Early-stage autism diagnosis using action videos and contrastive feature learning

Asha Rani1 · Pankaj Yadav2,3 · Yashaswi Verma1

Received: 24 March 2023 / Accepted: 28 June 2023 / Published online: 8 July 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract
Autism, also known as Autism Spectrum Disorder (or ASD), is a neurological disorder. Its main symptoms include difficulty in verbal/non-verbal communication and rigid/repetitive behavior. These symptoms are often indistinguishable from a normal (control) individual due to which this disorder remains undiagnosed in early childhood, thus leading to a delayed treatment. Since the learning curve is steep during the initial years, an early diagnosis of autism would allow to make an early intervention, which in turn would positively affect the growth of an autistic child. Further, the traditional methods of autism diagnosis require multiple visits to a specialized doctor; however, this process is generally time-consuming. In this paper, we present a learning-based approach to automate autism diagnosis using simple and small action video clips of subjects. This task is particularly challenging because the amount of annotated data available is small, and the variations among samples from the two categories (ASD and control) are generally indistinguishable. This is also evident from poor performance of a baseline binary classifier on this task. To address this, we propose to adopt contrastive feature learning for the first time on this task and demonstrate a significant increase in the prediction accuracy. We further validate this by conducting thorough experimental analyses in both self-supervised and supervised setups on two publicly available datasets. We have also released our codes and pre-trained models for reproducibility.

Keywords Autism diagnosis · Contrastive feature learning · Deep learning

1 Introduction
Contrastive feature learning is a popular machine learning technique that has been used in a variety of tasks in the recent times [1–4]. The main advantage of this approach is its ability to learn feature embeddings in a discriminative manner, which is particularly beneficial when the availability of labeled data is limited. This situation is often encountered in the medical domain where the process of data annotation is quite costly and time-consuming. Autism is one such medical condition where it is challenging to reliably collect and annotate data, thus leading to small datasets for training machine learning algorithms.

Autism, also known as Autism Spectrum Disorder (or ASD), is a neurological disorder which largely affects an individual’s cognitive capabilities, with the main symptoms being difficulty in social interaction and a rigid behavior. The number of individuals suffering from autism has increased over the last few years [5], due to which its diagnosis at an early stage can be helpful in an early intervention. Traditionally, this has been done by trained medical practitioners, which is time-consuming and may delay an early (correct) diagnosis and subsequent intervention.

In this work, our objective is to perform autism diagnosis in an automated manner. There are some recent machine-learning-based methods which have proposed to automate the ASD diagnosis process. These works have investigated...
data from different modalities, such as eye tracking data, gesture videos, MRI data, EEG data, etc. [6–16]. Earlier analysis revealed that individuals with autism have an atypical sight [17]. For example, [7, 18] worked on this objective, with [7] focusing on visual bias toward different objects, contrast, and color, and [18] on using these aspects to predict ASD and control subjects using deep visual models like VGG16 [19]. A different perspective to distinguish the two categories was introduced in [8] where the first-person view of a scene (or visual) of an individual is compared for analysis. A few works have used eye tracking data to distinguish on the basis of attention [10], hand gesture data (small action clips) [9, 11], EEG signals [12, 20, 21] and MRI data [14, 16].

In this paper, we approach the autism diagnosis task using simple and short action (color/gray scale) videos. This is because such data are much easier to acquire compared to other modalities such as EEG, MRI or eye-tracking data, and can also be readily used for training deep-learning-based models. However, it is difficult to obtain good prediction accuracy using such data because it exhibits small inter-class variability when acquired in a flexible/less-controlled environment; e.g., Fig. 1 shows a sample from both ASD and control categories of one such dataset (the Hand Gesture dataset [9]). We can observe that both the samples look quite identical, and if we attempt to differentiate them just by seeing them, it would be nearly impossible to identify the correct category of a sample. Because of this, the conventional deep learning models used in visual classification tasks may not be directly adopted for this task. This is also evident from our experimental analyses where a conventional deep binary classifier gives poor accuracy.

To address this challenge, we propose to use contrastive feature learning to learn minute distinctions between the two categories. It is also important to note that our application of contrastive feature learning is quite different from the conventional computer vision tasks since (a) our datasets contain a small number of samples, (b) the data are inherently complex being video data, and (c) the samples within a dataset may exhibit low inter-class variability and are nearly impossible to distinguish visually by a normal human being. We adopt both self-supervised and supervised contrastive feature learning techniques, and conduct thorough experiments to demonstrate that these techniques can be quite beneficial in such tasks, and the features, thus, learned lead to a significant increase in the prediction accuracy. As per our knowledge, ours is the first attempt that demonstrates the effectiveness of contrastive feature learning on the autism diagnosis task. 1

In the next section, we describe the related work in autism diagnosis and other relevant domains of our interest (i.e., action recognition in videos, and contrastive feature learning and its applications in biomedical domain). In Sect. 3, we describe the self-supervised and supervised contrastive feature learning techniques that we adopt to learn discriminative features using action videos for autism diagnosis. In Sect. 4, we provide the experimental details and discuss our empirical findings under different experimental setups, and finally we conclude in Sect. 5.

## 2 Related work

Since our focus is on the problem of ASD diagnosis using action videos and contrastive feature learning, we first present an overview of the related advancements in action recognition, followed by contrastive feature learning and its applications in biomedical problems, and finally ASD diagnosis.

### 2.1 Action recognition

In the recent times, one of the earlier works on action recognition involved using two-stream networks to learn the temporal information in the videos. Optical flow [22] was used instead of traditional features like IDT [23] to obtain an effective motion representation. The authors of [24] created a spatial and temporal stream based on human visual pathways. The spatial stream was used to capture the visual appearance, while the temporal stream used a stack of optical flow images to capture information in between frames. It outperformed the previously used handcrafted IDT features in terms of empirical results. As the two stream network used in [24] was shallow, the authors of [25] used a deep network, this did not improve the results due to overfitting on small datasets. To address this, the authors of [25] introduced cross-modality initialization, synchronization in

---

1 Our code and pre-trained models are available here https://github.com/asharani97/CLRE_autism for reproducibility.
Early-stage autism diagnosis using action videos and contrastive feature learning

Batch normalization, multi-scale cropping, data augmentation, and large dropout ratio using the VGG16 [19] model, which outperformed the previous large dataset methods like UCF101 [26] by a large margin. The two stream networks require spatio-temporal fusion, and the most widely used fusion was late fusion. The authors of [27] demonstrated that early fusion of spatial and temporal information improved feature learning, which was later adopted in [28–30]. Following this, the videos were investigated in [31] in a segment-by-segment manner, resulting in the temporal segment network. As one frame from each segment is used to learn the network, information from all the frames was used here. Later, TRN [32] was proposed to improve the reasoning ability by learning dependencies between video frames. The authors of [33] employed a slow pathway to analyze frames at a low frame rate and to capture spatial information, and a rapid pathway to obtain temporal information, since 2D networks used on video were ineffective at capturing all of this information. As a video contains a variety of data such as pose, object, audio, depth, etc., in [34, 35], multi-stream networks were used. The 3D CNN models were then extended in [36] to a deeper 3D network known as C3D. Later, many concepts emerged based on factorization of 2D and 3D CNNs to capitalize on the advantages of both the types of neural networks. One such model R(2+1)D was proposed in [37], in which the 2D convolution is spatial while the 1D convolution is temporal. In this paper, we have adopted this R(2+1)D network as the initial video encoder. Furthermore, EAN [38] suggested a strategy for dealing with the problem of action recognition in videos with many events. This consisted of three components: one to recognize the event, second to extract the related features using a 3D network, and third to aggregate the discovered events. The event was identified by collecting identical motions in the frames. In [39, 40], the authors solved this challenge without using labeled data, by employing self-supervised learning to anticipate the future frames in a video. After extracting the motion cues, the scattering transform was utilized to encode the local motion cues in the feature vector. In our work, we apply contrastive feature learning in both self-supervised and supervised setups coupled with the R(2+1)D network [37], and demonstrate its effectiveness on the autism diagnosis task using action videos. This way, while the data that we work with contain different types of actions, our goal is to identify whether a given subject is autistic irrespective of the action being performed.

2.2 Contrastive learning and biomedical applications

Contrastive feature learning is the process of learning discriminative features using similar and dissimilar pair of data points. During training, the model attempts to learn the parameters such that the similar data points can be brought closer to each other and dissimilar data points can be pushed farther. There are several techniques that use the principle of contrastive learning, including SimCLR [1], SupCLR [2] MoCo [3], and SimpleSiamese [4]. While some of these are unsupervised/self-supervised, others are supervised. MoCo [3] is an unsupervised technique that is based on momentum contrast and generates a large number of negative samples during training. In [4], a Siamese network is proposed which learns shared weights across the two encoder networks, and another approach [41] uses selective aggregation for video scaling.

The success of contrastive feature learning in biomedical applications has led to their widespread applications in the medical domain. Using the textual reports, the authors of [42] adopted a report-guided contrastive technique to learn distinctions in the pathology model. In [42], similarity between medical photos and their accompanying reports are learned using a contrastive triplet loss. In [43], the authors employ contrastive learning to generalize a multi-graph representation to diverse modalities such as neuroimaging data and functional connections to portray the brain’s structure. Motivated by these, we examine and demonstrate the effectiveness of contrastive learning in the autism diagnosis task using simple action videos.

2.3 ASD diagnosis

Several machine-learning-based techniques have been proposed in the recent years that attempt to perform autism diagnosis in an automated manner. These attempts have investigated data acquired in the form of different modalities, such as eye tracking data, gesture videos, MRI data, EEG data, etc. [6–16]. One such study reported that an autism affected person shows atypical visual saliency [17]. To quantify this observation, Wang et al. [18] used a three-layer model to discriminate between the ASD and control subjects based on their visual preferences using semantics and pixel information. However, they did not target the problem of automatically classifying the two categories. In [7], the authors used the same dataset but for a different task. Specifically, they used the fisher score to first select the most discriminative images to compare the visual attention of the ASD and control subjects, and then used the VGG16 [19] network followed by an SVM classifier for prediction. Along the similar lines, instead of working with the eye gaze data, Ruan et al. [8] concentrated on analyzing the visual attention of subjects by asking them to click pictures of a given scene, thus indirectly analyzing their visual attention. In [9], the authors used the idea of attention mechanism [44, 45] and focused on hand gesture videos. They use the knowledge that each video frame has a contextual relation to the one before it. In [11], the authors showed that the use of bi-linear...
pooling results in the loss of spatial information in a frame, thus making autism classification difficult, and used spatial attention bi-linear pooling to address this.

Some of the other approaches for ASD diagnosis include using data from other modalities. In [12], the authors proposed a method where EEG signals are first converted into spectrogram images and then used for automatic detection. Similar concept of using spectrogram images is also used in [20, 21]. In [13], the authors used both structural as well as functional MRI information for the study of ASD. To do so, they used stacked autoencoders for unsupervised classification and multi-layer perceptrons for supervised learning. In [14], the authors investigated the neural patterns which are most associated with ASD. They generalized complex patterns and used SVM along with an autoencoder model to distinguish between ASD and control subjects. In [15], a multi-channel convolutional neural network (CNN) was proposed based on a patch-level data-expanding method to diagnose ASD, where multivariate and high-dimensional data are reduced to two-dimensional features, and the functional connectivity pattern related to ASD is explored using a variational autoencoder model.

Unlike the existing works, for the first time, we demonstrate the effectiveness of contrastive feature learning in ASD diagnosis using action videos, which is possibly the simplest modality for collecting such a sensitive data.

### 3 Contrastive feature learning

Below we discuss the contrastive feature learning techniques we have adopted in this work. Specifically, we perform contrastive learning in both self-supervised [1] and supervised [2] setups. Lately, these techniques have been found to be useful in learning discriminative features by making use of additional data generated by doing minor variations in the original data, and then creating pairs of similar and dissimilar samples to learn discriminative features.

#### Notations:

The basic notations used in the paper are as follows:

- **Video / frame(s):** A pair of video (frames) is represented as $f_{ij}$ where $f_i$ and $f_j$ are the individual video / frames (samples) in a particular pair.
- **Transformation function:** The transformation function $a()$ is applied on a frame $f$ to produce its transformed version.
- **Encoder:** The neural network used for encoding an input video $f_i$ is represented as $e()$, and the output feature vector obtained is given by $x_i$.
- **Non-linear head:** The feature vector $x_i$ is passed through a non-linear head $h()$ (a multi-layer perceptron) to get a new feature vector $z_i$.

Fig. 2 illustrates one of the widely used network architectures for contrastive feature learning, which is also adapted in this work.

#### 3.1 Self-supervised contrastive learning

Contrastive learning generally requires a memory bank or a specialized network architecture to learn discriminative visual representations of objects. This leads to an increased memory requirement and/or a complex training procedure. To address this, in [1], the authors proposed a self-supervised contrastive learning technique named SimCLR which does not require any specialized architecture or memory bank. Instead, they integrated various existing and simple approaches in the literature for self-supervised learning, and achieved state-of-the-art results on the standard image classification task. Specifically, this approach uses simple transformations (e.g., vertical flip, addition of a small noise, etc.) to generate new samples, and uses them to create positive and negative pairs. A positive pair is created using a given sample and its transformed version, and a negative pair is created by combining a given sample (or its transformed version) with another sample (or its transformed version). During training, both the samples in a pair are passed through
Supervised contrastive learning

In [2], the authors extended the above self-supervised contrastive learning technique by introducing supervision in the form of labels available in the training data. This results in supervised contrastive learning (SupCLR), where the label information is used to identify multiple positive samples for a given sample in a batch. Unlike the self-supervised learning setup where the loss is calculated based on a given sample and its transformed version, SupCLR uses all the positive samples to create positive pairs (i.e., the samples that belong to the same category) in a mini-batch. As a result, while there is only one positive pair in case of self-supervised contrastive learning, there may be multiple such pairs in case of supervised contrastive learning. For a given sample, the loss in SupCLR is defined as follows:

\[
p_{sup} = \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(z_i, z_p) / \tau)}{\sum_a \exp(\text{sim}(z_i, z_a) / \tau)}
\]

where \(P(i)\) denotes the set of indices of positive samples with respect to the \(i^{th}\) sample in a mini-batch, and \(A\) denotes the set of indices in the mini-batch excluding \(i\). In the process of representation learning using the above loss function, clustering of data points is expected to improve in the presence of labels as this avoids data points from different categories to come closer.

Figure 3 compares the network architecture of a conventional deep binary classifier with those of SimCLR and SupCLR. Compared to the binary classifier, SimCLR makes use of a transformation function to generate a similar (or positive) sample, whereas SupCLR uses the category information to identify a positive sample with respect to a given sample in a mini-batch.

4 Experiments

4.1 Datasets

Our selection of datasets is inspired by two practical considerations: (a) In medical domains, datasets are generally small in size due to high annotation cost and sensitive nature of data. (b) Sometimes, it is not feasible to collect samples from all possible categories of a particular disease by one group of researchers, because of which the generated dataset contains
As we are working with video data, we use $R(2+1)D$ [37] as the base encoder network $e()$ for all the three methods. This network contains 18 layers, and its parameters are randomly initialized. A set containing a fixed number of ordered frames (i.e., a clip of a fixed length) is passed through the encoder $e()$ (similarly for an augmented/transformed version of frames in case of SimCLR). The output of the encoder is a 512-dimensional feature vector for the entire video (set of frames). This feature vector is then passed to a non-linear MLP layer $h()$ which gives a 256-dimensional feature vector as the output. Once this entire network is trained, the two contrastive learning methods SimCLR and SupCLR drop the MLP layer $h()$ and use only the learned encoder $e()$. On top of this pre-trained encoder, we add a linear classification layer and then re-train the entire network using the labeled training data following the procedure described in [1, 2].

### 4.3 Training details

As part of data pre-processing, each video is loaded in the form of a sequence of frames. We pick 16 frames in case of the Hand Gesture dataset and 10 frames in case of the Autism dataset following uniform sampling. We choose a different number of frames per clip for the two datasets since the videos in the Hand Gesture dataset contain 21–30 frames, whereas those in the Autism dataset contain 12–20 frames. In both the datasets, the training and test data consist of an equal number of samples from both the classes by creating a train–test split in the ratio 70:30. To create positive pairs in SimCLR, we perform horizontal flip to generate a transformed sample.

**Hyperparameters:** For training the binary classifier, we use the standard binary cross-entropy loss with the learning rate of $1e^{-3}$, weight decay of $1e^{-6}$, and the Adam optimizer. During the training of SimCLR, we use the learning rate of $1e^{-4}$, weight decay of $1e^{-5}$, and the Adam optimizer. In case of SupCLR, we use the learning rate of $1e^{-3}$, weight decay of $1e^{-6}$, momentum of 0.9, and the SGD optimizer.

### 4.4 Quantitative results

We compare the performance of binary classifier and contrastive learning methods on the Hand Gesture Dataset in Table 1 (top). Here, we observe that the classification accuracy improves drastically when we use contrastive learning compared to the binary classifier. Further, while the supervised contrastive learning method performs much better than the binary classifier, it is slightly outperformed by the self-supervised method by around 2%. Table 1 (bottom) shows the classification results on the Autism Dataset using all the three methods. Based on the average accuracy, we can observe that self-supervised contrastive learning outperforms both supervised contrastive learning and binary classifier by around 1.2% and 1.73% respectively.
4.5 Cross-dataset analysis

In this experiment, we perform pre-training (training of the encoder) and evaluation in a cross-dataset manner; i.e., we perform training using the train set of one dataset and evaluate on the test set of another dataset. Table 2 (top) shows results on the Hand Gesture Dataset when pre-training is done on the Autism Dataset. We observe that the accuracy obtained using self-supervised contrastive learning is not affected significantly, and it performs better than the other two methods. Table 2 (bottom) shows results on Autism Dataset when pre-training is done on the Hand Gesture Dataset with similar trends. Interestingly, in case of self-supervised contrastive learning, we observe that the baseline results are slightly inferior than those obtained using cross-data training.

4.6 Mixed dataset analysis

In real-life applications of autism diagnosis, the data points used in training and testing may come from different distributions. To analyze the effectiveness of different methods in such situations, we perform experiments on mixed dataset setups. For this purpose, we simulate three setups as described below. As before, the train–test split remains 70:30 in all the experiments.

4.6.1 Setups

1. Setup 1: In this case, the training set consists of ASD samples from both the datasets. However, the control samples come only from the train split of the Hand Gesture Dataset. For the test set, the following three cases are considered:
   - Test 1: Test set of the Hand Gesture Dataset with only ASD samples.
   - Test 2: Test set of the Autism Dataset with both the classes.
   - Test 3: Combination of the test sets of both Autism and Hand Gesture Datasets.

2. Setup 2: In this case, the training set consists of ASD samples from the train set of the Hand Gesture Dataset, and ASD samples from the train+test set of the Autism Dataset. For the test set, the following case is considered:
   - Test 1: Test set of the Hand Gesture Dataset.

3. Setup 3: In this case, the training set consists of the merged train sets of both Autism and Hand Gesture Datasets from both ASD and control categories. For the test set, the following cases are considered:
   - Test 1: Test set of the Hand Gesture Dataset.
   - Test 2: Test set of the Autism Dataset.
   - Test 3: Combination of the test sets of both Autism and Hand Gesture Datasets.

4.6.2 Results

Table 3 shows results for the three setups discussed above. We can observe that for Setup 1, the accuracy does not improve on the Hand Gesture Dataset. However, for the Autism dataset, the accuracy improves as compared to the baseline and cross-dataset results on the same test data. In case of Setup 2, the results obtained are reasonable; however, they are inferior to the baseline results obtained using self-supervised learning. In Setup 3, the training sets of both
For the Hand Gesture Dataset, the accuracy drops significantly for the self-supervised contrastive learning method. However, for the Autism Dataset, the accuracy is similar to that obtained in cross-dataset evaluation. On the other hand, using supervised contrastive learning, the accuracy is comparable to that when training alone on the dataset as in Sect. 4.4; however, it is better than the cross-dataset training. We also observe that since the third test setup (Test 3) is a combination of Test 1 and Test 2, it performs as expected; i.e., approximately average of the accuracies on the two datasets.

In general, these results indicate that contrastive feature learning methods may be preferred over simple binary classifier for small data problems such as autism diagnosis, which are frequently encountered in the healthcare domain.

### 4.7 Visualization using t-SNE

The visualization of data samples using t-SNE [49] helps us for a better explanation of the results obtained above. In case of Hand Gesture Dataset, we can see in Fig. 5 (top row) that the data points from the two classes are quite overlapping and clear separation is not possible in case of binary classifier. In SupCLR, the overlap is less in comparison to binary classifier, and is further reduced on using SimCLR. These distributions justify the quantitative results on this dataset as reported in Table 1. In case of Autism Dataset, we can see in Fig. 5 (bottom row) that the data points from both the classes are well separated and clustered. As a result, all the three methods achieve competitive and high classification accuracy.

### Table 3 Percentage classification accuracy for mixed dataset analysis (Sect. 4.6)

| Dataset  | Accuracy % | BinClassifier | SimCLR | SupCLR |
|----------|------------|---------------|--------|--------|
| Training | Testing    |               |        |        |
| Setup 1  | Test 1     | 54.37         | 67.02  | 59.53  |
|          | Test 2     | 100           | 100    | 100    |
|          | Test 3     | 74.93         | 83.25  | 77.77  |
| Setup 2  | Test 1     | 56.51         | 70.14  | 57.40  |
| Setup 3  | Test 1     | 55.44         | 61.50  | 60.96  |
|          | Test 2     | 99.07         | 99.74  | 97.22  |
|          | Test 3     | 80.47         | 82.37  | 81.61  |

The best results are highlighted in bold.

#### Fig. 5 t-SNE visualization of the learned features on the two datasets. Green dots represent ASD samples while orange dots represent control samples

Hand Gesture Dataset

- (a) BinClassifier
- (b) SimCLR
- (c) SupCLR

Autism Dataset

- (d) BinClassifier
- (e) SimCLR
- (f) SupCLR

the datasets are merged. For the Hand Gesture Dataset, the accuracy drops significantly for the self-supervised contrastive learning method. However, for the Autism Dataset, the accuracy is similar to that obtained in cross-dataset evaluation. On the other hand, using supervised contrastive learning, the accuracy is comparable to that when training alone on the dataset as in Sect. 4.4; however, it is better than the cross-dataset training. We also observe that since the third test setup (Test 3) is a combination of Test 1 and Test 2, it performs as expected; i.e., approximately average of the accuracies on the two datasets.

In general, these results indicate that contrastive feature learning methods may be preferred over simple binary classifier for small data problems such as autism diagnosis, which are frequently encountered in the healthcare domain.

#### 4.8 Analysis using confusion matrix

Figure 6 (top row) shows confusion matrix on the Hand Gesture Dataset using the three methods. We observe that both binary classifier and SimCLR are correctly predicting more ASD samples than Control samples; however, the difference in case of binary classifier is more, which is also evident.
from the achieved accuracy. The binary classifier predicts most of the Control samples as ASD, which may be because it is unable to learn the minute differences between the two categories (as also observed in the t-SNE plot Fig. 5a). On the other hand, SupCLR correctly predicts Control samples more as compared to ASD. Figure 6 (bottom row) shows confusion matrix on the Autism Dataset, where we observe that the performance of all the three models is quite comparable.

4.9 Qualitative results

Figure 7 shows some qualitative examples from the Hand Gesture Dataset considering different cases. Along with each example, we also show the confidence scores of all the three methods, where the left score (highlighted in green) denotes the confidence score of the ASD category and the right score (highlighted in red) denotes the confidence score of the Control category. The examples in the top row are misclassified by all the three methods. However, we observe that while the binary classifier misclassifies them with a high confidence, both SimCLR and SupCLR misclassify them with a low (near chance) confidence scores. The examples in the middle row are misclassified by both binary classifier and SupCLR but correctly classified by SimCLR. Here, we observe that while the binary classifier misclassifies these examples with a high confidence, SupCLR misclassifies them with a low confidence score. On the other hand, SimCLR correctly classifies them with a high confidence score. The examples in the third row are correctly classified by both SimCLR and SupCLR but misclassified by the binary classifier. Here, we observe that the confidence score for correct classification in case of SimCLR is higher than SupCLR, while the binary classifier misclassifies them with a high confidence score. Overall, these results demonstrate the superiority of both SupCLR and SimCLR compared to binary classifier, and also show that SimCLR may give some benefit over SupCLR.

5 Summary and conclusion

Diagnosing autism at an early age is a challenging yet crucial task, and automating this process can potentially help in an early intervention, thus significantly benefiting the growth of an affected child. Using simple action videos for this automated diagnosis would significantly impact its reach by eliminating the requirement of costly equipments and trained manpower for collecting data in other modalities such as EEG or fMRI. In this paper, we have made an initial attempt in this direction. Specifically, we have demonstrated the effectiveness of the recent contrastive feature learning techniques in performing autism diagnosis with high accuracy using simple action videos. We observed that self-supervised contrastive learning consistently outperforms not only the competitive deep binary classifier but also supervised contrastive learning. This is in contrast with the earlier findings reported on large(million)-scale datasets [2], and indicates
that self-supervised contrastive learning may be preferred over supervised contrastive learning while working with small-scale datasets with low inter-class variability as in our case. We hope that our study will also motivate researchers to examine these techniques on other biomedical applications that involve scarce data. To facilitate this, we have made our codes and pre-trained models publicly available.

Acknowledgements 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 (registration no.: IFA17-221). We acknowledge the usage of the computing facility at IIT Jodhpur for conducting computational experiments. We are grateful to the authors of [9, 47] for making their datasets available upon request, and all the subjects who contributed in building these datasets.

Author Contributions All the authors have contributed to the development of the presented idea. A.R. wrote codes, conducted experiments, and prepared all the figures. P.Y. and Y.V. examined the results. Y.V. examined the codes. All the authors have contributed to manuscript writing and reviewing.

Data availability We have shared the link for our codes and pre-trained models in Sect. 1. The two datasets used in the paper were obtained from the authors of [9] (Hand Gesture Dataset) and [47] (Autism Dataset) via a registration process.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethical standards This research study was conducted retrospectively using human subject data made available by the authors of [9] (Hand Gesture Dataset) and [47] (Autism Dataset) via a registration process. No additional ethical approvals were required.

References

1. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (ICML) (2020)
2. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Adv. Neural Inf. Process. Syst. 33, 18661–18673 (2020)
3. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
4. Chen, X., He, K.: Exploring simple Siamese representation learning. In: CVPR (2021)
5. Maenner, M., Shaw, K., Bakian, A., Bilder, D., Durkin, M., Esler, A., Furnier, S., Hallas-Muchow, L., Hall-Lande, J., Hudson, A., Hughes, M., Patrick, M., Pierce, K., Poynter, J., Salinas, A., Shenouda J, Vehorn A, Warren Z, Constantin J, Cogswell M.: Prevalence and Characteristics of Autism Spectrum Disorder among Children Aged 8 years - Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2018. Morbidity and
Early-stage autism diagnosis using action videos and contrastive feature learning

mortality weekly report. Surveillance summaries (Washington, D.C.; 2002) 70, 1–16 (2021)
6. Liu, W., Li, M., Yi, L.: Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. Autism Res. 9, 888–898 (2016)
7. Jiang, M., Zhao, Q.: Learning visual attention to identify people with autism spectrum disorder. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3287–3296 (2017)
8. Ruan, M., Webster, P., Li, X., Wang, S.: Deep neural network reveals the world of autism from a first-person perspective. Autism Res. 14(2), 333–342 (2021)
9. Zunino, A., Morerio, P., Cavallio, A., Ansuini, C., Podd, J., Battaglia, F., Veneselli, E., Becchini, C., Murino, V.: Video gesture analysis for autism spectrum disorder detection. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 3421–3426 (2018)
10. Tian, Y., Min, X., Zhai, G., Gao, Z.: Video-based early ASD detection via temporal pyramid networks. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 272–277 (2019)
11. Sun, K., Li, L., Li, L., He, N., Zhu, J.: Spatial attentional bi-linear 3d convolutional network for video-based autism spectrum disorder detection. In: ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3387–3391 (2020)
12. Tawhid, M.N.A., Siuly, S., Wang, H., Whittaker, F., Wang, K., Zhang, Y.: A spectrogram image based intelligent technique for automatic detection of autism spectrum disorder from eeg. PLOS One 16(6), 1–20 (2021)
13. Rakić, M., Cabezas, M., Kushibar, K., Oliver, A., Lladó, X.: Improving the detection of autism spectrum disorder by combining structural and functional MRI information. NeuroImage 25, 102181 (2020)
14. Heinsfeld, A.S., Franco, A.R., Craddock, R.C., Buchweitz, A., Meneguzzi, F.: Identification of autism spectrum disorder using deep learning and the abide dataset. NeuroImage 17, 16–23 (2018)
15. Sherkatghanad, Z., Akhondzadeh, M.S., Safari, S., Zomorodi-Moghadam, M., Abdar, M., Acharya, U.R., Khosrowabadi, R., Safari, V.: Automated detection of autism spectrum disorder using a convolutional neural network. Front. Neurosci. 13, 1325 (2019)
16. Kong, Y., Gao, J., Xu, Y., Pan, Y., Wang, J., Liu, J.: Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. Neurocomputing 324, 63–68 (2018)
17. Dawson, G., Webb, S.J., McPartland, J.: Understanding the nature of face processing impairment in autism: insights from behavioral and electrophysiological studies. Dev. Neurosci. 27(3), 403–424 (2005)
18. Wang, S., Jiang, M., Duchesne, X.M., Laugesen, E.A., Kennedy, D.P., Adolphs, R., Zhao, Q.: Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. Neuron 88(3), 604–616 (2015)
19. Liu, S., Deng, W.: Very deep convolutional neural network based image classification using small training sample size. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), pp. 730–734 (2015)
20. Bayrin, M., Dogan, S., Tuncer, T., Datta Barua, P., Faust, O., Arunkumar, N., Abdulhay, E.W., Emma Palmer, E., Rajendra Acharya, U.: Automated ASD detection using hybrid deep lightweight features extracted from EEG signals. Comput. Biol. Med. 134, 104548 (2021)
21. Alturki, F.A., Aljalal, M., Abdurraqeeb, A.M., Alsharabi, K., Al-Shamma’a, A.A.: Common spatial pattern technique with EEG signals for diagnosis of autism and epilepsy disorders. IEEE Access 9, 24334–24349 (2021)
22. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. 17, 185–203 (1981)
23. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV, pp. 3551–3558 (2013)
24. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS, pp. 568–576 (2014)
25. Wang, L., Wang, Z., Xiong, Y., Qiao, Y.: CUHK & SIAT submission for THUMOS’15 Action Recognition Challenge. In: THUMOS’15 Action Recognition Challenge. In Conjunction with CVPR’15 (2015)
26. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR arXiv: 1212.0402 (2012)
27. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: CVPR (2016)
28. Benaim, S., Ephrat, A., Lang, O., Mosseri, I., Freeman, W.T., Rubinstein, M., Irani, M., Dekel, T.: Speednet: learning the speediness in videos. In: CVPR (2020)
29. Feichtenhofer, C., Pinz, A., Wildes, R.P.: Spatiotemporal multiplier networks for video action recognition. In: CVPR (2017)
30. Wang, Y., Long, M., Wang, J., Yu, P.S.: Spatiotemporal pyramid network for video action recognition. In: CVPR (2017)
31. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L.V.: Temporal segment networks: towards good practices for deep action recognition. In: ECCV (2016)
32. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: ECCV (2018)
33. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV (2019)
34. Chéron, G., Laptev, I., Schmid, C.: P-CNN: Pose-based CNN features for action recognition. In: ICCV (2015)
35. Zolfaghari, M., Oliveira, G.L., Sedaghat, N., Brox, T.: Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection. In: ICCV (2017)
36. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: ICCV (2015)
37. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)
38. Tian, Y., Yan, Y., Zhai, G., Guo, G., Gao, Z.: EAN: event adaptive network for enhanced action recognition. Int. J. Comput. Vis. 130(10), 2453–2471 (2022)
39. Tian, Y., Che, Z., Bao, W., Zhai, G., Gao, Z.: Self-supervised motion representation via scattering local motion cues. In: ECCV (2020)
40. Tian, Y., Lu, G., Min, X., Che, Z., Zhai, G., Guo, G., Gao, Z.: Self-conditioned probabilistic learning of video rescaling. In: ICCV (2021)
41. Tian, Y., Yan, Y., Zhai, G., Chen, L., Gao, Z.: CLSA: a contrastive learning framework with selective aggregation for video rescaling. IEEE Trans. Image Process. 32, 1300–1314 (2023)
42. Seibold, C., Reiß, S., Sarfraz, M.S., Stiefelhagen, R., Kleesiek, J.: Breaking with fixed set pathology recognition through report-guided contrastive training. Med. Image Comput. Comput. Assist. Interv. 13435, 690–700 (2022)
43. Zhao, C., Zhan, L., Thompson, P.M., Huang, H.: Explainable contrastive multiview graph representation of brain, mind, and behavior. Med. Image Comput. Comput. Assist. Interv. 13431, 356–365 (2022)
44. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS, pp. 3104–3112 (2014)
45. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR, pp. 3156–3164 (2015)
46. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6450–6459 (2018)
47. Pandey, P., Prathosh, A.P., Kohli, M., Pritchard, J.: Guided weak supervision for action recognition with scarce data to assess skills of children with autism. Proc. AAAI Conf. Artif. Intell. 34, 463–470 (2020)
48. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.