Activity-based Early Autism Diagnosis Using A Multi-Dataset Supervised Contrastive Learning Approach

Asha Rani
IIT Jodhpur, India
rani.1@iitj.ac.in

Yashaswi Verma
IIT Jodhpur, India
yashaswi@iitj.ac.in

Abstract

Autism Spectrum Disorder (ASD) is a neurological disorder. Its primary symptoms include difficulty in verbal/non-verbal communication and rigid/repetitive behavior. Traditional methods of autism diagnosis require multiple visits to a human specialist. However, this process is generally time-consuming and may result in a delayed (early) intervention. In this paper, we present a data-driven approach to automate autism diagnosis using video clips of subjects performing simple activities recorded in a weakly constrained environment. This task is particularly challenging since the available training data is small, videos from the two categories ("ASD" and "Control") are generally perceptually indistinguishable, and there is no clear understanding of what features would be beneficial in this task. To address these, we present a novel multi-dataset supervised contrastive learning technique to learn discriminative features simultaneously from multiple video datasets with significantly diverse distributions. Extensive empirical analyses demonstrate the promise of our approach compared to competing techniques on this challenging task.

1. Introduction

Autism Spectrum Disorder (ASD), commonly known as *autism*, is a neurological disorder that is primarily characterized by difficulty in social interaction. ASD affects a person's ability to communicate properly with others and is reported to occur across all racial, ethnic and socio-economic groups. The symptoms of ASD usually start appearing in early childhood between the age of 1-3 years. These include a lack of proper eye contact, poor imitation skills and rigid/repetitive behavior. While the number of identified ASD cases has increased in the last decade, the current mean age of ASD diagnosis is reported as 3-12 years [18]. In the recent times, several measures have been adopted to reduce the mean age of diagnosis. This is particularly crucial since an early diagnosis of ASD leads to an early intervention and subsequent therapies that significantly benefit the growth of the affected child. Traditionally, ASD diagnosis has been performed physically by an expert medical practitioner. However, this requires multiple visits and is thus time-consuming as well as sometimes error-prone.

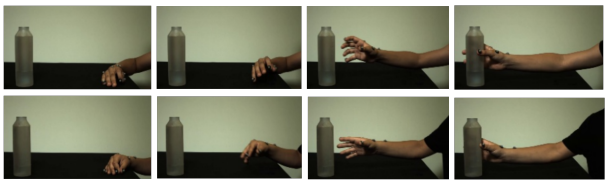


Figure 1. Sample frames from video clips of subjects from the ASD (top) and Control (bottom) categories in the Hand Gesture dataset [32]. Note that these video clips exhibit low inter-class variability and are difficult to classify using individual visual cues.

The limitations of physical ASD diagnosis can be addressed to a large extent by adopting an automated approach. In this paper, we present a data-driven approach for automated autism diagnosis using recorded activity (action) video clips of subjects. Each clip captures a subject performing some specific activity in a loosely controlled environment, and is much easier and less expensive to acquire compared to data in other modalities such as EEG, MRI or eye-tracking. However, it is difficult to obtain a good prediction accuracy with a machine learning model trained on such data because (a) the training data contains a small number of samples, (b) the data is inherently complex to comprehend, and (c) the video samples (clips) exhibit low inter-class variability. E.g., Figure 1 shows sample video frames corresponding to "ASD" and "Control" category in one such dataset (the Hand Gesture dataset [32]), where we can observe that both the samples look quite similar. Due to this, it is difficult to correctly classify samples based on individual visual cues. In our experiments also, we will show that a conventional deep neural network can not be directly adopted for this task for the same reasons.

To address these challenges, we propose to use contrastive feature learning [3, 12, 20] to learn distinctions be-

tween video clips of the two categories (ASD and Control) in a relative manner. Specifically, we present a novel multi-dataset supervised contrastive learning (MSupCL) technique that learns discriminative features by simultaneously using multiple (activity-based) video datasets from diverse distributions. Thorough empirical experiments on the two relevant and publicly available datasets (Hand Gesture [32] and Autism [19]) show that our proposed approach significantly outperforms competing contrastive learning techniques [3, 12, 20] on the challenging Hand Gesture dataset [32], and is comparable to the best approach on the (easier) Autism dataset [19]. For reproducibility, our code and pre-trained models are available at https://github.com/asharani97/MDSupCL.

2. Related Work

Several machine learning based techniques have been proposed in the recent years that aim at performing autism diagnosis in an automated manner. In general, most of these efforts have primarily investigated data acquired in the form of different modalities and posed it as a single-step classifier learning task [9, 11, 13, 15, 17, 21–26, 31, 32]. One of the early studies on autism diagnosis revealed that autistic individuals have atypical sight [6]. Later, the authors of [11, 30] further worked on this finding, with [11] focusing on visual bias towards different objects, their contrast and colour, while [30] focusing on using these aspects to predict ASD and control subjects using deep features [16]. Another approach to distinguish between the two categories was introduced in [22], where the first-person viewpoint of a scene by an individual is compared for analysis. A few studies that have investigated other modalities of data include eye-gaze data to distinguish on the basis of visual attention [26], EEG signal [1, 2, 25], and MRI data [9, 13]. Since these different modalities of data are difficult and expensive to acquire, a recent work [32] introduced the Hand Gesture dataset which contains short video clips that capture predefined activities performed by autistic/control subjects in a weakly controlled environment.

It is worth noting that while most of the existing machine learning based autism diagnosis approaches have analyzed the pros and cons of using features from different modalities as discussed above, eventually they pose this as a binary classification task and use a sample-level binary classifier such as SVM or a deep neural network. Unlike existing approaches, we aim at learning discriminative features in a relative manner by simultaneously using two diverse datasets in a contrastive learning based set-up. The broad idea of learning from multiple datasets simultaneously has recently gained popularity [5, 28], where a model is trained by integrating multiple datasets created for a specific task such as object detection [5], image segmentation [28], etc. In general, multi-dataset training aims at learning improved feature representations given the costly nature of annotation for diverse downstream applications. This also helps in minimizing the issue of domain shift across different datasets (also see our discussion in Section 4.4.1). As per our knowledge, ours is the first NT-Xent (normalized temperature-scaled cross entropy) loss based contrastive learning approach for a multi-dataset set-up, as well as the the first such attempt for the activity video-based autism diagnosis task.

3. Proposed Approach

Our approach consists of two steps: learning a deep feature encoder network using the proposed multi-dataset supervised contrastive learning approach, followed by training of a dataset-specific classifier for prediction.

3.1. Background and Motivation

Contrastive learning requires positive/similar and negative/dissimilar pairs of samples, and aims at pulling similar samples closer than dissimilar samples in the learned feature space. Our approach is motivated by the success of the recent contrastive learning techniques such as [3,4,8,12,20]. SSCL [3] is one of the earlier uni-modal approaches that is based on self-supervised contrastive learning. In this, a positive pair is generated using a given sample (anchor) and its transformed version (obtained using some non-learned transformation; e.g., horizontal flip), and a negative pair is generated using the same given sample and any other sample (either from the dataset or the transformed version of another sample). Extending SSCL, SupCL [12] uses external supervision in the form of category labels for creating positive pairs, and for each positive pair. In the recent multi-modal self-supervised approach MSSCL [20], the input data contains *paired* samples from two different modalities (video and audio) that are extracted from an audio-video recording, thus naturally giving cross-modal positive pairs. However, it may not always be feasible to acquire paired data. Unlike MSSCL, our approach does not require an explicit pairing of samples. Rather, we take motivation from [20] and make use of the categorical information to create multiple cross-dataset positive pairs corresponding to each anchor. Specifically, we assume that two datasets may be collected independently and follow different distributions, however both should contain samples from the same set of categories. This allows us to create cross-dataset positive pairs of samples based on categorical supervision without requiring an explicit pairing as in [20]. Below, we describe our first step, *i.e.*, generation of pairs for our contrastive learning based approach.

3.2. Pair Generation

In our task, we have two datasets $\mathcal{D}_1 = \{(v_i, y_i)\}$ and $D_2 = \{(v_j, y_j)\}$, each containing activity videos of ASD

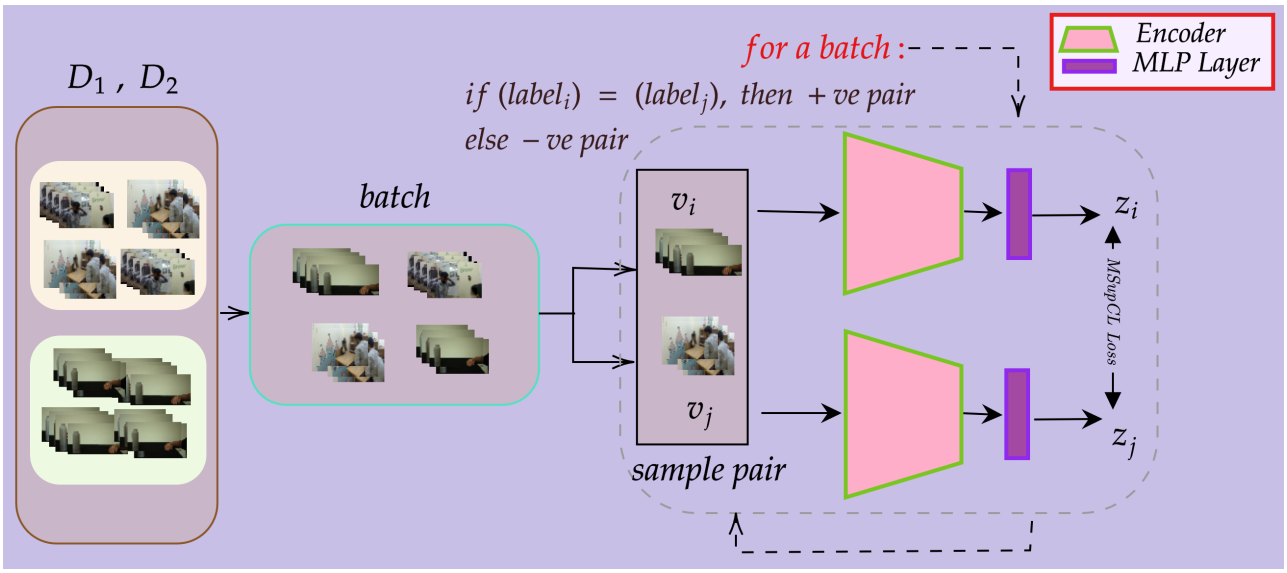


Figure 2. Block diagram of our network architecture and training process. We consider two different datasets D_1 and D_2 for the same task (*i.e.* ASD diagnosis in our case) from different distributions. In a batch, there are samples from both the datasets. For cross-dataset pair generation (Section 3.2), one sample is picked from each dataset (a transformation is applied on only one of these samples). Both the samples are fed to the encoder network $e()$ to generate the initial feature representations, which are then passed through a multi-layer perceptron head $h()$ giving us the embeddings z_i and z_j respectively (Section 3.3). These embeddings are then used to compute the proposed MSupCL loss (Section 3.4) and train the whole network in an end-to-end manner using a gradient descent method.

and control subjects collected separately; *i.e.*, by independent groups of researchers/clinicians under different conditions. During training, a batch consists of samples from both the datasets. In a batch, for a given sample (anchor) from one dataset, we create a positive cross-dataset pair by pairing it with a sample from the second dataset that belongs to the same category. To create a negative cross-dataset pair for the same anchor point, we pair it with a sample from the second dataset that belongs to another category. Specifically, consider an anchor point and its corresponding category (v_a, y_a) from one dataset. To create a positive pair, we pick a sample (v_p, y_p) from the second dataset such that $y_a = y_p$. To create a negative pair, we pick another sample (v_n, y_n) from the second dataset such that $y_a \neq y_n$. We perform this to obtain all possible pairs in the given batch, thus resulting in *multiple* cross-dataset positive and negative pairs corresponding to each anchor point. These pairs are then passed to a deep feature extraction (encoder) network, as described next.

3.3. Feature Extraction Network

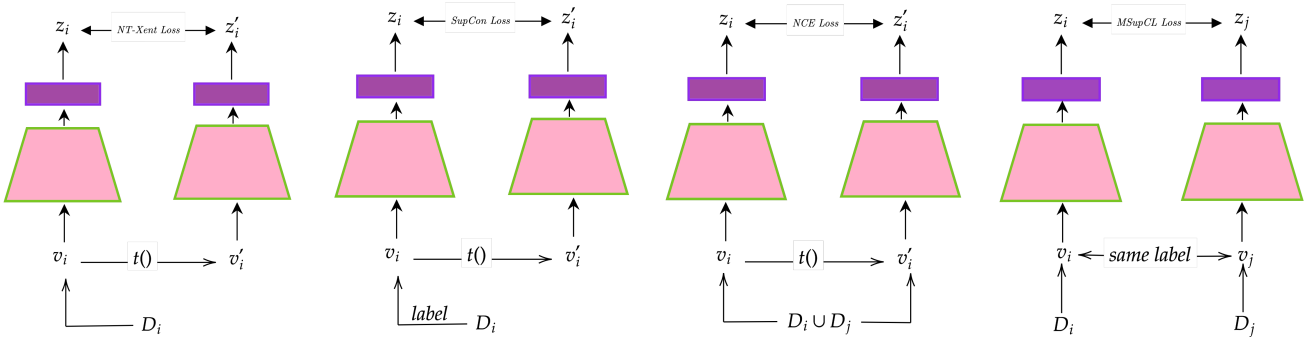
Given a (positive/negative) pair of videos v_i and v_j created from the datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, we pass them through feature encoder networks $e_i()$ and $e_j()$ to obtain their initial feature representations x_i and x_j respectively. These feature representations are then passed through fully-connected layers $h_i()$ and $h_j()$ and mapped to embeddings z_i and z_j respectively (Figure 2). To learn the parameters of this network, we propose a novel multi-dataset supervised contrastive loss function, which we describe next.

3.4. Multi-dataset Supervised Contrastive Loss

For a given anchor point v_a, we create a positive pair (v_a, v_p) and a negative pair (v_a, v_n) as discussed in the pair generation step. Let P_a and N_a denote the sets of all such positive and negative samples with respect to v_a, which are used to create positive and negative pairs respectively. These pairs are passed through the feature extraction network as described above to obtain their feature embeddings. For the anchor v_a, its positive sample v_p and its negative sample v_n, the feature embeddings are denoted by z_a, z_p and z_n respectively. Also, let $K_a = P_a \cup N_a$ be the set of all such positive and negative samples corresponding to v_a. Using these, we calculate the multi-dataset supervised contrastive loss for v_a as below:

$$L^a_{MSupCL} = -\frac{1}{|P_a|} \sum_{v_p \in P_a} \log \frac{\exp(z_a \cdot z_p/\tau)}{\sum_{v_k \in K_a} \exp(z_a \cdot z_k/\tau)} \quad (1)$$

Here, τ denotes the temperature parameter, and the $\cdot$ symbol denotes the inner (dot) product. This loss is averaged over all the samples in a batch by considering each sample as an anchor point at a time to obtain the total loss. This is then used to train the whole network (Figure 2) in an end-to-end manner using a gradient descent approach (Section 4.3).



(a) SSCL [3] (b) SupCL [12] (c) MSSCL [20] (d) MSupCL (Ours)

Figure 3. Schematic comparison of the proposed MSupCL approach with three contrastive learning based methods in terms of network architectures and training procedures. In all the methods, an input sample v_i is first passed through an encoder network $e()$ and then through a non-linear multi-layer perceptron head $h()$ to get the embedding z_i. $t()$ denotes a transformation function that generates a sample v'_i from v_i using a simple transformation, which is then used to create a positive pair during training. Here: (a) SSCL is a self-supervised uni-modal technique that uses an NT-Xent loss. (b) SupCL is a label-supervision based uni-modal technique that uses a supervised contrastive loss. (c) MSSCL is an extension of SSCL for multi-modal data $D_i \cup D_j$ and requires explicit pairing between cross-modal samples during training. (d) MSupCL (ours) uses a multiple unpaired datasets simultaneously, however the pairing of data points is done based on label information instead of explicit pairing as in MSSCL, and is trained using the loss function as described in Section 3.4.

It is worth noting that the loss function in Eq. 1 generalizes the supervised constrastive loss proposed in [12] to a multi-dataset set-up, and extends the multi-modal self-supervised contrastive loss proposed in [20] to benefit from categorical supervision. We would also like to emphasize that while we demonstrate our approach on a multi-dataset set-up, our approach may be easily adapted for data containing different modalities by plugging appropriate modality-specific encoder networks.

3.5. Classification

The next step is to train a dataset-specific classification model. Following earlier related papers [3, 12, 20], we take only the encoder network from the previously trained network (as shown in Figure 2) and add a new fully-connected classification layer with softmax activation. To train this layer of the updated network, we freeze the parameters of the encoder network and learn those in the classification layer using the standard cross-entropy loss. This network is then used for doing prediction on the unseen/test data. It should be noted that the classification network needs to be trained for each dataset individually [20].

4. Experiments

In this section, we discuss our experimental setup and present the empirical results.

4.1. Datasets

We use two activity video datasets in our experiments: Hand Gesture dataset [32] and Autism dataset [19]. As per our knowledge, these are the only relevant and publicly accessible activity video datasets for this task. Both the datasets contain short video clips of subjects where they are asked to perform some predefined activities in a loosely controlled environment.

Hand Gesture dataset [32]: This dataset contains video recordings of four activities (placing, pouring, pass to place, and pass to pour), each performed multiple times by 39 subjects (19 ASD and 20 Control). As discussed in Section 1 and illustrated in Figure 1, this is a challenging dataset with high intra-class and low inter-class variability.

Autism dataset [19]: This dataset contains video recordings corresponding to eight activities (move the table, touch ear, lock hands, touch head, touch nose, rolly polly, tapping, and arms up). Each activity is recorded from two camera orientations: tutor facing and child-facing. To create this dataset, the authors of [19] first recorded videos of ASD subjects, and then obtained samples of the "control" class by picking a similar number of videos corresponding to the most identical actions from the HMDB51 dataset [14]. Because of this, the distributions of the two classes are well-separated (Figure 4) in this dataset.

4.2. Compared Methods

To examine the effectiveness of the proposed MSupCL approach, we compare it with competing contrastive learning techniques including the uni-modal self-supervised approach SSCL [3], uni-modal supervised approach SupCL [12], and multi-modal self-supervised approach MSSCL [20], as illustrated in Figure 3. As a baseline, we also compare with a standard deep classifier (BinClass)

Dataset		% Accuracy				
Activity	#Samples	BinClass	SSCL	SupCL	MSSCL	MSupCL
Pass to Place	139	51.09	70.07	<u>70.80</u>	64.23	**89.78**
Pass to Pour	140	54.23	<u>73.24</u>	69.72	57.04	**81.69**
Placing	140	51.43	<u>73.57</u>	67.14	55.71	**80.71**
Pouring	142	52.11	68.31	<u>69.01</u>	52.11	**85.92**
Average	561	52.23	<u>71.30</u>	69.16	57.22	**84.49**

Table 1. Activity-wise and average classification accuracy on the Hand Gesture dataset. In each row, the best result is highlighted in ***bold***, while the second best is *<u>underlined</u>*.

Dataset		% Accuracy				
Activity	#Samples	BinClass	SSCL	SupCL	MSSCL	MSupCL
Touch Nose - Eat	101	**100**	**100**	**100**	<u>98.02</u>	**100**
Touch head - Shoot Ball	106	<u>98.11</u>	**100**	97.17	91.51	<u>98.11</u>
Touch ear - Situp	74	97.30	**100**	95.95	**100**	<u>98.65</u>
Tapping - Chew	144	97.91	**99.31**	96.53	<u>98.61</u>	**99.31**
Rolly Polly - Flic flac	59	**100**	**100**	**100**	**100**	**100**
Move the table - Push	101	<u>99.01</u>	**100**	99.00	96.04	**100**
Lock Hands - Shake Hands	85	<u>98.82</u>	**100**	**100**	97.65	**100**
Arms Up - Fall Floor	85	<u>98.82</u>	**100**	97.65	95.29	96.47
Average	755	98.67	**99.87**	98.14	96.95	<u>99.07</u>

Table 2. Activity-wise and average classification accuracy on the Autism dataset. In each row, the best results is highlighted in ***bold***, while the second best is *<u>underlined</u>*.

trained using the binary cross-entropy loss.

4.3. Implementation Details

We first pre-process each activity video by extracting key frames while keeping the sequential information intact. We uniformly pick 16 and 10 frames from each video sample of the Hand Gesture and Autism dataset respectively, since these datasets contain 21-30 frames and 12-20 frames per video clip respectively. We use the ResNet-based R(2+1)D-18 [27] network as the encoder network $e()$ in all the compared methods, which is the most widely used feature encoder for activity/action datasets. It maps an input video (a sequence of sampled frames) into a 512-dimensional feature vector. We pass this feature vector through a fully-connected layer $h()$ with ReLU activation, giving a 256-dimensional feature vector. This is then normalized using the L_2-norm and is used to compute the loss function. In our case, since both the modalities are videos (though obtained from different sources and thus following different distributions), we use a duplicated encoder network. For classification, we keep only the encoder network and add a fully-connected classification layer with softmax activation. For fair comparisons, we use the same approach for classifier training in the three contrastive learning techniques (SSCL, SupCL and MSSCL) as used in the proposed MSupCL. In all the experiments, we keep the train-test ratio as 70:30.

Compute Environment: The experiments were conducted on a server with shared access, having 8 GTX 1080 Ti 12GB GPUs, Intel Xeon E5-2650 2.20GHz processors, and 256GB RAM. For training MSSCL and MSupCL, 4 GPUs were used, while for other methods, 2 GPUs were used.

4.4. Results and Discussion

We first compare the classification accuracy of all the methods on the Hand Gesture dataset in Table 1. We observe that on this challenging dataset, the baseline binary classifier is insufficient to learn discriminative features, thus leading to a poor (near-chance) accuracy. Compared to this, both SupCL and SSCL achieve significantly higher accuracy. Interestingly, we notice that SSCL performs slightly better than SupCL even without using categorical information. We believe this is because of the inherent visual complexity of these task (perceptually indistinguishable variations among samples from the two categories), coupled with the technical distinctions between the two approaches with respect to their loss functions and the way they create positive/negative pairs (for further details, we request the reader the refer to the respective papers). This also indicates that categorical information may not significantly affect the accuracy of uni-modal approaches in tasks like ours where the number of samples is small and datasets depict low inter-class variability. However, the performance drops drastically in case of MSSCL, which indicates that self-supervised learning may not benefit from uncoupled/unpaired multi-source data as in our case (note that the original paper [20] used *paired* multi-modal data). For the same reason, we acknowledge that MSSCL is not directly comparable to our approach. Finally, we can notice that the proposed MSupCL approach achieves the maximum accuracy, which is around 13% (absolute) more than the second-best method (SSCL). This indicates that multiple unpaired but labelled datasets can be effectively used for multi-dataset contrastive learning by creating (positive/negative) pairs based on categorical information, and thus validates the utility of our approach over the compared techniques.

Next, we compare the accuracy of all the methods on the Autism dataset in Table 2. As emphasized in Section 4.1, the distributions of the two categories in this dataset are well separated. This becomes evident in the empirical results where all the compared methods achieve a near-perfect accuracy with statistically insignificant differences in their predictions (based on the t-test, with t-value 0.00).

4.4.1 Relation with Knowledge Distillation

Knowledge distillation [7, 10] is a well-studied idea where the broad objective is to transfer the knowledge of one or more teacher models to a student model (*i.e.*, *model-level* knowledge distillation) for a particular dataset, where the teacher model is generally assumed to be of more/equal learning capacity compared to the student model. In our

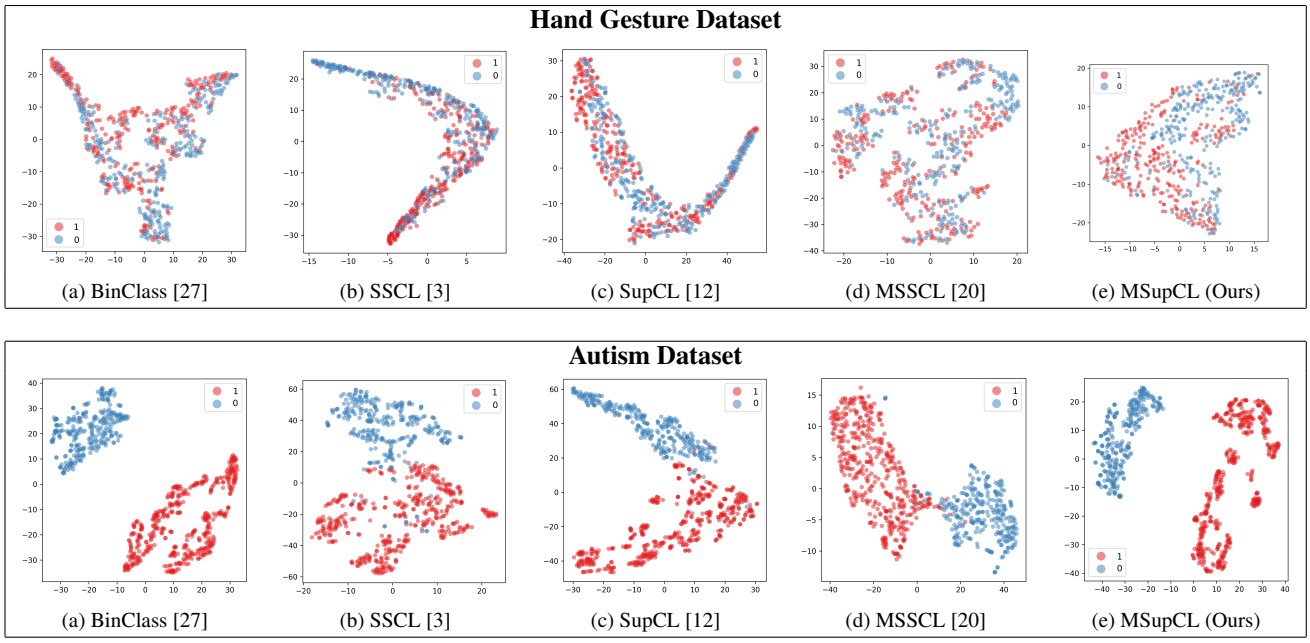


Figure 4. t-SNE visualization of features learned using different methods on the Hand Gesture dataset (top row) and the Autism dataset (bottom row). The points in blue denote Control samples while those in red denote ASD samples. (Best viewed in colour.)

task, we have two datasets that belong to the same (video) modality, are developed for the same task (*i.e.*, ASD diagnosis), and also have the same output space (ASD versus Control). However, these datasets are collected separately by different groups of researchers under different conditions. Due to this, they lie in completely different spaces and their sample distributions are non-comparable, thus making this a multi-dataset task. Further, one (Autism) dataset is *easy* for which we can train a model that achieves high accuracy (*c.f.*, Table 2), and the other (Hand Gesture) dataset is *difficult* for which the accuracy is relatively low (*c.f.*, Table 1), while considering the same network architecture (learning capacity) for both. The quantitative results discussed above indicate that while the classification accuracy is nearly saturated on the Autism dataset, it can contribute in boosting the accuracy on the more challenging Hand Gesture dataset through our contrastive learning based multi-dataset approach. In other words, our approach seamlessly distills (extracts and propagates) task-specific knowledge from the easy dataset to the model trained for the difficult dataset and improves its accuracy, thus resulting in *data-level* knowledge distillation. As per our knowledge, this is the first such attempt on this task, and we believe that our approach may also benefit other healthcare/biomedical applications.

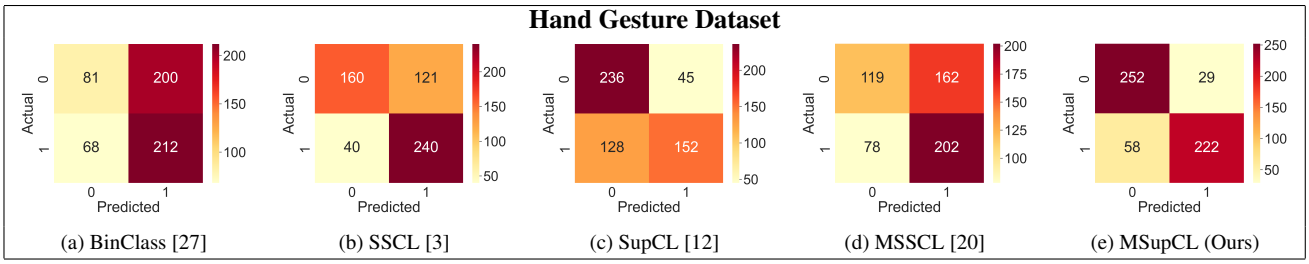
4.5. Analyses

The above results are supported by the t-SNE [29] visualization of the features learned by different methods on the two datasets as shown in Figure 4. Here, we can observe that for the Hand Gesture dataset, the samples from the two classes (ASD and Control) are best separated using our MSupCL approach. However, for the Autism dataset, the samples are well-separated in all the cases, thus leading to a high classification accuracy.

Figure 5 shows the confusion matrix for all the methods on the Hand Gesture dataset. (we do not include the confusion matrix for the Autism dataset as all the methods achieve a near perfect accuracy on this dataset with statistically insignificant difference). We can observe that the binary classifier classifies most of the samples as ASD, which means it cannot differentiate between the ASD and Control samples. On the other hand, while SupCL and MSSCL correctly predict majority of Control samples, the number of incorrect ASD predictions is relatively higher. SSCL performs good for ASD samples but also misclassifies a large number of Control samples to ASD. MSupCL improves upon SSCL as it correctly classifies more number of Control samples than any other method, while the number of correct predictions for ASD is comparable to SSCL.

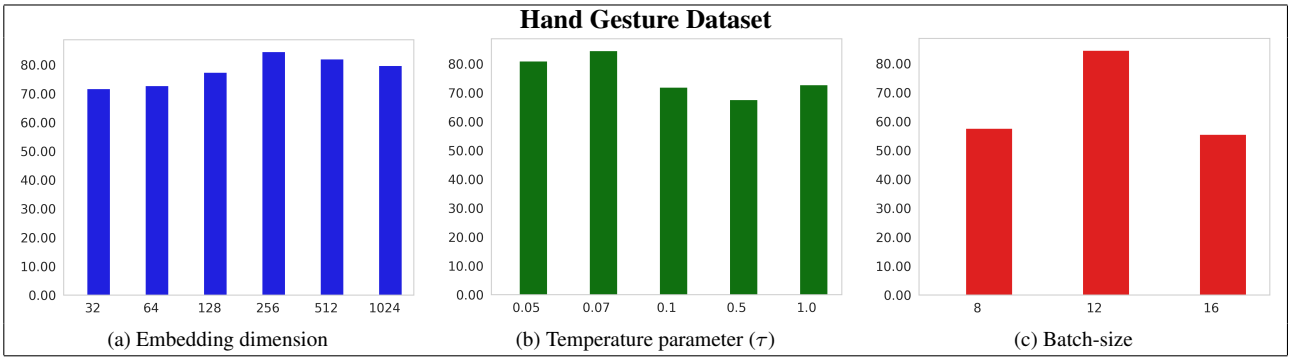
4.6. Ablation Study

In Figure 6, we analyze the impact of different hyper-parameters on MSupCL's performance using the Hand Gesture dataset (we use the same set of hyper-parameters for both the datasets). In Figure 6(a), we first study the importance of the final feature embedding size (z) by varying it in the range $\{32, 64, 128, 256, 512, 1024\}$. We observe that initially the accuracy increases as we increase the em-



(a) BinClass [27] (b) SSCL [3] (c) SupCL [12] (d) MSSCL [20] (e) MSupCL (Ours)

Figure 5. Confusion matrix for different methods on the Hand Gesture dataset. Here, '0' corresponds to the Control category and '1' corresponds to the ASD category.



(a) Embedding dimension (b) Temperature parameter (τ) (c) Batch-size

Figure 6. Ablation study on the Hand Gesture dataset by varying different hyper-parameters. In each plot, the vertical axis denotes percentage classification accuracy.

bedding size, and then it starts to drop. Next, we study the impact of the temperature parameter τ in Figure 6(b), where we observe that a lower value of τ gives better results. In Figure 6(c), we compare the model's performance using different batch-sizes and find that the accuracy first improves and then declines on increasing the batch-size. It is interesting to note that our observations on projection dimension and temperature are in-line with those reported in [3, 12]. However, we observe a decline in performance on increasing the batch-size from 12 to 16. We believe this is because of the characteristics of this dataset where we have only two categories to distinguish (ASD and Control), and a medium batch-size of 12 is possibly optimum to segregate a group of similar and dissimilar pairs in a relative manner.

4.7. Qualitative Results

In Figure 7, we compare the predictions made by the two top-performing methods SSCL and MSupCL on examples from the Hand Gesture dataset, for both ASD and Control categories. Along with each example, we also show the confidence score of these methods, and whether the prediction was correct/incorrect. Here, the first row shows examples that are correctly classified by both the methods, the middle row shows examples that are correctly classified by MSupCL but misclassified by SSCL, and the last row shows examples that are misclassified by both the methods. In general, we observe that MSupCL has a relatively high confidence score in cases where it makes correct predictions, and a low (near chance) confidence when it makes an incorrect prediction. On the other hand, SSCL has a relatively low confidence score in cases where it makes correct predictions, and a high confidence when it makes an incorrect prediction. Overall, these results validate the promise of the presented MSupCL approach on this challenging task.

5. Summary and Conclusion

Automated ASD diagnosis is a challenging and long-standing research problem. Over the last few years, different groups of researchers have independently collected various datasets to demonstrate the effectiveness of existing machine learning techniques on this task. These datasets generally contain a small number (a few hundreds) of samples having low inter-class and high intra-class variability. In this paper, we have made an attempt towards addressing these challenges by integrating knowledge from two independently collected and significantly diverse video datasets in a contrastive learning set-up. To do so, we have presented a multi-dataset supervised contrastive learning technique, and empirically demonstrated its superiority over the competing techniques such as [3, 12, 20]. On a general note, our experiments demonstrate that contrastive learning techniques, that learn discriminative features in a relative man-

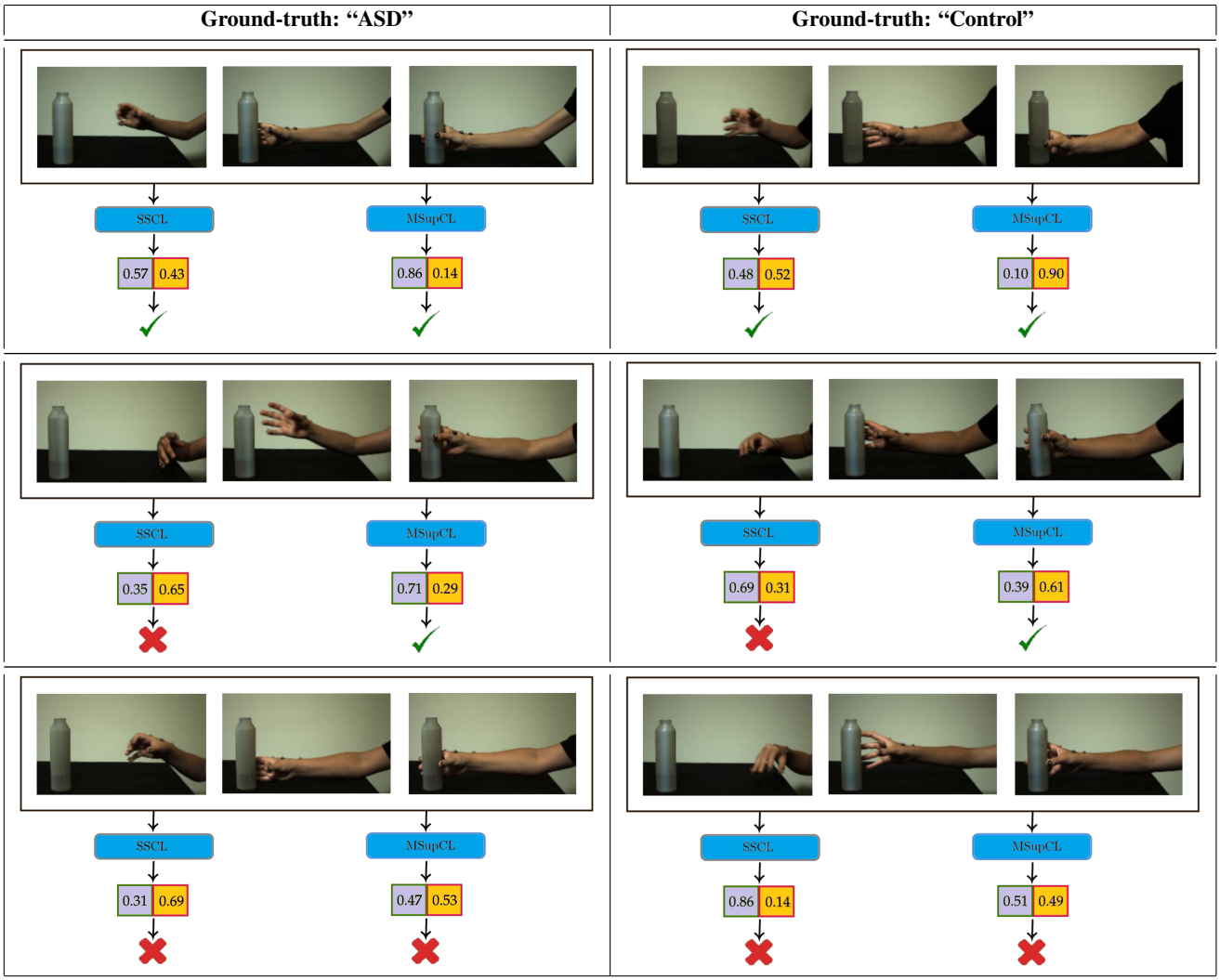


Figure 7. Qualitative comparisons from the Hand Gesture dataset. The left and right columns show examples from the ASD and Control categories respectively. Along with each example, we show the confidence score of SSCL [3] and MSupCL (ours).

ner, can be quite beneficial in automating healthcare-related tasks that suffer from the above challenges.

Limitations and Potential Negative Social Impact

While our approach outperforms competing techniques and achieves compelling results given the challenges involved in this task, one important limitation of all the examined techniques is their substantially high false-positive and false-negative rates (Figure 5). Because of this, we believe more research efforts will be required to make such systems deployable. Also, since ours is a computational learning based study, our experiments consumed energy produced by burning of fossil fuels and warmed our planet.

Compliance with ethical standards

This research study was conducted retrospectively using human subject data provided by the authors of [32] (Hand Gesture dataset) and [19] (Autism dataset) via a registration process. No additional ethical approvals were required.

Acknowledgments: The authors would like to thank the Ministry of Education (India) for financial support. YV would like to thank the Department of Science and Technology (India) for the INSPIRE Faculty Award 2017.

References

[1] Fahd A. Alturki, Majid Aljalal, Akram M. Abdurraqeeb, Khalil Alsharabi, and Abdullrahman A. Al-Shamma'a. Common spatial pattern technique with eeg signals for diagnosis of autism and epilepsy disorders. *IEEE Access*, 9:24334–24349, 2021.

[2] Mehmet Baygin, Sengul Dogan, Turker Tuncer, Prabal Datta Barua, Oliver Faust, N. Arunkumar, Enas W. Abdulhay, Elizabeth Emma Palmer, and U. Rajendra Acharya. Auto-

mated asd detection using hybrid deep lightweight features extracted from EEG signals. *Computers in Biology and Medicine*, 134:104548, 2021.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020.

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[5] Yanbei Chen, Manchen Wang, Abhay Mittal, Zhenlin Xu, Paolo Favaro, Joseph Tighe, and Davide Modolo. Scaledet: A scalable multi-dataset object detector. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[6] Geraldine Dawson, Sara Jane Webb, and James McPartland. Understanding the nature of face processing impairment in autism: Insights from behavioral and electrophysiological studies. *Developmental Neuropsychology*, 27(3):403–424, 2005.

[7] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[9] Anibal Solon Heinsfeld, Alexandre Rosa Franco, Richard Cameron Craddock, Augusto Buchweitz, and Felipe Meneguzzi. Identification of autism spectrum disorder using deep learning and the abide dataset. *NeuroImage : Clinical*, 17:16 – 23, 2018.

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[11] M. Jiang and Q. Zhao. Learning visual attention to identify people with autism spectrum disorder. In *IEEE International Conference on Computer Vision*, 2017.

[12] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

[13] Yazhou Kong, Jianliang Gao, Yunpei Xu, Yi Pan, Jianxin Wang, and Jin Liu. Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. *Neurocomputing*, 324:63–68, 2019.

[14] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *IEEE International Conference on Computer Vision*, 2011.

[15] Jung Hyuk Lee, Geon Woo Lee, Guiyoung Bong, Hee Jeong Yoo, and Hong Kook Kim. End-to-end model-based detection of infants with autism spectrum disorder using a pretrained model. *Sensors*, 23(1):202, 2023.

[16] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *IAPR Asian Conference on Pattern Recognition*, pages 730–734, 2015.

[17] Wenbo Liu, Ming Li, and Li Yi. Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*, 9, 2016.

[18] MJ Maenner and Kelly A. Shaw. Prevalence and characteristics of autism spectrum disorder among children aged 8 years. *Report*, 2021. Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2018.

[19] Prashant Pandey, A P Prathosh, Manu Kohli, and Josh Pritchard. Guided weak supervision for action recognition with scarce data to assess skills of children with autism. In *AAAI Conference on Artificial Intelligence*, 2020.

[20] Mandela Patrick, Yuki M. Asano, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. In *IEEE International Conference on Computer Vision*, 2021.

[21] Mladen Rakic, Mariano Cabezas, Kaisar Kushibar, Arnau Oliver, and Xavier Llado. Improving the detection of autism spectrum disorder by combining structural and functional mri information. *NeuroImage: Clinical*, 25:102181, 2020.

[22] Mindi Ruan, Paula J. Webster, Xin Li, and Shuo Wang. Deep neural network reveals the world of autism from a first-person perspective. *Autism Research*, 14(2):333–342, 2021.

[23] Zeinab Sherkatghanad, Mohammad Sadegh Akhondzadeh, Soorena Salari, Mariam Zomorodi-Moghadam, Moloud Abdar, U. Rajendra Acharya, Reza Khosrowabadi, and V. Salari. Automated detection of autism spectrum disorder using a convolutional neural network. *Frontiers in Neuroscience*, 13, 2019.

[24] K. Sun, L. Li, L. Li, N. He, and J. Zhu. Spatial attentional bilinear 3D convolutional network for video-based autism spectrum disorder detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[25] Md. Nurul Ahad Tawhid, Siuly Siuly, Hua Wang, Frank Whittaker, Kate Wang, and Yanchun Zhang. A spectrogram image based intelligent technique for automatic detection of autism spectrum disorder from eeg. *PLOS ONE*, 16(6):1–20, 06 2021.

[26] Yuan Tian, Xiongkuo Min, Guangtao Zhai, and Zhiyong Gao. Video-based early asd detection via temporal pyramid networks. *IEEE International Conference on Multimedia and Expo*, 2019.

[27] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[28] Constantin Ulrich, Fabian Isensee, Tassilo Wald, Maximilian Zenk, Michael Baumgartner, and Klaus H. Maier-Hein. MultiTalent: A multi-dataset approach to medical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2023.

[29] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[30] Shuo Wang, Ming Jiang, Xavier Morin Duchesne, Elizabeth A. Laugeson, Daniel P. Kennedy, Ralph Adolphs, and

Qi Zhao. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, 88(3):604–616, 2015.

[31] Hao Zhu, Jun Wang, Yin-Ping Zhao, Minhua Lu, and Jun Shi. Contrastive multi-view composite graph convolutional networks based on contribution learning for autism spectrum disorder classification. *IEEE Transactions on Biomedical Engineering*, 70(6):1943–1954, 2023.

[32] A. Zunino, P. Morerio, A. Cavallo, C. Ansuini, J. Podda, F. Battaglia, E. Veneselli, C. Becchio, and V. Murino. Video gesture analysis for autism spectrum disorder detection. In *IAPR International Conference on Pattern Recognition*, 2018.