Contrastive Representation Learning for Voice-Based Autistic Trait Identification

Hajarimino Rakotomanana and Ghazal Rouhafzay *

Posted Date: 15 April 2026

doi: 10.20944/preprints202604.1071.v1

Keywords: Autism Spectrum Disorder; contrastive learning; representation learning; Time-Frequency Consistency; supervised contrastive loss; SupCon; ReCANVo

 doi:10.20944/preprints202604.1071.v1

Article

Contrastive Representation Learning for Voice-Based Autistic Trait Identification

Hajarimino Rakotomanana and Ghazal Rouhafzay *

Department of Computer Science, Université de Moncton, NB, E1A 3E9, Canada
* Correspondence: ghazal.rouhafzay@umoncton.ca

Abstract

Early identification of Autism Spectrum Disorder (ASD) traits in infants is crucial for early intervention, which can greatly improve the child's quality of life. Solutions that use voice analysis offer a promising non-invasive way to detect ASD. However, most current studies depend on extracting specific voice markers from certain datasets and do not include validation across different groups. In this paper, we propose a supervised contrastive learning method for identifying ASD based on infant vocalizations. We extend the Time-Frequency Consistency (TF-C) framework from self-supervised learning to a contrastive approach that uses labels. Our method takes advantage of both time-related and frequency-related data through a dual-branch encoder. It applies supervised contrastive constraints during pre-training to reduce variation within classes while boosting separation between different classes in the embedding space. We pre-train the model using diagnostic labels on a dataset that includes typically developing (TD), Attention-Deficit Hyperactivity Disorder (ADHD), and ASD infants from an open-access dataset, and then fine-tune it with a simple classification head. Evaluation on a cross-cohort group of participants shows the model generalizes well and can distinguish ASD from non-ASD infants, achieving up to 100.00 % accuracy on non-verbal vocalizations.

Keywords: Autism Spectrum Disorder; contrastive learning; representation learning; Time-Frequency Consistency; supervised contrastive loss; SupCon; ReCANVo

1. Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that can impact social and communication behaviours, intellectual performance and other mental or physical health aspects of life. Early diagnosis is critical, as timely intervention can significantly improve the well-being and health of the child. However, current diagnostic procedures rely primarily on behavioral assessments conducted by trained clinicians, which are time-consuming, resource-intensive, and often inaccessible in remote areas or require a long waiting time. These limitations have motivated many researchers to explore the potential of leveraging Artificial Intelligence-driven solutions to support early and scalable ASD screening.

Among potential biomarkers that can be easily sensed and explored using AI, vocal characteristics have particularly raised interest as a promising diagnostic indicator. Differences in prosody, pitch variability, temporal dynamics, and spectral features are suggested to be distinctive features in the speech of children with ASD [3–5]. These research works further suggest that such distinctive vocal characteristics are not limited to verbal speech but might also be available in pre-linguistic and non-verbal voice patterns. Accordingly, voice biomarkers can have strong potential to provide early diagnostic indicators, even before full language development in children.

Recent advances in deep learning, such as the development of more efficient neural network architectures and improved training algorithms, have significantly improved automatic speech and audio classification systems. Convolutional and recurrent architectures, known for their ability to capture temporal and spatial features, have demonstrated competitive performance for ASD detection

 doi:10.20944/preprints202604.1071.v1

when trained on spectrogram representations. [7,8]. However, many of these approaches rely on supervised classification objectives that do not explicitly structure the embedding space, which may limit the model's ability to generalize to unseen individuals. This challenge is particularly critical in clinical settings, where factors such as inter-speaker variability, inconsistent recording conditions, and cohort heterogeneity can significantly degrade model performance by introducing noise and variability that the model has not been trained to handle.

Self-supervised representation learning, a burgeoning field in machine learning, has gained significant attention for its ability to effectively utilize vast amounts of unlabeled data, offering a promising alternative to traditional supervised methods. The Time–Frequency Consistency (TF-C) framework [1] demonstrated that enforcing consistency between temporal and spectral views of a signal improves representation robustness. Similarly, contrastive learning methods, particularly Supervised Contrastive Learning (SupCon) [2], have demonstrated that directly incorporating label information into the contrastive objective enhances class separation and generalization by creating more distinct feature representations than those achieved through standard cross-entropy training.

Building upon these advances, this work extends the TF-C architecture from a self-supervised paradigm to a fully supervised contrastive learning framework. This framework is specifically tailored for the detection of ASD from infant vocalizations. We take advantage of complementary temporal and spectral representations and integrate diagnostic labels during representation learning. This approach aims to create embeddings that are both discriminative and robust to inter-speaker variability. The proposed approach is evaluated on multi-cohort infant datasets, which include both structured speech and spontaneous non-verbal vocalizations. The final performance is assessed on completely unseen individuals to rigorously measure generalization capability.

2. Literature Review and Key Issues

2.1. Vocal Biomarkers for ASD Detection

Acoustic analysis has been extensively investigated as a potential tool for ASD screening. Early studies relied on handcrafted features such as Mel-frequency cepstral coefficients (MFCCs), pitch statistics, jitter, shimmer, and prosodic descriptors [9,14]. Traditional machine learning classifiers, including Support Vector Machines, Random Forest and XGBoost achieved reported accuracies ranging from 65% to 90% depending on dataset size and task formulation [10,15,16].

More recent works have adopted deep learning approaches using spectrogram-based convolutional neural networks (CNNs) or hybrid CNN–RNN models [11,12]. Reported accuracies in controlled experimental settings often range between 75% and 90%. However, performance frequently drops when models are evaluated on strictly unseen subjects, highlighting generalization challenges. Clinical-grade studies emphasize the importance of subject-independent splits to avoid overestimating performance [17,18].

Importantly, several studies have shown that ASD-related acoustic markers are present not only in structured speech tasks but also in spontaneous and non-verbal vocalizations [14]. This finding is particularly relevant for infant cohorts and pre-linguistic populations, where structured speech may not be available [19]. Datasets such as ReCANVo (Real-World Communicative and Affective Nonverbal Vocalizations) [23] focus specifically on spontaneous infant vocalizations, providing ecologically valid but acoustically heterogeneous data.

2.2. Self-Supervised and Contrastive Representation Learning

Contrastive learning has emerged as a powerful paradigm for representation learning. The SimCLR framework [20] demonstrated that instance discrimination via data augmentation can yield high-quality embeddings without labels. Building on this idea, Supervised Contrastive Learning (SupCon) [2] introduced label-informed contrastive objectives, showing improved robustness and higher accuracy compared to cross-entropy baselines across multiple benchmarks.

 doi:10.20944/preprints202604.1071.v1

In the neuroimaging domain, self-supervised and contrastive graph transformer methods have demonstrated strong potential for extracting meaningful representations from brain connectivity data, improving the detection of autism spectrum disorder [21]. Inspired by these advances, in the audio domain, the TF-C framework [1] enforces consistency between temporal and frequency representations of time-series data, leading to enhanced downstream classification performance. However, TF-C was originally designed for a purely self-supervised setting and does not leverage label information during representation learning. This omission potentially limits the model's ability to achieve enhanced performance through supervised contrastive objectives, which could otherwise offer significant improvements in representation quality.

Recent medical AI studies suggest that embedding-level supervision can enhance class separability and improve performance under limited data regimes [2]. Nevertheless, applications of supervised contrastive learning to infant vocal ASD detection remain limited.

2.3. Generalization to Unseen Individuals: A Critical Challenge

A central issue in clinical machine learning is the challenge of generalizing to unseen individuals, due to the inherent variability in patient data and the complexity of medical conditions. Many reported high accuracies are obtained using random sample splits rather than subject-independent partitions, potentially inflating results. Studies that enforce strict subject-level separation often report reduced performance, underscoring the importance of robust embedding learning strategies.

Inter-speaker variability, differences in recording environments, heterogeneity in developmental stages, and class imbalance introduce noise and bias, complicating the accurate detection of ASD from vocal signals. Models trained solely with cross-entropy loss may inadvertently focus on spurious correlations or speaker-specific artifacts, as this loss function does not inherently promote generalization across diverse data. Contrastive objectives help structure the embedding space to enforce intra-class compactness and inter-class separability, which may enhance robustness under distribution shifts.

2.4. Key Issues Addressed in This Work

Based on the literature, several key challenges remain:

- **Limited robustness to inter-speaker variability:** Existing supervised classifiers may overfit to cohort-specific patterns.
- **Under-explored use of supervised contrastive learning in ASD vocal analysis.**
- **Need for cross-context modeling combining structured and spontaneous vocalizations.**

To address these issues, we extend the TF-C architecture into a supervised contrastive framework that integrates diagnostic labels during representation learning. By jointly optimizing temporal and spectral representations under label-informed contrastive constraints, the proposed approach aims to learn discriminative and generalizable embeddings suitable for clinical deployment. Final performance is evaluated on completely unseen infants to rigorously assess real-world applicability.

3. Framework

Our framework builds upon the Time-Frequency Consistency (TF-C) architecture originally introduced by Zhang et al. (2022) [1], extending it from a self-supervised representation learning setting to a fully supervised contrastive learning paradigm. Specifically, we leverage the complementary nature of temporal and spectral voice representations to enforce cross-view consistency while explicitly incorporating diagnostic labels during training.

The proposed objective is designed to identify traits associated with Autism Spectrum Disorder (ASD) by simultaneously minimizing intra-class variability and maximizing inter-class separability in the learned embedding space. By integrating label-informed contrastive constraints, the model promotes more discriminative representations and improves robustness to inter-speaker variability, thereby enhancing generalization to previously unseen individuals.

 doi:10.20944/preprints202604.1071.v1

Figure 1 provides an overview of the proposed pipeline. The architecture consists of three main stages: (i) time–frequency feature extraction and dual-branch encoding, (ii) supervised contrastive representation learning, and (iii) downstream fine-tuning with a lightweight classification head. The figure illustrates how temporal and spectral representations are jointly optimized under contrastive supervision before being transferred to the diagnostic classification task.

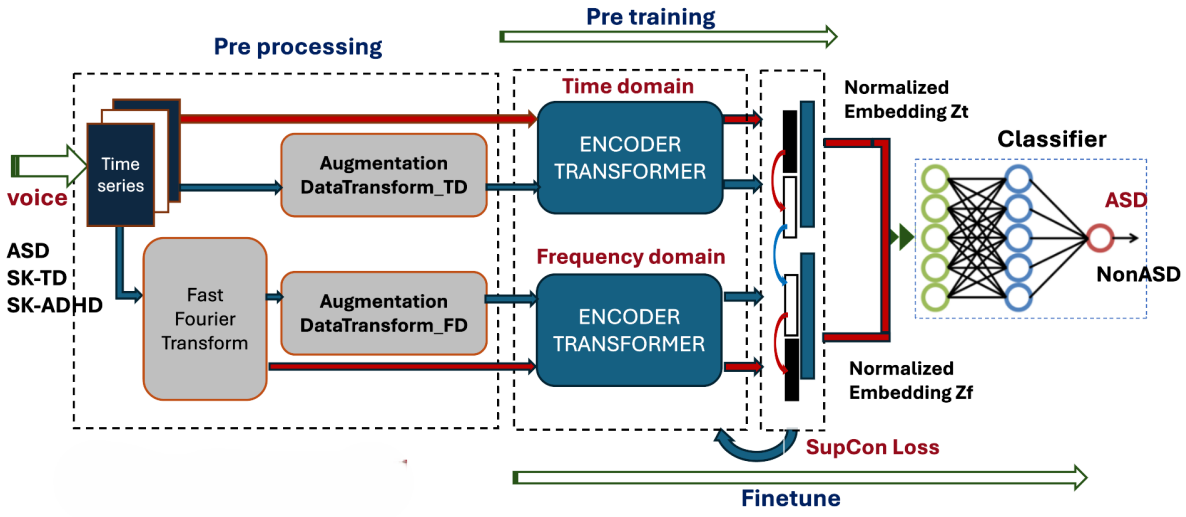

Figure 1. Overview of the proposed supervised Time–Frequency Consistency framework. The model jointly learns temporal and spectral embeddings under supervised contrastive constraints, followed by fine-tuning with a task-specific classifier.

3.1. Dataset

We create a multi-cohort pediatric corpus composed of publicly available speech and voice datasets. Recordings for ASD, ADHD, and TD children are primarily drawn from the SK-Dutch dataset [22], which consists of structured Dutch child speech productions. Publicly available datasets containing complete voice recordings of children are extremely rare. To strengthen the research design, we therefore incorporate the ReCANVo dataset [23], a publicly available resource in autism research that contains non-verbal vocalizations from children with ASD. This dataset is used to extend our experimental evaluation and to assess the robustness of the proposed network more reliably. This multi-cohort design allows us to force the model to learn more generalizable acoustic markers of neurodevelopmental vocal patterns in ASD. To ensure methodological rigor and prevent data leakage, we adopt an individual-level partitioning strategy, where training, validation, and test splits are performed at the child level. Consequently, no segments from the same participant appear across different subsets. Furthermore, we enforce a balanced class distribution such that:

$$|D_{ASD}| = |D_{ADHD} \cup D_{TD}| \tag{1}$$

This ensures that the model learns discriminative features rather than frequency biases inherent in unbalanced datasets.

3.2. Preprocessing

The preprocessing pipeline transforms raw audio into structured inputs for the dual-encoder:

1. **Digitization:** Audio signals are resampled at 16,000 Hz using the Librosa library.
2. **Segmentation:** Continuous recordings are sliced into 0.5s chunks with a 0.1s hop length, creating a high-density time series.
3. **Domain Transformation:** Each time-series segment is converted into its frequency-domain representation via Fast Fourier Transform (FFT).

 doi:10.20944/preprints202604.1071.v1

4. **Data Augmentation:** Stochastic augmentations are applied to both domains to increase the diversity of the contrastive pairs.
5. **Shuffling & Stratification:** Data is shuffled and split using a stratified approach to preserve class distribution across batches ($B = 256$).

Participant Distribution and Data Split.

To ensure strict subject-level generalization, the dataset was partitioned at the individual level. A total of 28 participants which represent 21.7% of the participants in the training were held out exclusively for the inference experiment: 7 SK-TD, 7 SK-ADHD, 7 SK-ASD (Dutch), and 7 ReCANVo-ASD. This subject-wise split prevents data leakage and ensures that no vocal segments from the same child appear across experimental stages. The remaining participants were used for pre-training and fine-tuning.

Table 1 summarizes the cohort distribution and data partitioning strategy.

Table 1. Participant Distribution and Individual-Level Data Split.

Group	Total	Train/Val	Inference
SK-TD	38	31	7
SK-ADHD	37	30	7
SK-ASD (Dutch)	46	39	7
ReCANVo-ASD	8	-	7
Total	129	100	28

3.3. Contrastive Model Architecture

The model architecture consists of a dual-stream Transformer encoder designed to capture the cross-domain consistency.

- **Pre-training Phase:** Two separate Transformer-based encoders process the time-series and the frequency spectra. The resulting latent vectors are normalized and projected into a joint embedding space. Instead of a self-supervised loss, we implement a **Supervised Contrastive (SupCon) Loss**, which encourages embeddings of the same class (e.g., ASD) to cluster together while pushing different classes apart.
- **Fine-tuning Phase:** The pre-trained encoders are coupled with a linear classifier. The final classification (presence or absence of ASD) is performed by concatenating the temporal and spectral representations, optimized again via the SupCon objective to refine decision boundaries.

3.3.1. Supervised Contrastive Loss (SupCon) in Pre-Training

Unlike traditional self-supervised contrastive learning (e.g., SimCLR) which only considers two augmented versions of the same sample as a positive pair, the Supervised Contrastive (SupCon) loss leverages label information to define positive pairs as all samples belonging to the same class within a multi-view batch.

For a batch of N samples, let $i \in I \equiv \{1 \dots 2N\}$ be the index of an augmented sample. Let z_i be its latent representation and y_i its label. The SupCon loss is formulated as:

$$\mathcal{L}_{SupCon} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i) \setminus \{i\}} \exp(z_i \cdot z_a / \tau)} \tag{2}$$

where:

- $P(i) = \{p \in A(i) \setminus \{i\} : y_p = y_i\}$ is the set of indices of all positives in the multiview batch relative to anchor i.
- $A(i)$ is the set of all indices in the batch.
- $\tau \in \mathbb{R}^+$ is a scalar temperature parameter that controls the difficulty of the contrastive task.

 doi:10.20944/preprints202604.1071.v1

By encouraging $|P(i)|$ many positive pairs to cluster together, the model learns a representation space where ASD-positive voice segments are intrinsically separated from TD and ADHD samples, facilitating the downstream classification task.

3.3.2. Fine-Tuning and Classification

Following the supervised contrastive pretraining stage, the learned encoder is fine-tuned for the downstream diagnostic classification task. The objective of this stage is to map the learned time–frequency embeddings to clinically meaningful class labels (e.g., ASD vs. non-ASD, or multi-class settings including ADHD and TD), while preserving the discriminative structure enforced during contrastive training.

Downstream Classifier Architecture.

The downstream classifier is implemented as a lightweight multi-layer perceptron (MLP) placed on top of the frozen or partially fine-tuned encoder. The final decision is based on the concatenated vector $Z_{total} = [z_t; z_f]$. Let $\mathbf{z} \in \mathbb{R}^{2\times128}$ denote the joint time–frequency embedding produced by the backbone network, where the dimensionality reflects the concatenation of temporal and spectral representations. The classifier first flattens this representation into a vector $\mathbf{z}_{flat} \in \mathbb{R}^{256}$:

$$\mathbf{z}_{flat} = \text{reshape}(\mathbf{z}). \tag{3}$$

The classification head consists of two fully connected layers:

$$\mathbf{h} = \sigma(\mathbf{W}_1\mathbf{z}_{flat} + \mathbf{b}_1), \quad \mathbf{W}_1 \in \mathbb{R}^{64\times256}, \tag{4}$$

$$\hat{\mathbf{y}} = \mathbf{W}_2\mathbf{h} + \mathbf{b}_2, \quad \mathbf{W}_2 \in \mathbb{R}^{C\times64}, \tag{5}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function, $\mathbf{h} \in \mathbb{R}^{64}$ is the hidden representation, and C corresponds to the number of target classes. The final output $\hat{\mathbf{y}}$ represents the unnormalized logits used for supervised optimization via the cross-entropy loss.

This architecture is intentionally compact to reduce the risk of overfitting, particularly given the moderate size of pediatric voice datasets. By limiting the classifier depth and parameter count, we ensure that most representational capacity remains within the contrastively trained encoder, while the MLP serves primarily as a task-specific decision layer.

Fine-tuning Strategy.

During fine-tuning, the encoder and classifier are jointly optimized using the Adam optimizer. The learning rate is initialized at 1×10^{-3}, balancing convergence speed and training stability. The model is trained for 25 epochs with a batch size of 256. Hardware acceleration is enabled to ensure efficient computation.

To prevent data leakage and ensure generalization to unseen individuals, training, validation, and test splits are performed at the subject level. The contrastive temperature parameter $\tau = 0.07$, selected during the representation learning phase, is maintained to preserve embedding geometry consistency.

3.4. Experimental Configuration

Table 2 summarizes the key experimental settings used throughout training and fine-tuning. We resample audio signals at $f_s = 16KHz$. Each recording is segmented into 0.5-second chunks with a hop length of 0.1 seconds, enabling partial overlap and increasing the number of training samples while preserving temporal continuity. This parameter choice is intended to allow the network to capture short-term acoustic and prosodic patterns, including brief pre-linguistic and paralinguistic vocal events, while preserving sufficient temporal structure for effective representation learning. The use of overlapping windows increases sample coverage without compromising subject-level independence, since the train–test splits are performed at the subject level. A fixed random seed (42) is used to ensure deterministic initialization and data shuffling, facilitating reproducibility across runs. The

 doi:10.20944/preprints202604.1071.v1

combination of standardized preprocessing, controlled optimization settings, and explicit architectural design ensures that the experimental protocol can be reliably replicated.

Table 2. Summary of Experimental Hyperparameters and Configuration.

Hyperparameter	Value
Sampling Rate (f_s)	16,000 Hz
Chunk Duration	0.5 s
Hop Length	0.1 s
Batch Size	256
Number of Epochs	25
Optimizer	Adam
Initial Learning Rate	1×10^{-3}
Temperature (τ)	0.07
Random Seed	42
Hardware Acceleration	NVIDIA CUDA (Enabled)

4. Results and Discussion

This section reports the experimental findings obtained during the pre-training and fine-tuning phases. The final evaluation, conducted on completely unseen individuals through model inference, is presented in the last subsection.

All experiments were conducted on a cohort of 129 pediatric participants. The non-ASD group consisted of 75 children, including 38 SK-TD and 37 SK-ADHD participants. The ASD group comprised 54 children, including 46 SK-ASD participants from the Dutch dataset and 8 ASD participants from the ReCANVo dataset.

Since the recording duration varied across participants, a dedicated preprocessing algorithm was employed to standardize the data. Specifically, this algorithm ensured a balanced dataset by controlling the number of audio chunks extracted from each participant for use in the main experiments.

4.1. Pre-Training Results

The primary objective of the pre-training phase is to investigate the behavior of the supervised contrastive framework when trained exclusively on the Dutch dataset and to analyze the influence of key hyperparameters, particularly the learning rate. In a second step, ASD vocalizations from the ReCANVo dataset are incorporated during pre-training in order to evaluate the impact of non-verbal vocal signals on representation quality and downstream generalization.

Training Dynamics.

During pre-training, only the encoder parameters are optimized using the supervised contrastive loss (SupCon). As illustrated in Figure 1, the downstream classifier remains inactive at this stage. The Fast Fourier Transform (FFT)-based preprocessing and data augmentation modules are activated throughout training to enhance representation robustness and promote invariance to acoustic variability, thereby supporting improved generalization during inference.

Learning Rate Impact.

The choice of learning rate plays a critical role in convergence behavior and generalization capacity. As shown in Figure 2, a learning rate of 3×10^{-4} achieves a favorable balance between rapid convergence and training stability. Higher values (e.g., 5×10^{-4}) lead to unstable optimization and reduced inference performance, whereas lower values (e.g., 1×10^{-4}) produce rapid convergence but increase the risk of overfitting. Consequently, 3×10^{-4} was selected as the optimal learning rate for subsequent experiments.

 doi:10.20944/preprints202604.1071.v1

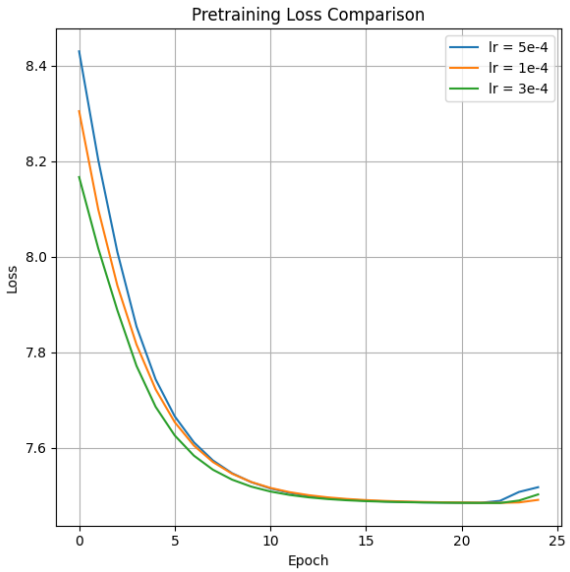

Figure 2. Comparison of loss curves by learning rate during pre train.

4.2. Fine-Tuning Results

The fine-tuning phase employs the same training participants used during pre-training. The encoder weights corresponding to the minimum supervised contrastive loss are retrieved and used to initialize the downstream classifier.

Extensive experiments were conducted to identify the optimal combination of hyperparameters capable of maximizing generalization across heterogeneous vocal profiles and recording conditions. The comparative results obtained with different ASD data configurations and learning rates are reported in Table 3.

For example, the first row of Table 3 corresponds to a configuration where SK-TD and SK-ADHD participants form the negative (non-ASD) class and SK-ASD (Dutch) participants form the positive (ASD) class. The learning rate is set to 3×10^{-4}. The first line reports fine-tuning performance, whereas the bold values correspond to the inference stage (82.14%). For example, increasing the learning rate to 5×10^{-4} reduces accuracy to 64.28%.

Table 3. Performance Comparison in Finetune and Inference across Learning Rate and Data in Training Configurations with mixed data (Verbal + Non-Verbal) in testing

Data Used in Training	LR	Accuracy	Precision	Recall	F1
Dutch	3E-04	99.62% **82.14**%	100.00% **76.47**%	100.00% **92.85**%	100.00% **83.87**% *
Dutch	1E-04	99.50% **75.00**%	97.50% **70.58**%	99.50% **85.71**%	99.50% **77.41**% *
Dutch + ReCANVo	1E-04	99.84% **82.14**%	100.00% **100.00**%	100.00% **64.24**%	100.00% **78.26**% *
Dutch	5E-04	97.30% **64.28**%	98.55% **59.09**%	98.55% **92.85**%	98.55% **72.22**%*

* Performance obtained during the inference stage.

Embedding Space Analysis.

Figure 3 presents t-SNE visualizations of the embedding space under different learning rate configurations. A clear separation between ASD and non-ASD clusters is observed, demonstrating that supervised contrastive pre-training effectively structures the latent space. In all experiments, the highest fine-tuning accuracy achieved is 99.89%.

 doi:10.20944/preprints202604.1071.v1

However, high validation accuracy alone does not guaranty robust real-world performance. For this reason, an additional inference evaluation is conducted on completely unseen individuals to confirm the sensitivity of generalization to optimization stability.

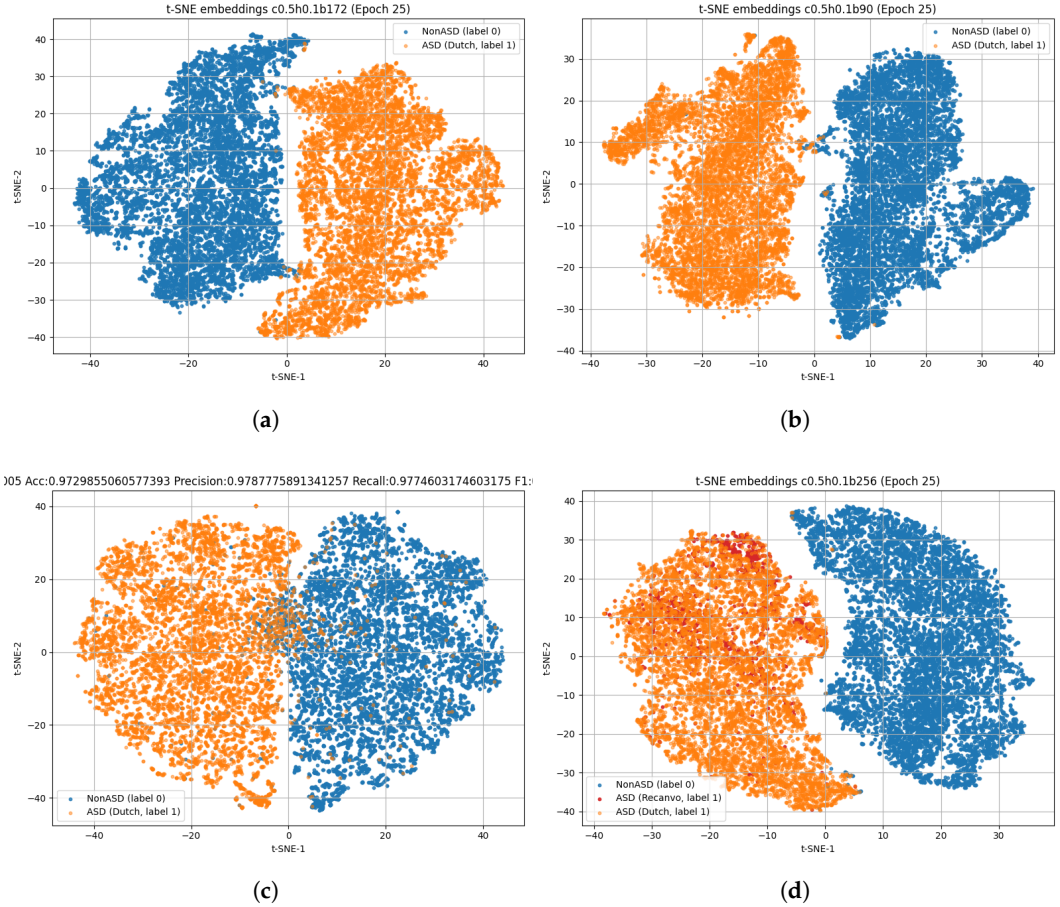

Figure 3. t-SNE visualization of embeddings for different learning rate "lr" and "ASD data" used during the finetune: (**a**) lr = 3e-4, Dataset = Dutch. (**b**) lr = 1e-4, Dataset = Dutch. (**c**) lr = 5e-4, Dataset = Dutch. (**d**) lr = 3e-4, Dataset = Dutch + ReCANVo.

4.3. Inference Performance

To evaluate the generalization capability of the proposed framework, the inference dataset was composed exclusively of participants who were not involved in either the pre-training or fine-tuning stages. This strict separation ensures that the reported results accurately reflect the model's ability to generalize to entirely unseen individuals. Two main experimental settings are considered, differing in the dataset composition used during the pre-training and fine-tuning stages.

4.3.1. Data Used in Training: Dutch + ReCANVo

The best inference performance is reported in Table 4. The proposed framework achieves a perfect classification accuracy of 100.00%, corresponding to 18 correctly classified children out of 18. This result is obtained with a learning rate of 3×10^{-4} when ASD participants in the testing phase produce non-verbal vocalizations as shown in Figure 4.

However, when the testing data include both articulated speech and non-verbal vocalizations, the accuracy decreases to 88.00%. This performance variation suggests that non-verbal vocalizations provide particularly discriminative acoustic cues for ASD detection.

 doi:10.20944/preprints202604.1071.v1

Table 4. Best Performance across Testing Dataset type (**Verbal vs Non-Verbal Vocalizations**) in Inference, with LR = 3E-04 and **Data used in Training : (Dutch + ReCANVo)**.

Data used in Testing	Accuracy	Precision	Recall	F1
ASD:ReCANVo + (TD,ADHD):Dutch	100.00%	100.00%	100.00%	100.00%
ASD:Dutch + (TD,ADHD):Dutch	82.14%	76.47%	92.85%	83.87%
ASD:(ReCANVo,Dutch) + (TD,ADHD):Dutch	88.00%	100.00%	72.72%	84.21%

```
Individuals evaluated: 18 (14 NonASD, 4 ASD)
MLP Testing: Acc=100.0000| Precision = 100.0000 | Recall = 100.0000 | F1 = 100.0000 | AUROC= 100.0000 | AUPRC=100.0000
...
adhd20: ✓ Pred=0 (True=0) Files=10Confidence=0.62 Agreement=100.0%
P03: ✓ Pred=1 (True=1) Files=115Confidence=0.73 Agreement=96.5%
P02: ✓ Pred=1 (True=1) Files=77Confidence=0.72 Agreement=87.0%
P11: ✓ Pred=1 (True=1) Files=106Confidence=0.74 Agreement=95.3%
td12: ✓ Pred=0 (True=0) Files=10Confidence=0.67 Agreement=100.0%
td29: ✓ Pred=0 (True=0) Files=10Confidence=0.69 Agreement=100.0%
adhd16: ✓ Pred=0 (True=0) Files=9Confidence=0.64 Agreement=77.8%
td17: ✓ Pred=0 (True=0) Files=10Confidence=0.64 Agreement=100.0%
td22: ✓ Pred=0 (True=0) Files=10Confidence=0.69 Agreement=100.0%
P06: ✓ Pred=1 (True=1) Files=74Confidence=0.76 Agreement=91.9%
td19: ✓ Pred=0 (True=0) Files=10Confidence=0.60 Agreement=100.0%
adhd19: ✓ Pred=0 (True=0) Files=10Confidence=0.51 Agreement=50.0%
adhd17: ✓ Pred=0 (True=0) Files=9Confidence=0.57 Agreement=77.8%
td33: ✓ Pred=0 (True=0) Files=10Confidence=0.60 Agreement=90.0%
adhd12: ✓ Pred=0 (True=0) Files=10Confidence=0.50 Agreement=60.0%
td16: ✓ Pred=0 (True=0) Files=10Confidence=0.57 Agreement=90.0%
adhd33: ✓ Pred=0 (True=0) Files=10Confidence=0.60 Agreement=80.0%
adhd37: ✓ Pred=0 (True=0) Files=10Confidence=0.61 Agreement=90.0%
```

Figure 4. Prediction Individual List with learning rate 3e-4, finetune with mixed train dataset and ReCANVo as testing dataset: 14 Non-ASD and 4 ASD

4.3.2. Data Used in Training: Dutch Only

In this experimental setting, the model is trained exclusively on the Dutch dataset, comprising SK-ASD, SK-ADHD, and SK-TD participants. Unlike the previous configuration, no samples from the ReCANVo cohort are included during training.

The results summarized in Table 5 indicate that the overall inference accuracy reaches 78.57% and 82.14% for verbal (Dutch) and mixed testing configurations, respectively. When evaluated on non-verbal vocalizations from the ReCANVo dataset, the model achieves a higher accuracy of 90.90%, suggesting that certain ASD-related acoustic patterns generalize across datasets even without explicit exposure during training.

Table 5. Performance across Testing Dataset type (**Verbal vs Non-Verbal Vocalizations**) in Inference, with LR = 3E-04 and **Data used in Training = Dutch only**

Data used in testing	Accuracy	Precision	Recall	F1
ASD:ReCANVo + (TD,ADHD):Dutch	100.00%	100.00%	100.00%	100.00%
ASD:Dutch + (TD,ADHD):Dutch	78.57%	78.57%	78.57%	78.57%
ASD:(ReCANVo,Dutch) + (TD,ADHD):Dutch	82.14%	76.47%	92.85%	83.87%

However, despite this relatively strong performance on non-verbal data, a noticeable performance gap remains compared to the configuration where both Dutch and ReCANVo datasets are used during training (Table 4).

In other words, the inclusion of both structured speech (Dutch dataset) and spontaneous non-verbal vocalizations (ReCANVo dataset) appears to enhance the robustness, stability, and discriminative power of the learned representations.

 doi:10.20944/preprints202604.1071.v1


```
Individuals evaluated: 21 (14 NonASD, 7 ASD)
MLP Testing: Acc=100.0000| Precision = 100.0000 | Recall = 100.0000 | F1 = 100.0000 | AUROC= 100.0000
P03: ✓ Pred=1 (True=1) Files=30Confidence=0.51 Agreement=43.3%
P01: ✓ Pred=1 (True=1) Files=29Confidence=0.66 Agreement=75.9%
adhd33: ✓ Pred=0 (True=0) Files=10Confidence=0.78 Agreement=100.0%
adhd20: ✓ Pred=0 (True=0) Files=10Confidence=0.61 Agreement=100.0%
P08: ✓ Pred=1 (True=1) Files=28Confidence=0.55 Agreement=57.1%
td12: ✓ Pred=0 (True=0) Files=10Confidence=0.70 Agreement=100.0%
adhd37: ✓ Pred=0 (True=0) Files=10Confidence=0.53 Agreement=50.0%
td33: ✓ Pred=0 (True=0) Files=9Confidence=0.54 Agreement=44.4%
td16: ✓ Pred=0 (True=0) Files=10Confidence=0.51 Agreement=40.0%
td19: ✓ Pred=0 (True=0) Files=10Confidence=0.65 Agreement=70.0%
P05: ✓ Pred=1 (True=1) Files=28Confidence=0.51 Agreement=57.1%
td29: ✓ Pred=0 (True=0) Files=10Confidence=0.70 Agreement=80.0%
td17: ✓ Pred=0 (True=0) Files=10Confidence=0.66 Agreement=70.0%
P06: ✓ Pred=1 (True=1) Files=18Confidence=0.60 Agreement=61.1%
adhd19: ✓ Pred=0 (True=0) Files=10Confidence=0.62 Agreement=80.0%
P11: ✓ Pred=1 (True=1) Files=27Confidence=0.57 Agreement=59.3%
adhd16: ✓ Pred=0 (True=0) Files=9Confidence=0.74 Agreement=100.0%
td22: ✓ Pred=0 (True=0) Files=10Confidence=0.81 Agreement=90.0%
adhd17: ✓ Pred=0 (True=0) Files=8Confidence=0.68 Agreement=75.0%
adhd12: ✓ Pred=0 (True=0) Files=10Confidence=0.58 Agreement=70.0%
P02: ✓ Pred=1 (True=1) Files=21Confidence=0.59 Agreement=57.1%

***Individual-level testing unseen data:   Number of individus: 21 (14 NonASD, 7 ASD)
Achieved at epoch 25
-----MLP Best Testing Perf : Acc=100.0000| Precision = 100.0000 | Recall = 100.0000 | F1 = 100.0000 |
✅ Fichier individus sauvegardé : /content/drive/MyDrive/TFC/Save_dataset/saved_models_supCon_predict

📊 Rapport d'analyse pour : predictions_new_data_indiv.csv
--------------------------------------------------
                          Total  Corrects  Accuracy (%)
group_type
DUTCH ( SK_TD + SK_ADHD )    14        14         100.0
RECANVO (ASD)                 7         7         100.0
```

Figure 5. Prediction Individual List with learning rate 3e-4, finetune with Dutch train dataset and ReCANVo as testing dataset: 14 Non-ASD and 7 ASD

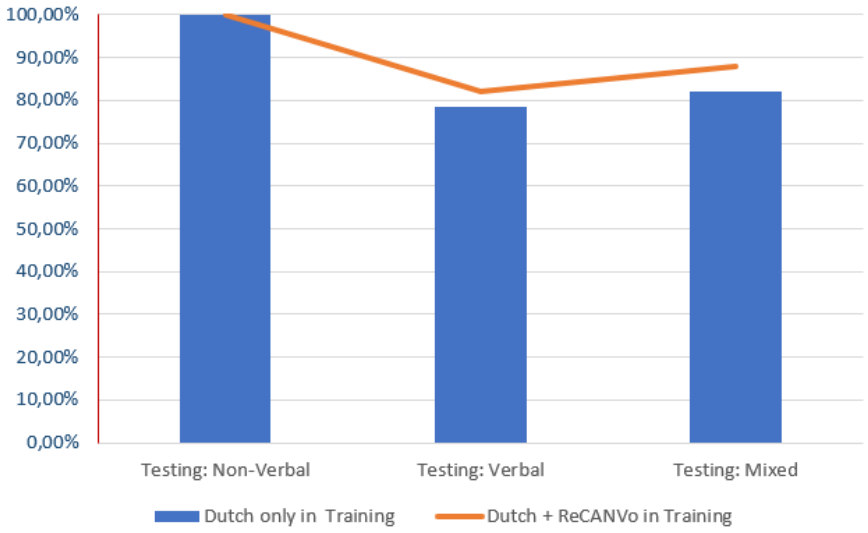

Figure 6. Accuracy in Inference across Dataset type (Verbal vs Non-Verbal Vocalizations)

4.4. Comparison with Vocal Biomarker Techniques

Previous research on ASD detection using vocal signals has largely focused on handcrafted acoustic and prosodic features, commonly referred to as vocal biomarkers [10,14–16]. These features include measurable characteristics such as pitch, jitter, shimmer, formant frequencies, and spectral

 doi:10.20944/preprints202604.1071.v1

energy, which identify anomalies in speech production that may indicate atypical vocal patterns associated with autism spectrum disorder (ASD), independently of linguistic content.

Support Vector Machines (SVM), Logistic Regression, Naive Bayes, Random Forest, K-Nearest Neighbours (KNN), and Decision Trees are just a few of the classical machine learning models that have been widely used to classify ASD using these vocal biomarkers.

To provide a comparative evaluation with traditional vocal biomarker approaches, experiments were conducted using the ReCANVo and Dutch datasets under similar training and testing conditions. The results obtained with these classical methods are summarized in Table 6.

When the Dutch dataset is used for training, the performance remains relatively limited across all testing configurations. The accuracy ranges between 52.47% and 67.71%, with particularly low recall values for ASD detection in non-verbal speech (6.45%). These results suggest that handcrafted acoustic features extracted from verbal speech alone may not capture sufficiently discriminative patterns for reliable ASD classification.

When both datasets are combined during training, the performance improves compared to single-dataset training, reaching an accuracy of 73.15% and an F1-score of 65.33% under mixed testing conditions. Nevertheless, the results remain inferior to those achieved by the proposed supervised contrastive learning framework.

Table 6. Performance across Training and Testing Dataset type : **Verbal (Dutch) vs Non-Verbal Vocalizations (ReCANV0)** in Inference for unseen data.

DUTCH in Training	**Accuracy**	**Precision**	**Recall**	**F1**
ASD:ReCANVo + (TD,ADHD):Dutch	67.71%	50.00%	6.45%	11.43%
ASD:Dutch + (TD,ADHD):Dutch	52.47%	55.00%	33.08%	41.31%
ASD:(ReCANVo,Dutch) + (TD,ADHD):Dutch	55.64%	57.65%	38.58%	46.23%
MIXED in Training	**Accuracy**	**Precision**	**Recall**	**F1**
ASD:ReCANVo + (TD,ADHD):Dutch	92.71%	86.36%	91.94%	89.06%
ASD:Dutch + (TD,ADHD):Dutch	53.23%	55.43%	38.35%	45.33%
ASD:(ReCANVo,Dutch) + (TD,ADHD):Dutch	73.15%	90.28%	51.18%	65.33%

The proposed method, on the other hand, works much better in most testing situations. For example, it gets up to 100% accuracy when tested on non-verbal vocalizations and 88.00% accuracy when tested in mixed conditions. These results show the advantage of representation learning approaches that jointly capture temporal and spectral patterns directly from raw vocal signals rather than relying solely on predefined acoustic descriptors.

Overall, the comparison demonstrates that supervised contrastive learning provides a more powerful and robust framework for modeling ASD-related vocal biomarkers, particularly in heterogeneous and cross-context evaluation settings involving both verbal (dutch dataset) and non-verbal (ReCANVo dataset) vocalizations.

4.5. Discussion

- Regularization through Data Heterogeneity and Learning Rate Calibration

The t-SNE visualization of embeddings in Figure 3 shows an important trade-off in model optimization. A lower learning rate (*lr*) initially creates a clearer separation between the two classes, but it also increases the risk of overfitting because it lowers the variance within each class, which makes generalization worse in the end. To fix this, Figure 3d shows that keeping the learning rate at its average value strikes a better balance between class separability and architectural stability. Adding ReCANVo's non-verbal vocalizations to the articulated speech dataset makes this balance even stronger. The inclusion of diverse ASD profiles guarantees a more robust representation that encompasses the entire spectrum of the condition while maintaining categorical distinctiveness.

 doi:10.20944/preprints202604.1071.v1

- Analysis of Model Generalizability across Heterogeneous Cohorts

The results in Table 3 prove that the model works well even in new situations. Although it was only trained on a small group of Dutch speakers telling stories, it still performs highly when tested on a much broader dataset. This success is significant because the test data includes different recording methods, environments, and speech types. Most importantly, the model remained accurate when tested on 14 subjects with Autism Spectrum Disorder (ASD), half of whom were from a completely different dataset (ReCANVo). Since these new recordings came from families in the United States, the results confirm that the model can handle changes in language, geography, and recording conditions without losing its effectiveness.

- Optimization of Class Separability and Intra-class Variance

The integration of heterogeneous ASD vocalizations—specifically non-verbal samples from the ReCANVo dataset—substantially enhances the modeling of intra-class variability while preserving robust inter-class separability. As evidenced by the t-SNE projections in Figure 3, these different sounds belong to the same cluster rather than form separate groups. This proves that the "voice biomarkers" of ASD stay the same whether a child is speaking or making non-verbal sounds, reinforcing the validity of the learned feature space.

- Better prediction with non-verbal vocalizations

Both the proposed framework and the traditional vocal biomarker approaches perform best when analyzing non-verbal vocalizations. The results in Table 4 highlight a significant correlation between non-verbal vocalizations and increased model precision. This is because non-verbal sounds are usually continuous and do not have the long pauses found in regular speech. Because there is more "active" sound and less silence or background noise, the data are cleaner. This high density of information makes it much easier for the model to detect vocal patterns linked to ASD.

5. Conclusions

This study presented a supervised contrastive learning framework for the detection of ASD in infant vocalizations, based on the Time-Frequency Consistency Architecture (TF-C) and extending it beyond its original self-supervised formulation.

Heterogeneous vocal traits associated with ASD were successfully captured by integrating complementary temporal and spectral representations within a dual-branch encoder. The learned embeddings' discriminative ability was further reinforced by the supervised contrastive objective, which enhanced their robustness to inter-speaker variability.

The final evaluation on completely unseen individuals demonstrated strong generalization performance, demonstrating that the model learns diagnostically significant patterns rather than just memorizing cohort-specific characteristics.

Additionally, the findings imply that both spontaneous nonverbal vocalizations and structured speech exhibit vocal traits linked to autism. Thus, incorporating a variety of vocal profiles enhances the robustness of the representation and encourages cross-contextual generalization. In particular, the use of nonverbal vocalizations during the prediction phase significantly improves the model's ability to correctly identify autism spectrum disorders (ASDs).

Nevertheless, several limitations remain. The dataset size, although multi-cohort, is still limited relative to large-scale deep learning benchmarks. Future work should explore larger and more diverse populations, longitudinal data, and cross-linguistic robustness.

In short, the proposed framework advances the modeling of vocal biomarkers for ASD and represents a promising step towards the development of reliable, non-invasive and scalable early screening systems suitable for real-world clinical applications.

 doi:10.20944/preprints202604.1071.v1

Code Availability

To promote transparency and facilitate reproducibility, the full implementation of the proposed framework, including data preprocessing pipelines, model architectures, training scripts, and evaluation procedures is publicly available on GitHub ASD_SupCon_TFC_Repository. The repository provides detailed documentation, dependency specifications, and instructions for replicating the experimental setup reported in this study. In addition, configuration files are provided to reproduce the hyperparameter settings summarized in Table 2, ensuring that independent researchers can replicate and extend the presented results.

Institutional Review Board Statement: Ethical review and approval were waived for this study because the analysis was conducted exclusively on publicly available datasets. No new data were collected, and all datasets were originally obtained with ethical approval and informed consent by the respective investigators.

Informed Consent Statement: The datasets used in this study consist of secondary analyses of publicly available datasets. Informed consent was obtained from participants by the original investigators in accordance with the ethical approvals and study protocols of the respective studies.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, X.; Zhao, Z.; Tsiligkaridis, T.; Zitnik, M. Self-Supervised Contrastive Pre-Training For Time Series via Time-Frequency Consistency. *Proc. Neural Inf. Process. Syst. (NeurIPS)* **2022**.
2. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised Contrastive Learning. *arXiv* **2020**, arXiv:2004.11362. https://arxiv.org/abs/2004.11362.
3. Hu, C.; Thrasher, J.; Li, W.; Ruan, M.; Yu, X.; Paul, L.; Wang, S.; Li, X. Exploring speech pattern disorders in autism using machine learning. *arXiv* **2024**, https://doi.org/10.48550/arXiv.2405.05126.
4. Sai, K.; Krishna, R.; Radha, K.; Rao, D.; Muneera, A. Automated ASD detection in children from raw speech using customized STFT–CNN model. *Int. J. Speech Technol.* **2024**, *27*, 701–716. https://doi.org/10.1007/s10772-024-10131-7.
5. Vacca, J.; Brondino, N.; Dell'Acqua, F.; Vizziello, A.; Savazzi, P. Automatic Voice Classification of Autistic Subjects. *arXiv* **2024**, https://doi.org/10.48550/ARXIV.2406.13470.
6. Deng, S.; Kosloski, E.; Patel, S.; Barnett, Z.; Nan, Y.; Kaplan, A.;Aarukapalli, S.; Doan, W.T.; Wang, M.; Singh, H.; et al. Hear Me, See Me, Understand Me: Audio–Visual Autism Behavior Recognition. *arXiv* **2024**, https://doi.org/10.48550/ARXIV.2406.02554.
7. Chi, N.; Washington, P.; Kline, A.; Husic, A.; Hou, C.; He, C.; Dunlap, K.; Wall, D. Classifying autism from crowdsourced semi-structured speech recordings: A machine learning approach. *arXiv* **2022**, https://doi.org/10.48550/arXiv.2201.00927.
8. Murugaiyan, S.; Uyyala, S. Aspect-based sentiment analysis of customer speech data using deep CNN and BiLSTM. *Cogn. Comput.* **2023**, *15*, 914–931. https://doi.org/10.1007/s12559-023-10127-6.
9. Rakotomanana, H.; Rouhafzay, G. A Scoping Review of AI-Based Approaches for Detecting Autism Traits Using Voice and Behavioral Data. *Bioengineering* **2025**, *12*, 1136. https://doi.org/10.3390/bioengineering12111136. :contentReference[oaicite:0]index=0
10. Briend, F.; David, C.; Silleresi, S.; Malvy, J.; Ferré, S.; Latinus, M. Voice acoustics allow classifying autism spectrum disorder with high accuracy. *Transl. Psychiatry* **2023**, *13*, 250. https://doi.org/10.1038/s41398-023-02554-8.
11. Lee, J.; Lee, G.; Bong, G.; Yoo, H.; Kim, H. Deep-learning–based detection of infants with autism spectrum disorder using autoencoder feature representation. *Sensors* **2020**, 20, 6762. https://doi.org/10.3390/s20236762.
12. Li, M.; Tang, D.; Zeng, J.; Zhou, T.; Zhu, H.; Chen, B.; Zou, X. An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder. *Comput. Speech Lang.* **2019**, *56*, 80–94. https://doi.org/10.1016/j.csl.2018.11.002.
13. Rosales-Pérez, A.; Reyes-García, C.; Gonzalez, J.; Reyes-Galaviz, O.; Escalante, H.; Orlandi, S. Classifying infant cry patterns by the genetic selection of a fuzzy model. *Biomed. Signal Process. Control* **2015**, *17*, 38–46. https://doi.org/10.1016/j.bspc.2014.10.002.
14. Asgari, M.; Chen, L.; Fombonne, E. Quantifying voice characteristics for detecting autism. *Comput. Speech Lang.* **2021**, *12*, 665096. https://doi.org/10.3389/fpsyg.2021.665096.

 doi:10.20944/preprints202604.1071.v1

15. Lehnert-LeHouillier, H.; Terrazas, S.; Sandoval, S. Prosodic Entrainment in Conversations of Verbal Children and Teens on the Autism Spectrum. *Front. Psychol.* **2020**, *11*, 582221. https://doi.org/10.3389/fpsyg.2020.582221.
16. Mohanta, A.; Mittal, V. Analysis and classification of speech sounds of children with autism spectrum disorder using acoustic features. *Comput. Speech Lang.* **2022**, 72, 101287. https://doi.org/10.1016/j.csl.2021.101287.
17. Nature Digital Medicine. Reproducibility, external validation, and generalization in clinical machine learning. *Nat. Digit. Med.* **2023**. https://doi.org/10.1038/s41746-023-00845-4.
18. Molecular Psychiatry. Cross-cohort generalization challenges in psychiatric biomarkers. *Mol. Psychiatry* **2023**. https://doi.org/10.1038/s41398-023-02554-8.
19. Laguna, A.; Pusil, S.; Paltrinieri, A. L.; Orlandi, S. Automatic Cry Analysis: Deep Learning for Screening of Autism Spectrum Disorder in Early Childhood. *J. Autism Dev. Disord.* **2025**. https://doi.org/10.1007/s10803-025-06811-1.
20. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *Proc. Mach. Learn. Res.* **2020**, *119*, 1597–1607. https://arxiv.org/abs/2002.05709.
21. Leng, Y.; Anwar, S. M.; Rekik, I.; He, S.; Lee, E.-J. Self-Supervised Graph Transformer with Contrastive Learning for Brain Connectivity Analysis Towards Improving Autism Detection. In *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*; IEEE: Houston, TX, USA, 2025; pp. 1–5. https://doi.org/10.1109/ISBI60581.2025.10981292. :contentReference[oaicite:0]index=0
22. MacWhinney, B. *The CHILDES Project: Tools for Analyzing Talk*, 3rd ed.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2000. Available online: https://talkbank.org.
23. Narain, J.; Johnson, K. T. ReCANVo: A Dataset of Real-World Communicative and Affective Nonverbal Vocalizations. *Zenodo Dataset* **2021**. https://doi.org/10.5281/zenodo.5786860.