primary paired attribution comparison was therefore not performed. The analysis does not support a conclusion about whether the gender-aware model differed from baseline in its relative reliance on the detected facial region and surrounding image content.

4.6 Future directions

Alam et al. (20) reported substantially higher facial-image classification performance, including ~ 98.9% accuracy, sensitivity, and specificity, using preprocessing and augmentation. Their Grad-CAM examples highlighted facial regions including the eyes, nose, lips, and forehead. These results cannot be directly compared numerically with the present study because the image preparation, test sets, and evaluation procedures differed. However, they emphasize an unresolved question in facial-image classification, which is whether apparent performance depends on facial information that generalizes across independently collected images. Highlighting a facial region alone does not establish that the model has identified an ASD-specific facial feature.

The planned quantitative Grad-CAM comparison did not satisfy its paired-coverage

requirement, leaving the relative contributions of facial and surrounding image content unresolved. Future work could directly compare otherwise matched models using either the inner facial region alone or only the image content outside a masked facial region. Performance based on surrounding content would indicate that nonfacial image information predicts labels in this dataset, without establishing why. A separate color-versus-grayscale comparison could test whether recorded color contributes to classification. Any such effect would need to be distinguished from lighting, camera processing, skin-tone differences, and other image-acquisition factors and would not establish differences in facial blood flow or neurotransmitter activity.

The shuffled-label finding also motivates additional tests of the auxiliary-task explanation. Because the annotation workflow produced predicted age as well as predicted gender, future work could compare a predicted-age auxiliary task with a shuffled-age control under the same matched protocol. Another approach could train separate ASD-label and predicted-gender models and then fit a small combining model using their output scores. Comparing this model with one receiving an uninformative or shuffled predicted-gender score could test whether the gender prediction adds information to the final classification decision. Such a model should be fitted using validation or out-of-fold predictions and evaluated on untouched test images. Neither approach should be assumed to improve specificity, because an additional task or score could either have no effect or could change performance.

A classifier fitted to frozen ImageNet-pretrained image features would provide an additional baseline. This could show how much ASD classification performance is available without fine-tuning the image-processing network on this dataset, although the classifier itself would still be trained using the dataset labels. Comparing this baseline with the fine-tuned and multitask models at matched sensitivity could help identify which training steps are associated with the observed performance differences.

The maximum of 20 training epochs and early stopping leave open the question whether model behavior would change with longer training. Extended training and validation curves for the baseline, genuine-label, and shuffled-label models could test for delayed improvement in held-out performance. The present results provide no evidence that such delayed generalization occurred, and longer training could instead encourage memorization of shuffled auxiliary labels. Future models should ultimately be evaluated on independently collected, clinically verified, subject-separated images. Repeated random seeds assess stability across training runs on the current fixed split, whereas grouped cross-validation or an independent test set would assess stability when the held-out children change.

4.7 Limitations

The dataset itself limits the weight of the results and how strongly they can be interpreted. Images used were from a public facial image dataset rather than a clinically controlled autism dataset with verified diagnostics, which may have led to the labels within the public dataset not having the same reliability as labels collected through formal clinical methods. Consequently, the findings cannot confirm

whether the observed classification patterns would generalize to clinically confirmed populations.

The public dataset may also have image-based patterns that are unrelated to autism, including differences in lighting, background, facial pose, the quality of the facial image, hair, clothing, or camera angle. The clinical provenance and diagnostic verification of the dataset-provided autism labels could not be independently confirmed. Therefore, the reported performance may reflect dataset-specific correlations and should not be interpreted as evidence that ASD can be identified from facial appearance.

Automated face/head-box cropping reduced, but did not completely eliminate, surrounding image content. Crop quality was not manually adjudicated image by image in the final logged run, and the detector-derived boxes varied in coverage. The residual background information along with any crop errors and other visual cues specific to the dataset may have influenced the models.

Predicted gender was used instead of verified demographic labels. Due to the public dataset not containing gender labels, an off-the-shelf demographic model was used to produce only predicted gender. The original off-the-shelf model produced both age and gender predictions. All gender labels were predicted based on facial appearance, which may have caused classification errors. If some of the resulting predictions were assigned incorrectly, that may have affected both the training process for the gender-specific model and the subgroup evaluation. Prediction errors may also have affected the auxiliary task used by the gender-

aware model. Therefore, the predicted labels should not be seen as verified biological sex.

The relatively small size of the balanced validation and test set also introduced another limitation. Although balancing made comparison between models more uniform, it also reduced the number of unique images available for evaluation. The original balanced test set consisted of a total of 172 images containing 43 images in each crossed subgroup. After four images with duplicate or near-duplicate overlap were removed, the cleaned primary test set consisted of 168 images. Subject-level identity verification was not available, so additional overlap that could not be identified may have been present. The cleaning reduced overlap concerns, yet due to the small scale of the actual dataset the small number of changed predictions could affect the results.

The training set was balanced by replacement resampling every crossed group to 815 rows, rather than only by oversampling underrepresented groups. Consequently, some selected image paths appeared repeatedly, while 763 of the 2,536 source training paths were not selected. Repeated exposure to selected images may have increased the risk of overfitting or altered the effective training distribution, while omission of other source images reduced the diversity of the training data.

Forty matched seeds reduced dependence on one random initialization. However, repeated-seed evaluation does not replace repeated cross-validation. Additionally, repeated-seed evaluation does not replace testing on a completely independent external dataset. Ensemble analysis also reduced variability across individual trained models, but due to

being evaluated on the same test set as the main analysis it cannot establish external generalizability.

The shuffled-label control was performed after the cleaned 168-image test set had already been evaluated in Stage 2. Its findings provide exploratory follow-up evidence within an already exposed evaluation set and should not be interpreted with the same level of confirmatory status as the locked Stage 2 comparison.

The quantitative Grad-CAM analysis produced the planned attribution measurements, but nine test images did not meet the minimum paired-coverage requirement of 36 usable matched seed pairs. The planned primary paired comparison was therefore not performed, hence the analysis does not support a conclusion about whether the baseline and gender-aware models differed in their relative attribution to facial and surrounding image regions.

The representation subspace ablation was also performed on the previously evaluated cleaned 168-image test set and was treated as exploratory. The active interpretation additionally reflects the post-hoc exclusion of one unreliable generic-subspace control, and the exact fitted attribute basis matrices were not persisted. These factors limit reproducibility and the strength of mechanistic conclusions drawn.

Seven performance metrics were compared across the matched seed pairs using unadjusted paired tests. Therefore, the individual p values should be interpreted cautiously because testing multiple outcomes increases the possibility of a statistically significant result occurring by chance. A *post hoc* Holm-adjusted sensitivity

analysis retained the significant differences in accuracy, specificity, precision, and F1-score. Because this adjustment was not prespecified, it should be interpreted as a robustness check without being part of the primary locked analysis.

Thresholds were chosen using a prespecified validation rule in order to create a consistent comparison between models. A different threshold objective could produce a different balance between sensitivity and specificity, which means the thresholds selected cannot be interpreted as the clinically optimal cutoffs.

Complete end-to-end retraining reproducibility is limited because the exact Stage 2 cloud software environment, locally archived final checkpoints, and original gender-annotation and balancing scripts are not fully available. The automated cropping implementation, crop-coordinate log, and cropped image lineage were recovered, and the Stage 2 manifests and saved prediction artifacts corresponded to the same cropped image set used in Stage 1, although the historical cloud copy itself could not be freshly hash verified. Saved per-image predictions allowed all reported numerical results to be recalculated and numerically verified. The original split-generation script was unavailable. However, the saved training manifest was reproduced at the filepath-multiset level by applying replacement resampling to every crossed training group with random state 42. This was a forensic reconstruction from retained artifacts rather than direct verification of the original code.

Because the separate gender-specific system and shared gender-aware model were evaluated in different experimental stages, the results do

not provide a direct locked comparison of the two approaches. The strongest direct comparison was between the shared gender-aware model and its matched baseline under the locked repeated-seed protocol. The paired seed-level tests quantify consistency across random initializations on this fixed test set and should not be interpreted as population-level inference.

Regarding clinical applicability, facial images alone are not sufficient for ASD diagnosis. This study compared model designs within one dataset, and its findings should not be interpreted as evidence that the model is ready for clinical use.

5. Conclusion

This study shows why an apparent improvement from a predicted-gender task should be tested before being attributed to gender information. Separate predicted-gender classifiers did not demonstrate an advantage over a general classifier. A shared multitask model improved accuracy and specificity relative to its matched baseline, but a model trained with shuffled gender labels also improved these measures and did not differ significantly from the genuine-label model. The principal contribution is therefore a controlled demonstration that better classification after adding a gender task does not, by itself, establish that meaningful gender

information produced the benefit.

The exploratory subspace analysis adds a distinct finding: information associated with predicted gender was encoded in representation directions whose removal impaired classification more than removal of the comparison-attribute subspaces on average. This does not contradict the shuffled-label result. Information can be present in a model's representation without explicit supervision for that information providing an additional performance advantage. Nor does removing those representation directions establish that gender itself caused the observed classification behavior.

Together, these experiments distinguish three questions often conflated in interpreting machine-learning results: whether an additional task improves performance, whether its meaningful labels explain that improvement, and whether associated information is encoded in the model. Within this dataset, the first and third received support, while the second did not. This distinction provides a useful framework for evaluating auxiliary tasks beyond predicted gender. Its generality requires independent testing, and the present findings establish neither autism-specific facial features nor clinical diagnostic validity.

Ethics Statement

This study analyzed an existing, publicly available dataset and involved no recruitment, participant interaction, or collection of new images or health information. No institutional review board review was sought. The available dataset documentation did not provide subject-level identifiers or independently verifiable clinical records, and documentation regarding consent and ethical oversight for the original image collection was not available. The dataset-provided autism and typical labels were therefore analyzed as given, and no clinical or diagnostic claim is made. No facial images are reproduced or redistributed in this article or its supplementary materials. Because the source images depict identifiable minors and the available source documentation did

not establish research consent, the results are reported strictly as a comparison of model designs and should not be used to support any screening or diagnostic application.

Data Availability

The source facial images were obtained from Autistic Children Facial Image Dataset on Kaggle (7) and are not redistributed by this study. Predicted-gender annotations and balanced split manifests were derived for this work. Manifests containing source-relative image identifiers, labels, and derived annotations can be shared without bundling image files, subject to review of the source dataset terms and other policies.

6. References

1. Centers for Disease Control and Prevention. (2025). *Clinical testing and diagnosis for autism spectrum disorder*. <https://www.cdc.gov/autism/hcp/diagnosis/index.html>
2. Mujeeb Rahman, K. K., & Subashini, M. M. (2022). Identification of autism in children using static facial features and deep neural networks. *Brain Sciences*, 12(1), 94. <https://doi.org/10.3390/brainsci12010094>
- 3*. Ahmad, I., Rashid, J., Faheem, M., Akram, A., Khan, N. A., & Ul Amin, R. (2024). Autism spectrum disorder detection using facial images: A performance comparison of pretrained convolutional neural networks. *Healthcare Technology Letters*, 11(4), 227–239. <https://doi.org/10.1049/htl2.12073> **Retracted** August 17, 2026, owing to concerns about diagnostic-label validity and consent for children's images. Retraction notice: <https://doi.org/10.1049/htl2.70095>
4. Shaw, K. A., Williams, S., Patrick, M. E., Valencia-Prado, M., Durkin, M. S., Howerton, E. M., Ladd-Acosta, C. M., Pas, E. T., Bakian, A. V., Bartholomew, P., Nieves-Muñoz, N., Sidwell, K., Alford, A., Bilder, D. A., DiRienzo, M., Fitzgerald, R. T., Furnier, S. M., Hudson, A. E., Pokoski, O. M., . . . Maenner, M. J. (2025). Prevalence and early identification of autism spectrum disorder among children aged 4 and 8 years—Autism and Developmental Disabilities Monitoring Network, 16 sites, United States, 2022. *MMWR Surveillance Summaries*, 74(2), 1–22. <https://doi.org/10.15585/mmwr.ss7402a1>
5. Maddox, B. B., Brodtkin, E. S., Calkins, M. E., Shea, K., Mullan, K., Hostager, J., Mandell, D. S., & Miller, J. S. (2017). The accuracy of the ADOS-2 in identifying autism among adults with complex psychiatric conditions. *Journal of Autism and Developmental Disorders*, 47(9), 2703–2709. <https://doi.org/10.1007/s10803-017-3188-z>
6. Li, B., Mehta, S., Aneja, D., Foster, C., Ventola, P., Shic, F., & Shapiro, L. (2019). *A facial affect analysis system for autism spectrum disorder* [Preprint]. arXiv.

<https://arxiv.org/abs/1904.03616>

7. Das, P. (n.d.). *Autistic children facial image dataset* [Data set]. Kaggle. <https://www.kaggle.com/datasets/prayashdas/autistic-children-facial-image-dataset>
8. Saxena, U. (2024). *Age & gender detection | Multi-model prediction* [Computer code]. Kaggle. <https://www.kaggle.com/code/utkarshsaxenadn/age-gender-detection-multi-model-prediction>
9. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6105–6114). PMLR. <https://proceedings.mlr.press/v97/tan19a.html>
10. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252. <https://doi.org/10.1007/s11263-015-0816-y>
11. Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *7th International Conference on Learning Representations*. <https://openreview.net/forum?id=Bkg6RiCqY7>
12. Fagerland, M. W., Lydersen, S., & Laake, P. (2013). The McNemar test for binary matched-pairs data: Mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13, Article 91. <https://doi.org/10.1186/1471-2288-13-91>
13. Caruana, R. (1997). Multitask learning. *Machine Learning*, 28, 41–75. <https://doi.org/10.1023/A:1007379606734>
14. Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations*. <https://openreview.net/forum?id=r1Ddp1-Rb>
15. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). IEEE. <https://doi.org/10.1109/ICCV.2017.74>
16. Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020). Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7237–7256). Association

for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.647>

17. Elazar, Y., Ravfogel, S., Jacovi, A., & Goldberg, Y. (2021). Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9, 160–175. https://doi.org/10.1162/tacl_a_00359

18. Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109–130. [https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1)

19. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), Article 20150202. <https://doi.org/10.1098/rsta.2015.0202>

20. Alam, M. S., Rashid, M. M., Faizabadi, A. R., Mohd Zaki, H. F., Alam, T. E., Ali, M. S., Gupta, K. D., & Ahsan, M. M. (2023). Efficient deep learning-based data-centric approach for autism spectrum disorder diagnosis from facial images using explainable AI. *Technologies*, 11(5), 115. <https://doi.org/10.3390/technologies11050115>

***RETRACTION:** I. Ahmad, J. Rashid, M. Faheem, A. Akram, N. A. Khan, and R. Amin, “Autism Spectrum Disorder Detection Using Facial Images: A Performance Comparison of Pretrained Convolutional Neural Networks,” *Healthcare Technology Letters* 11, no. 1 (2024): 227–239, <https://doi.org/10.1049/htl2.12073>.

The above article, published online on 08 January 2024 in Wiley Online Library (wileyonlinelibrary.com), has been retracted by agreement between the journal Editor-in-Chief, Christopher James; The Institution of Engineering and Technology; and John Wiley & Sons Ltd. The retraction was agreed upon due to concerns regarding both the reliability of the study and significant privacy issues. The dataset used in the study contains images of children purported to have, or not have autism spectrum disorder (ASD), sourced from the internet without any documented clinical history or independent verification of an ASD diagnosis, raising concerns about the scientific validity of the dataset. In addition, the images of minors have been used and presented in the article without consent from the individuals or their parents/legal guardians. For these reasons, the article has been retracted, and the corresponding images of children have been redacted. The authors have been informed of the decision, and they disagree with the retraction.