

Autism Spectrum Disorder Screening on Home Videos Using Deep Learning

Anjali Singh

Department of Computer Science, Indira Gandhi Delhi Technical University for Women, Kashmere Gate, New Delhi, 110006, India

E-mail: anjalisinghgsm@gmail.com

ORCID iD: <https://orcid.org/0009-0003-7958-9945>

Abha Rawat

Department of Computer Science, Indira Gandhi Delhi Technical University for Women, Kashmere Gate, New Delhi, 110006, India

E-mail: abharawat8jan@gmail.com

ORCID iD: <https://orcid.org/0009-0003-3835-2327>

Mitali Laroia

Department of Computer Science, Indira Gandhi Delhi Technical University for Women, Kashmere Gate, New Delhi, 110006, India

E-mail: 04042001mitali@gmail.com

ORCID iD: <https://orcid.org/0009-0008-3953-6706>

Seeja K. R. *

Department of Computer Science, Indira Gandhi Delhi Technical University for Women, Kashmere Gate, New Delhi, 110006, India

Email: seeja@igdtuw.ac.in

ORCID iD: <https://orcid.org/0000-0001-6618-6758>

*Corresponding Author

Received: 27 October, 2023; Revised: 30 December, 2023; Accepted: 16 January, 2024; Published: 08 August, 2024

Abstract: Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by difficulty in social interactions, communication, and repetitive behaviors. Protocols like ADOS (Autism Diagnostic Observation Schedule) and ADI-R (Autism Diagnostic Interview Revised) are used by experts to assess the subject's behavior which is time-consuming. Over the decade, researchers have studied the application of various Machine Learning techniques for ASD diagnosis through facial feature analysis, eye movement tracking, questionnaire analysis, functional magnetic resonance imaging (fMRI) analysis, etc. However, these techniques are not helpful for the parent or guardian of the child to perform an initial screening. This research proposes a novel deep learning model to diagnose ASD using general videos of the subject performing some tasks with the parent/guardian. Since there is no publicly available dataset on ASD videos, a dataset is created by collecting the videos of autistic children performing some activities with parents/guardians from YouTube from different demographic locations. These videos are then converted to skeletal key points to extract the child's engagement and social interaction in a given task. The proposed CNN-LSTM model is trained on 80% of the collected videos and then tested on the remaining 20%. The experiment results on various combinations of pre-trained CNN models and LSTM/BiLSTM show that the proposed model can be used as an initial autism screening tool. Among the different combinations, the MobileNet and Bi-LSTM combo achieved the best test accuracy of 84.95% with 89% precision, recall and F1-score.

Index Terms: Autism Spectrum Disorder, Deep Learning, CNN, LSTM

1. Introduction

According to the National Institute of Mental Health (NIMH)[1], the prevalence of ASD in the United States is estimated to be 1 in 54 children. Diagnosis of ASD is complex, and no particular medical or genetic test is available for its diagnosis. Instead, it is diagnosed based on a comprehensive evaluation of a person's behaviour, communication, and developmental history. The diagnostic criteria for ASD are outlined in the Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5) [2], published by the American Psychiatric Association (APA). Other commonly used tools for the assessment of ASD include the ADOS [3] and the Childhood Autism Rating Scale (CARS)[4]. Early diagnosis and intervention are crucial for improving outcomes in individuals with ASD, making accurate diagnosis critical.

Traditional diagnostic methods rely on observing behavioral and developmental symptoms, which can be subjective and time-consuming. Recent advances in machine learning (ML) have provided new opportunities for improving the accuracy and efficiency of ASD diagnosis. ML algorithms like artificial neural networks, decision trees, and support vector machines are proposed to analyze large datasets of behavioural and biological markers to identify patterns indicative of ASD. All the existing ML-based ASD detection techniques are either cost and resource-intensive or prone to bias towards a particular race or color due to the dataset being based on children from one demographic location. To make the whole early detection process more feasible and user-friendly, an optimal approach to use the video recording of the subject doing some physical activity has been adopted in this research. The results obtained are more focused on general factors and behaviors that convey the subject being Autistic rather than being more influenced by any specific characteristics present in a particular race.

This research proposes a CNN-LSTM model for detecting autism from home videos of suspected children doing some tasks with the parent/guardian. The proposed model can be used as an initial screening tool and will be helpful for parents/guardians to decide whether further investigation is required.

The significant contributions of this research are

- A dataset has been created from publicly available YouTube videos and labelled manually based on the video's description.
- The videos in the dataset are then converted into skeletal ones, which reduces the computational complexity as well as protects the privacy of the children in the videos
- A CNN-LSTM model is proposed in which the CNN extracts the spatial features from each frame while the LSTM captures the temporal features from consecutive frames.
- The proposed model achieved a testing accuracy of 84.95% with 89% precision, recall and F1-score.

2. Literature Review

In recent years, various ML approaches on different modalities like eye tracking, facial images, medical signals like Electroencephalogram (EEG), video, etc, are being explored for ASD diagnosis. Mahmoudi et al. (2019)[5] developed an ML algorithm that analyzed EEG data to classify children with ASD and typically developing children. The results showed that the algorithm could accurately distinguish between the two groups, demonstrating the potential of EEG data for improving ASD diagnosis.

Another promising ML approach is the use of eye-tracking data, which is a reliable indicator of ASD. Zhao et al. [6] investigated eye-tracking data from conversations in person. For the dataset, 19 ASD and 20 TD children were interviewed with an eye tracker set up on their heads. These children showed no hyperactivity disorder or schizophrenia. The conversations obtained were divided into four sessions with some questions/conservation types to examine the behavioral gaze, nodding, initiating conversation, etc. The aim was to observe how children talk comfortably about things that interest them. It was observed that ASD individuals had a more extended interview duration than TD participants. Various ML classifiers were implemented on visual fixation features and interview duration as input. This approach must still be confirmed on a larger sample under varying disorder severity and balanced sex ratio. Guillon et al. (2014) [7] also used eye-tracking data to develop an ML algorithm that accurately classified children with ASD and typically developing children.

Xie et al.[8] has used a Two-stream End-to-End Deep Learning Network to study the visual attention of ASD and TD subjects. To create the dataset, 20 ASD and 19 TD Subjects were gathered based on IQ, race, age and gender. They were given images randomly from the OSIE database [9] for 3 sec each, and in this duration, a Human Fixation Map(HFM) representing their attention pattern on various pixels based on time was generated. In their model, VGGNet, which takes Natural Image and HFM as input, is used for object recognition separately. After which, the output for the two was concatenated and given to ASDNet for autism prediction. This study lacks clarity on the integration process of information from different parts of the image. Liaqat et al. [10] collected data on children freely viewing natural images (scan-path) for identifying autism using Machine Learning and Deep Learning. Scan paths consist of the fixation points' 2-D coordinates (where the subject's eyes have focused) and their duration generated by STAR-FC (Wloka et al. [11]), a

generative model of synthetic saccade patterns. This Image-based approach gave an accuracy of 55.31% and 61.62% on the test and validation datasets, respectively.

Abbas et al. [12] collected data from parent questionnaires and home videos for building the model. The video of the subject is recorded for a short duration in a natural environment. This video is then reviewed by an expert and filled with questionnaires based on the video. The output of the two models was combined to give the screening results. Both datasets are recorded in a highly clinical environment using AdI-R and ADOS protocols; the testing data is in a non-clinical environment. Also, the model gives an “Inconclusive” result in case it finds the input data challenging to avoid any wrong result like false-positive or false negative. Tariq et al. [13] used video raters who rated the video dataset for 30 different behavioural features like eye contact, passing of smile, etc., using 30 different questions. These raters were students or working professionals who didn't have any expertise in the autism domain. Then, they trained eight different ML models on these ratings. The top-performing classifier was logistic regression. With only five features, it showed an accuracy of 88.9%.

Hosseini et al. [14] have used facial feature extraction and image classification to detect ASD. Here, MobileNet was used, and it was observed that the classification results improved in terms of speed. Still, false positives, non-availability of metadata about images like their medical history, etc. and verification of images remain a hindrance. Khosla et al. [15] aimed to identify autistic children using facial expressions. This study considered a dataset from Kaggle containing 1468 images of each autistic and non-autistic group. The duplicate image was removed using the MD5 Hash Algorithm (Wang et al. [16]). Near-exact images are generated using the pHash Algorithm. The limitation of this study is the generalization Gap in validation and test accuracy. Lu and Perkowski [17] investigated ASD detection in minority groups by combining the images of minority groups with the Kaggle dataset.

Liao et al. [18] employed a small subset of ABIDE [19] data, which was collected by two universities, the California Institute of Technology and the University of Michigan, solely to accomplish the categorization by using the community structure of the brain. Herein, the study has calculated the correlation coefficients between two ROIs (Regions of interest) using Pearson correlation, and the community structure was discovered using the QCut detection method. Liu et al. [20] aim to address the problem of poor reproducibility and generalizability in neuroimaging investigations for possible biomarkers. They used ABIDE data to conduct an attentional connection analysis. They initially built separate brain networks for every person, then used the CC200 atlas to extract connection properties between ROIs.

Kojovic et al. [21] investigated the usability of nonverbal social interaction in ASD diagnosis. The dataset consists of videos of a child interacting with an adult, where the child was the test subject who could be ASD or TD. A distinct preprocessing used in this study was extracting skeletal key points with a black background for the human figures in the video. This reduced the dimensionality of the dataset. After this, CNN-LSTM architecture was deployed to recognise action, and VGG-16 was used for feature extraction. This study was able to reach a maximum accuracy of 80.9%. The proposed model is inspired by this study and aimed at the behavioural analysis of the subject for autism diagnosis.

3. Materials and Methods

3.1. CNN-LSTM

Convolutional Neural Network (CNN) [22] is a well-known machine learning model that uses convolution for pixel-related analysis. It works to break down the complex image data into a simpler form that still has all the essential data but is easier to process. Meanwhile, Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) that are suitable for sequence analysis problems [23]. RNNs are known for their ability to keep past information while processing the next step. This helps learn the current context, but when there is a time lag of 5-10 time steps, RNN fails to learn the context, resulting in the vanishing gradient problem [24]. LSTM was designed to address these issues [25] using an LSTM unit consisting of three activation functions known as gates [26], as they control the entire flow of information inside an LSTM unit. The Forget gate is responsible for choosing only the relevant information from the previous step, while the Input gate does the processing of current information and learns it. Lastly, the Output gate has the task of passing the updated information to the next timestamp. A useful and important variation of LSTM is the Bi-LSTM, consisting of two LSTMs (forward and backwards), increasing the amount of information available to the model. This leads to improvement in the context understanding of the model [27].

Combining these two architectures has helped to achieve a hybrid model that can be used for sequence prediction problems with spatial inputs, such as images or videos. The CNN in CNN-LSTM structure works upon spatial feature extraction of the input data, while LSTM helps to recognise sequential patterns. Thus, the CNN-LSTM combination creates a spatially and temporally deep model. These models have much potential, as much data is available and processed in visual formats.

3.2. Dataset

The dataset used in this study is self-curated. 128 Videos were collected from YouTube, 66 of which belong to the ASD class and 63 belong to Typically Developing (TD). The videos were collected considering the factors the ML model could use to make predictions. The selected videos capture the social interaction of the subject with a guardian when assigned a task. Since the data was sorted out from public platforms, the surroundings and assigned activity differ

from subject to subject. The videos are labelled as ASD and TD based on the description of the videos. The skeletal information was extracted from the clips using OpenPose to eliminate the variation in the surroundings. Figure 1 shows a snap from a video in the dataset and the corresponding skeletal structure created using OpenPose.

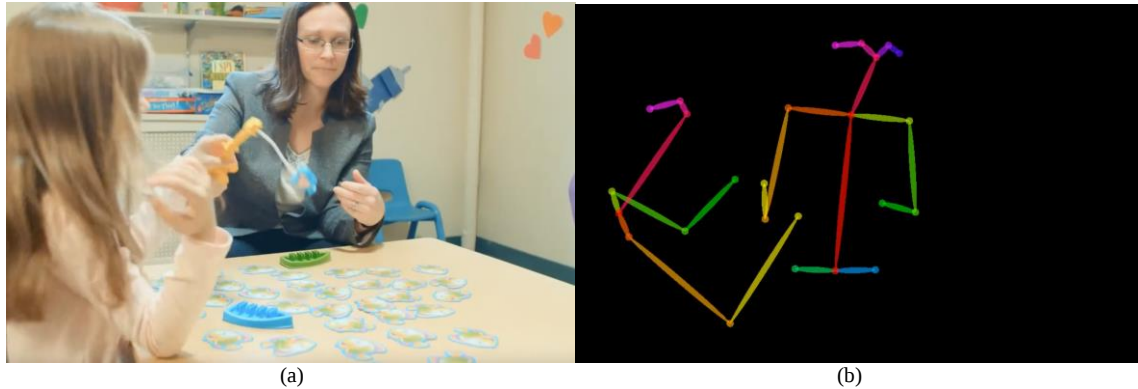


Fig. 1. (a) Snap from the video of a medical expert interacting with the subject through some games[28] (b) Snap from the video converted to skeletal structure using 2D OpenPose.

The videos consist of daily vlogs by parents/guardians performing some tasks with the subject or a medical expert interacting with children using some games for autism diagnosis or regular checks. Various tasks or activities include reading a picture book, painting, playing with a ball, soap bubbles, toys or games like building blocks, puzzles, etc. The goal of the activity is to engage the subject, study their attention towards tasks and objects around and social interaction with the guardian. Thus, the collected data captures the children's interest in objects and guardians. The dataset was preprocessed to remove unnecessary portions. Each video is further split into 5-second durations, and thus, 469 videos are created and labelled manually. The dataset is further divided into training and testing data in the ratio of 80:20. The dataset statistics are given in Table 1.

Table 1. Dataset Statistics

S No.	Criteria	ASD	TD	Total
1	No. of videos collected from YouTube	65	63	128
2	No. of videos after 5-second split	316	153	469
3	No of videos in the train dataset	254	121	375
4	No of videos in the test dataset	61	32	93

3.3. Proposed Methodology

In the proposed methodology, the videos containing ASD and TD children performing some activity with an adult are collected from YouTube. The relevant parts of each video are cropped out, and the skeletal information of the persons in the video is extracted using OpenPose. Since the deep learning model needs a large amount of data, each video is further split into 5-second video segments. The preprocessed videos were converted into frames using OpenCV [29] for feeding the data to the proposed CNN-LSTM model. The proposed methodology is outlined in Figure 2.

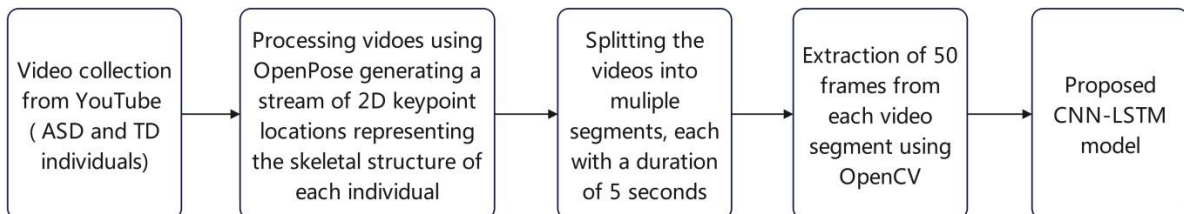


Fig. 2. Proposed methodology

3.4 Proposed CNN-LSTM Model

Due to a lack of benchmark video datasets for ASD diagnosis, the required videos were collected from YouTube, which was then passed through 2-D OpenPose technology to extract information on the skeletal structure of each individual. The continuous stream of skeletal key points was extracted from the video, capturing the gestures and interests of the individual in the task. The standard protocol for diagnosing autism [3, 30] mentions that the autistic

individual tends to show less interest in the surroundings and lack attention in the activities with a guardian or medical expert. Hence, feeding the acquired data to the CNN and LSTM captures the kinesics of the subject to classify autistic and non-autistic individuals. The pretrained CNN models like VGG16 [31], InceptionV3[32], Xception[33], MobileNet[34] are selected as the CNN. Preprocessed frames are fed to the pre-trained CNN model with “ImageNet” weights for spatial feature extraction, and hence, the fully connected layers at the top are not included. The output tensor is further reduced by adding one 2D convolution layer with 64 3x3 filters and then flattened to get the final feature vector. These feature vectors corresponding to each frame in the 5-second video are passed to LSTM to capture the sequential patterns in the video. Since the data is sequential and LSTM can learn long-term dependencies, it is suitable for video analysis. The proposed CNN-LSTM model is shown in Figure 3.

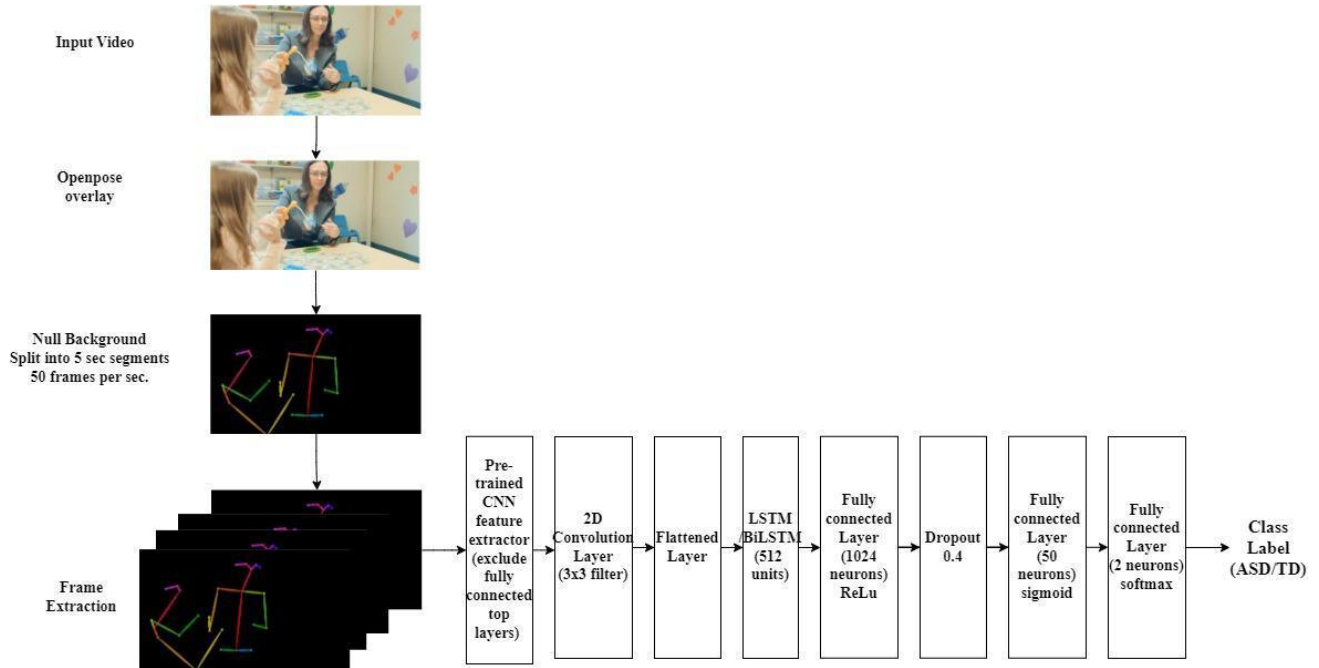


Fig. 3. Proposed CNN-LSTM model

4. Implementation

The proposed model uses several computer vision and deep learning libraries in Google Colab. The OpenCV library reads the input video file and extracts frames from it. Keras, a high-level deep learning library, is used to load the pre-trained CNN models and define the LSTM and BiLSTM model architectures. NumPy is used for array manipulation and storage of feature vectors. The h5py library efficiently stores feature vectors in .h5 file format. The TensorFlow library is used as the backend for Keras.

For dataset creation, YouTube videos were collected that showcase a subject performing an activity with a guardian. To capture this data, the relevant parts of each video were cropped, and skeletal information of the activity was extracted using OpenPose [35], a deep learning-based 2D pose estimator that can identify multiple people. The videos were split into 5-second segments using FFMPEG [36]. The preprocessed videos were converted into frames using OpenCV's VideoCapture() function. The frames are extracted in a loop until there are no more frames to read, and each frame is saved as an image using the imwrite() function. To avoid unnecessarily extracting too many frames and increasing the dataset size, the frames are extracted at a fixed interval of every ten frames using the modulo operator. The extracted frames are also preprocessed by resizing them to a fixed size of 224X224 pixels (for VGG16 & MobileNet) or 299X299 pixels (for Xception and InceptionV3) using the cv2.resize() function before being saved as individual images.

The proposed model uses pre-trained CNNs for feature extraction from the input video frames. The pre-trained CNN models selected are VGG16, InceptionV3, Xception and MobileNet with “ImageNet” weights, excluding the top fully connected layers. These pre-trained CNN models are loaded using the Keras library. The dimensions of output tensors of the CNN models are further reduced by adding a 2D convolution layer of 64 filters of size 3 X 3. The dimensions of the output tensor and the feature vector are shown in Table 2.

Table 2. Feature Vector Dimension

Sl.No	Pre-Trained CNN Model	Output Tensor Dimension	Tensor Dimension after reduction using 2D Conv.	Feature Vector Dimension
1.	VGG16	7 X 7 X 1024	5 X 5 X 64	1600
2.	MobileNet	7 X 7 X 1024	5 X 5 X 64	1600
3.	Xception	10 X 10 X 2048	8 X 8 X 64	4096
4.	InceptionV3	8 X 8 X 2048	6 X 6 X 64	2304

The resulting feature vectors are saved as NumPy arrays in two .h5 files - training and testing data. The training and testing dataset sizes are given in Table 1. The .h5 file format is a hierarchical data format that supports efficient storage of large multidimensional datasets. To store the training and testing data in .h5 files, the h5py library is used.

The feature vectors corresponding to each frame of a 5-second video are grouped and fed into the LSTM or BiLSTM layer with 512 units for learning the temporal patterns in the input video frames. The LSTM layer is followed by two fully connected layers (1024 neurons and 50 neurons) with ReLu and sigmoid activation, respectively. A dropout Layer with a 0.5 dropout rate is added between the two dense layers. Lastly, the output classification layer with two neurons and Softmax activation is added for classification. The model is compiled using mean_squared_error loss and Adam optimizer with a learning rate 0.001. The model is trained for 100 epochs with a batch size of 50. The model is evaluated on various combinations of pre-trained CNN models and LSTM or BiLSTM. All the parameters are kept the same for both LSTM and BiLSTM layers.

5. Results and Discussion

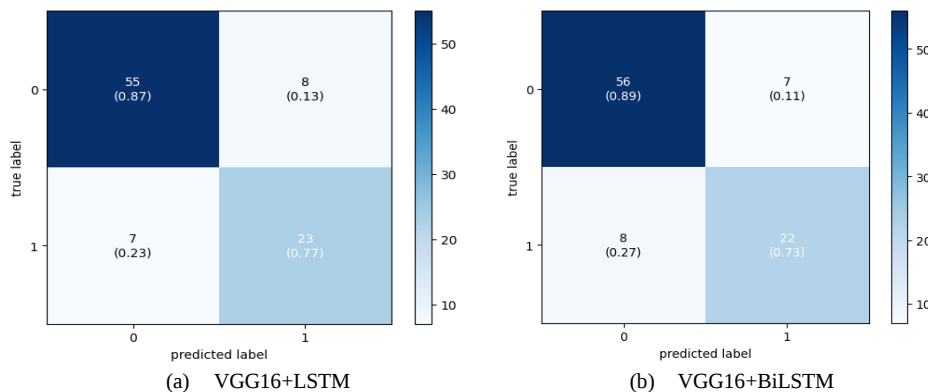
The proposed model is evaluated by experimenting with different combinations of pre-trained CNN models and LSTM or BiLSTM. The various metrics used for the performance evaluation of the models are Accuracy, precision, recall and F1 Score. The performance of the models on these metrics is shown in Table 3.

Table 3. Model Performance

Sl.No	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
1.	VGG16 + LSTM	83.87	84	84	84
2.	MobileNet + LSTM	83.87	84	84	84
3.	InceptionV3 + LSTM	68.82	69	69	69
4.	Xception + LSTM	81.72	82	82	82
5.	VGG16 + BiLSTM	83.87	84	84	84
6.	MobileNet + BiLSTM	84.95	85	85	85
7.	InceptionV3 + BiLSTM	67.74	68	68	68
8.	Xception + BiLSTM	81.72	82	82	82

The best accuracy is obtained with MobileNet and the BiLSTM model. VGG16 and Xception have given the same accuracy with LSTM and BiLSTM. Almost similar results have been obtained on LSTM and BiLSTM throughout the models, as shown in Table 3. InceptionV3 has demonstrated the least accuracy among all the models.

Since the final dataset is imbalanced due to the short video length of the TD class, to check whether the model is biased towards the majority class, a confusion matrix has been drawn and is shown in Figure 5. The confusion matrix demonstrates that the model is not biased towards the majority class. False Negatives obtained on Xception and InceptionV3 are more than the other two models. This means that these models will result in more diagnoses claiming that autistic subjects are ASD-negative and, hence, unsuitable for real-world deployment. MobileNet and VGG16 are giving remarkable performance with both LSTM and BiLSTM.



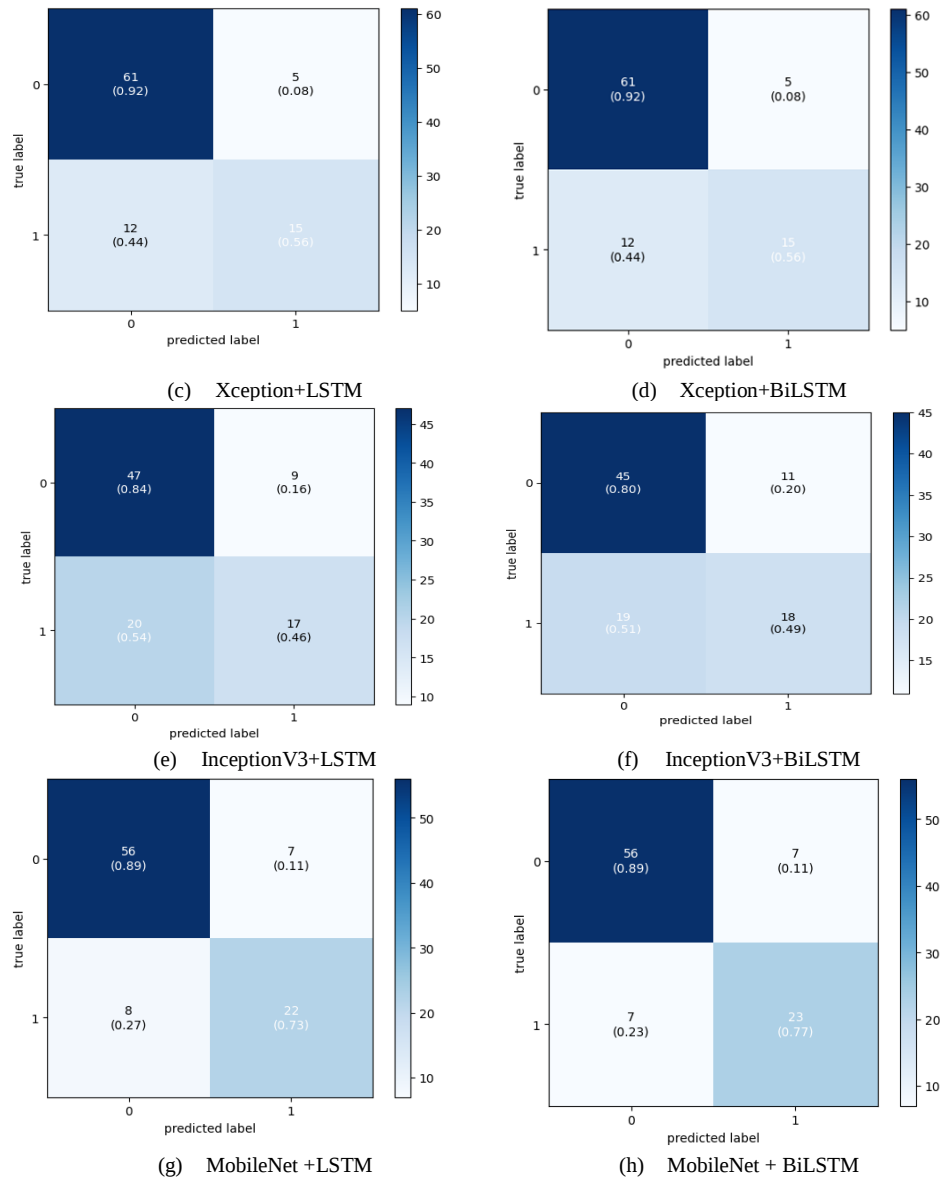


Fig. 4. Confusion Matrix

5.1 Comparison with state-of-the-art models

There is not much research on video analysis for ASD detection. The only research found in the literature is the work done by Kojovic et al. [21], which is similar to this study. However, the video dataset for their model was prepared in a controlled environment for hours. The expert performed activities with the subject based on ADOS assessment. The task given to each participant was cohesively designed to capture the focus and social interaction of the subject. Moreover, the dataset is created from children from a particular demographic region. In the proposed study, the data is collected from public sources, which are not structured and contain videos created on children from different demographics worldwide. Thus, the dataset used in this proposed study is more challenging, and the model trained on this dataset is expected to be more robust. The proposed model's performance is compared with the model proposed by Kojovic et al. [21] and is shown in Table 4.

Table 4. Comparison with state-of-the-art

Sl.No	Reference	Dataset			Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
		Method of Dataset Creation	Total duration of all the videos	Number of 5-sec segments					
1	Kojovic et al., [21]	Created in a Controlled environment	47.782 hours	34340	VGG16 + LSTM	80.9	78.4	85.4	81.8
2	Proposed Methodology	Created from YouTube videos	30 minutes	469	Mobile Net+BiLSTM	84.95	85	85	85

6. Conclusion

This paper proposes a CNN-LSTM model that is efficient and robust for ASD diagnosis using simple Home videos. The dataset is created by collecting YouTube videos and contains noise and inconsistency. The acquired dataset aimed to capture the social interaction between the subject and the guardian. Since the model is trained on a small and noisy dataset, it is robust towards the subject, environment, and activity performed. The videos are first converted to skeletal key points to reduce the data complexity and protect the privacy of the subjects. VGG16, InceptionV3, Xception and MobileNet are used for obtaining features, and their output is then passed into the LSTM/BiLSTM. LSTM and BiLSTM have shown the same performance throughout the models. MobileNet with BiLSTM gave the best accuracy of 84.95%. Studying the confusion matrix shows that InceptionV3 and Xception are unsuitable for real-world application as the False negatives obtained on them are high and may lead to undesirable misdiagnosis. However, the results show the potential of video-based assessment in ASD prediction. A software tool based on the proposed model may help parents perform an initial ASD screening cost-effectively.

Conflict of Interest

The authors declare no conflict of interest.

References

- [1] National Institute of Mental Health (NIMH), Autism Spectrum Disorder, <https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd>
- [2] DSM Library, Diagnostic and Statistical Manual of Mental Disorders, <https://www.psychiatry.org/psychiatrists/practice/dsm>
- [3] Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. L. (2012). Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) manual (Part I): Modules 1-4. Western Psychological Services.
- [4] Schopler, E., Reichler, R. J., & Renner, B. R. (1988). Childhood Autism Rating Scale (CARS). Western Psychological Services.
- [5] Mahmoudi, M., Ahmadi, M., Jahani, S., Aarabi, A., & Khaleghi, A. (2019). EEG-based classification of autism using dynamic functional connectivity: a deep learning approach. *Journal of neural engineering*, 16(5), 056026.
- [6] Zhao, Z., Tang, H., Zhang, X., Qu, X., Hu, X., & Lu, J. (2021). Classification of children with autism and typical development using eye-tracking data from face-to-face conversations: Machine learning model development and performance evaluation. *Journal of Medical Internet Research*, 23(8), e29328. <https://doi.org/10.2196/29328>
- [7] Guillon, Q., Hadjikhani, N., Baduel, S., & Rogé B. (2014). Visual social attention in autism spectrum disorder: insights from eye tracking studies. *Neuroscience & Biobehavioral Reviews*, 42, 279-297.
- [8] Xie, J., Wang, L., Webster, P., Yao, Y., Sun, J., Wang, S., & Zhou, H. (2019). A two-stream end-to-end deep learning network for recognizing atypical visual attention in autism spectrum disorder. *arXiv preprint arXiv:1911.11393*. <https://doi.org/10.48550/arXiv.1911.11393>
- [9] Mit Saliency Benchmark, Springer International Publishing. <http://saliency.mit.edu/datasets.html>, Accessed 26 August 2022
- [10] Wu, C., Liaqat, S., Cheung, S. C., Chuah, C. N., & Ozonoff, S. (2019, July). Predicting autism diagnosis using image with fixations and synthetic saccade patterns. In 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) (pp. 647-650). IEEE. <https://doi.org/10.1109/ICMEW.2019.00125>
- [11] Wloka, C., Kotseruba, I., & Tsotsos, J. K. (2017). Saccade sequence prediction: Beyond static saliency maps. *arXiv preprint arXiv:1711.10959*. <https://doi.org/10.48550/arXiv.1711.10959>
- [12] Halim Abbas, Ford Garberson, Eric Glover, Dennis P Wall, Machine learning approach for early detection of autism by combining questionnaire and home video screening, *Journal of the American Medical Informatics Association*, Volume 25, Issue 8, August 2018, Pages 1000–1007, <https://doi.org/10.1093/jamia/ocy039>
- [13] Tariq, Q., Daniels, J., Schwartz, J. N., Washington, P., Kalantarian, H., & Wall, D. P. (2018). Mobile detection of autism through machine learning on home video: A development and prospective validation study. *PLoS medicine*, 15(11), e1002705.
- [14] Hosseini, M. P., Beary, M., Hadsell, A., Messersmith, R., & Soltanian-Zadeh, H. (2021). Deep Learning for Autism Diagnosis and Facial Analysis in Children. *Frontiers in Computational Neuroscience*, 15. <https://doi.org/10.3389/fncom.2021.789998>
- [15] Y. Khosla, P. Ramachandra and N. Chaitra. (2021) . Detection of autistic individuals using facial images and deep learning. *IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, 2021, pp. 1-5, <https://doi.org/10.1109/CSITSS54238.2021.9683205>
- [16] Wang, X., & Yu, H. (2005, May). How to break MD5 and other hash functions. In *Annual international conference on the theory and applications of cryptographic techniques* (pp. 19-35). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11426639_2
- [17] Lu, A., & Perkowski, M. (2021). Deep Learning Approach for Screening Autism Spectrum Disorder in Children with Facial Images and Analysis of Ethnoracial Factors in Model Development and Application. *Brain Sciences*, 11(11), 1446.
- [18] Liao, D., & Lu, H. (2018, March). Classify autism and control based on deep learning and community structure on resting-state fMRI. In 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI) (pp. 289-294). IEEE. <https://doi.org/10.1109/ICACI.2018.8377471>
- [19] Welcome to the Autism Brain Imaging Data Exchange!, Child Mind Institute, https://fcon_1000.projects.nitrc.org/indi/abide/. Accessed 26 August 2022
- [20] Liu, Y., Xu, L., Li, J., Yu, J., & Yu, X. (2020). Attentional connectivity-based prediction of autism using heterogeneous rs-fMRI data from CC200 atlas. *Experimental Neurobiology*, 29(1), 27. <https://doi.org/10.5607/en.2020.29.1.27>
- [21] Kojovic, N., Natraj, S., Mohanty, S.P. *et al.* Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children. *Sci Rep* 11, 15069 (2021). <https://doi.org/10.1038/s41598-021-94378-z>

- [22] Analytics Vidya, Introduction to Convolution Neural Network(CNN), <https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/>
- [23] Machine Learning Mastery, A Gentle Introduction to Long-Short Term Memory Networks by the Experts, <https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/>
- [24] Felix A. Gers, Jürgen Schmidhuber, Fred Cummins; Learning to Forget: Continual Prediction with LSTM. *Neural Comput* 2000; 12 (10): 2451–2471. doi: <https://doi.org/10.1162/089976600300015015>
- [25] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke and J. Schmidhuber, "A Novel Connectionist System for Unconstrained Handwriting Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855-868, May 2009, doi: 10.1109/TPAMI.2008.137.
- [26] Analytics Vidya, Learn About Long-Short Term Memory (LSTM) Algorithms, <https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/>
- [27] Paperswithcode, Bidirectional LSTM, <https://paperswithcode.com/method/bilstm>
- [28] Sample YouTube Video: <https://www.youtube.com/watch?v=pSGVb60-BSw> Accessed 26 January 2023
- [29] Pulli, K., Baksheev, A., Korniyakov, K., & Eruhimov, V. (2012). Real-time computer vision with OpenCV. *Communications of the ACM*, 55(6), 61-69. <https://doi.org/10.1145/2184319.2184337>
- [30] Western Psychological Services. Schopler, E., Reichler, R. J., & Renner, B. R. (1988). Childhood Autism Rating Scale (CARS). Western Psychological Services.
- [31] Tammina, S. (2019). Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *International Journal of Scientific and Research Publications (IJSRP)*, 9(10), 143-150. <http://dx.doi.org/10.29322/IJSRP.9.10.2019.p9420>
- [32] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826). <https://doi.org/10.48550/arXiv.1512.00567>
- [33] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258). <https://doi.org/10.48550/arXiv.1610.02357>
- [34] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*. <https://doi.org/10.48550/arXiv.1704.04861>
- [35] Viso.ai, The Complete Guide to OpenPose in 2023 , <https://viso.ai/deep-learning/openpose/#:~:text=OpenPose%20is%20a%20real%2Dtime,a%20total%20of%20135%20keypoints>
- [36] FFmpeg, ffmpeg Documentation, <https://ffmpeg.org/ffmpeg.html>

Authors' Profiles



Anjali Singh is a B.Tech Computer Science & Engineering graduate from Indira Gandhi Delhi Technical University for Women in 2023. She currently serves as a Software Development Engineer I at JP Morgan Chase & Co, specializing in backend development. Her research interest include machine learning, deep learning and health informatics.



Abha Rawat is currently a Software Development Engineer I at JP Morgan Chase & Co. She graduated from Indira Gandhi Delhi Technical University for Women with B.Tech Computer Science & Engineering in 2023. Her research interest include machine learning, deep learning and health informatics.



Mitali Laroia graduated in 2023 with a B.Tech degree in Computer Science & Engineering from Indira Gandhi Delhi Technical University for Women. She has experience in the domain of backend engineering and DevOps. She is presently employed as Software Development Engineer I at JP Morgan Chase & Co. Her research interest include machine learning, deep learning and health informatics.



Seeja K. R. is Professor in the Department of Computer Science and Engineering, Indira Gandhi Delhi Technical University for Women (IGDTUW), Delhi. She received Ph.D. degree in Computer Science from Jamia Hamdard, New Delhi, India. Her research interest includes Data Mining, Intelligent Computing, Biometrics, Health Informatics and Algorithm Design. She has a teaching experience of 25 years and has published many research papers in various international journals and conference proceedings.

How to cite this paper: Anjali Singh, Abha Rawat, Mitali Laroia, Seeja K. R. , "Autism Spectrum Disorder Screening on Home Videos Using Deep Learning", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.16, No.4, pp. 106-115, 2024. DOI:10.5815/ijigsp.2024.04.08