Vision-Based Early Autism Detection Using Repetitive Behavior Analysis

Vaishnav Raj
Department of Computing Technologies
SRM Institute of Science and Technology
SRM Nagar, Kattankulathur, Tamil Nadu 603203, India
vr8324@srmist.edu.in

Aashish Kumar
Department of Computing Technologies
SRM Institute of Science and Technology
SRM Nagar, Kattankulathur, Tamil Nadu 603203, India
ak6916@srmist.edu.in

Dr.R.Jeya,
Associate Professor
Department of Computing Technologies,
Faculty of Engineering and Technology,
SRM Nagar, Kattankulathur, Tamil Nadu 603203, India
jeyar@srmist.edu.in

***Abstract*— Early identity of Autism Spectrum Disorder (ASD) is vital as it enables timely behavioral and educational intervention at some stage in the maximum critical developmental ranges. but conventional autism screening strategies are often subjective, time-extensive, and depending on skilled clinicians. This paper offers an imaginative and prescient-based totally deep analyzing framework for early autism screening thru studying repetitive motion patterns from brief motion pictures. The proposed technique makes a distinctiveness of typically located repetitive behaviors which encompass hand flapping, frame rocking, and spinning. A - degree pipeline is designed: first, a actual-time object detector is used to localize the child and decrease history interference; second, a spatiotemporal transformer version learns motion dynamics to classes repetitive movement styles. The shape combines YOLO-primarily based completely detection with VideoMAE-based totally temporal example getting to know, allowing strong motion facts even beneath partial occlusion and noisy backgrounds. Experimental evaluation on repetitive gesture samples demonstrates that the proposed method achieves excessive class performance and gives a normal and scalable possibility for early screening. The device is software program-based totally definitely and can be deployed on low-charge gadgets, making it appropriate for useful resource-limited environments.**

Keywords— autism spectrum disorder, repetitive behavior, YOLO, VideoMAE, computer vision, deep learning, early screening.

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition generally characterized with the aid of challenges in social communication and the presence of constrained or repetitive behaviours. a main mission in ASD care is that medical analysis frequently takes place past due, in spite of early signs and symptoms being seen in adolescence. delayed prognosis reduces the effect of early intervention, which is widely known to enhance lengthy-time period cognitive, emotional, and social consequences.

In cutting-edge practice, autism screening is in general accomplished via dependent behavioural statement, interviews, and developmental exams. whilst those methods are clinically dependable, they require trained experts, repeated sessions, and widespread time. additionally, results can range due to observer subjectivity and environmental elements those obstacles encourage the development of automated, objective screening systems which could manual clinicians and decorate accessibility.

Most of the earliest seen behavioural markers of ASD are repetitive motor styles consisting of hand flapping, rocking, and spinning. these behaviours are observable using trendy video recordings, which makes computer vision-based screening a sensible course. With current advances in deep learning, video know-how fashions can research complicated movement representations, making it possible to stumble on repetitive gestures without wearable sensors.

This paper proposes an early autism screening framework based on repetitive movement analysis the use of brief movies. The proposed method makes use of video as a low-price and non-intrusive input modality, which may be captured in domestic environments or medical settings with minimum setup by that specialize in repetitive movement cues, the framework targets a key early behavioural indicator that is frequently seen in young youngsters, and which may be measured without direct interaction or specialised clinical equipment.

One extra important point is that repetitive gestures in ASD are not continually consistent, and their intensity can range for the duration of the day relying on mood, environment, or stimulation level. due to this, screening structures must be able to operating with brief, imperfect, and real-world video clips rather than best laboratory recordings. in sensible situations, motion pictures may additionally include historical past clutter, one of a kind digital camera angles, and partial occlusion, which makes the problem more difficult however also extra significant for actual deployment.

The overall goal of the proposed system is to help early-stage screening through automatic recognition of repetitive gestures from short video segments. Such an technique can reduce reliance on prolonged announcement durations, reduce screening time, and offer greater goal outputs that can help clinicians. further, a video-primarily based totally pipeline is greater scalable for huge-scale deployment, specially in regions where expert clinicians and specialised diagnostic offerings are restricted.

The precept contributions of this observation are:

1. A stage imaginative and prescient pipeline that mixes infant detection and movement-based totally gesture magnificence.

2. A transformer-based totally temporal version that captures repetitive movement continuity.

3. A deployed form appropriate for the lowest price gadgets and scalable screening.

II. LITERATURE REVIEW

Early autism detection has been broadly explored using medical equipment, sensor-primarily based tracking, and laptop imaginative and prescient approaches [3], conventional diagnostic frameworks along with dependent remark and behavioural scoring continue to be the medical gold general, however they require expert involvement and are tough to scale in lots of areas [1].

Current studies have looked at how device learning can spot autism-related patterns by way of monitoring behaviour and physiological signals. Researchers have used wearable sensors like accelerometers and gyroscopes to measure movement and pick out up on repetitive actions. however, whilst those sensors do a very good job recording movement, they tend to be intrusive for youngsters. They need the kid to truely put on the tool and comply with commands, which isn't continually smooth. Plus, the use of these systems for huge-scale screening simply doesn't appear sensible. [3], [13].

Laptop imaginative and prescient has emerged as a non-contact alternative [13], [14]. earlier vision-primarily based techniques trusted hand-made features which include optical waft, skeletal joint tracking, or motion electricity pics. even as those procedures can perceive movement modifications, they regularly fail below real-international conditions including terrible lighting fixtures, historical past clutter, and occlusion [12], [13].

Current deep getting to know techniques enhance this by way of mastering features directly from information [9]–[11]. object detection networks including YOLO have enabled accurate character localization in actual time, permitting systems to isolate the challenge and reduce irrelevant history noise [5], [6]. For temporal knowledge, 3-d convolutional networks and transformer-primarily based architectures have shown strong performance in motion reputation tasks [8], [11]. lately, masked autoencoder models for video (inclusive of VideoMAE) have demonstrated the capability to study motion continuity and temporal shape by reconstructing missing video regions that is mainly beneficial in repetitive conduct popularity, in which the continuity of motion is a key cue [7].

Notwithstanding those advances, an opening stays in combining green toddler localization with robust temporal modelling for repetitive gesture detection in autism screening [12], [13]. Many present works either focus only on detection or handiest on category, without designing a unified quit-to-cease architecture suitable for realistic deployment [12], [13].

Video-based totally autism screening isn't as easy because it sounds. You run into all kinds of troubles exclusive digital camera angles, kids transferring in unpredictable ways, age-related differences, and all the random stuff taking place inside the background. in case you don't layout the machine cautiously, overall performance can certainly tank. On top of that, things like hand flapping or rocking frequently appearance a lot like normal play, which means that the gadget would possibly flag usual conduct as a trouble. To keep away from that, you actually need stable temporal modeling and correct tracking. That's the way you make certain the patterns you select up really display repetitive conduct, now not just ordinary kid interest [7], [11].

III. OVERALL ARCHITECTURE

The proposed gadget is designed as a level pipeline that converts uncooked video input into a repetitive movement prediction

A. System Pipeline

The shape is fabricated from the subsequent essential modules:

1. Video Acquisition

- The video can be recorded in natural settings like homes, clinics, or schools, with only a few policies.
- The technique doesn't want any wearable sensors or precise scientific machine, so it's smooth to apply in actual lifestyles.
- The input period is saved short whilst you do not forget that repetitive behaviors normally display up within just a few seconds of recording.

2. Frame Extraction and Preprocessing

- Motion images are resized, normalized, and damaged down into frames at a difficult and rapid frame rate.
- This step adjustments the uncooked video right proper into a modern format so the deep studying modules can address it without problems.
- Resizing cuts down at the computing strength preferred and enables the manner art work properly on less luxurious devices.
- Normalization helps reduce down on versions due to lights and camera see in movies recorded at domestic.

3. YOLO-based Child Localization

- A YOLO model spots the child in every frame and draws bounding packing containers round them.
- The detection module is vital due to the truth that hundreds of films have things like history motion, furniture, or particular human beings in them.
- By using using specializing in the baby location, the pipeline stays clear of choosing up beside the point info from the encompassing environment.
- YOLO changed into decided on because it runs in actual time and works properly on regular hardware with no need anything particular.

4. Cropping the region and Dimming the information

- The detected location is cropped to reduce down on litter and maintain the point of interest on the movement.
- This step makes movement assessment higher with the useful resource of clearing out historic past noise and slicing down on fake styles.
- Cropping makes certain the temporal model will pay interest through the kid's frame and hand movements.

5. Temporal movement reading VideoMAE version

- Temporal motion analyzing the use of the VideoMAE model captures how subjects circulate through the years in films. It allows the machine recognize the drift of motion thru analysing sequences, making it higher at responsibilities concerning movies.
- The cropped video section is sent to a VideoMAE version to research the manner movement works.
- VideoMAE learns each region and time skills and is able to capture repeated movement styles, which frequently show up in ASD repetitive behaviors.

6. Repetitive Gesture Class

- a totally final first-classifier makes a decision if the motion suits one of the following:
- Hand flapping is a repetitive movement wherein a person brief shakes their hands or palms. it's frequently visible in children, especially those with autism or sensory processing versions.
- A smooth interest you don't do time and again all over again every day
- Which encompass a class that isn't repetitive is important to prevent fake alarms because of the truth youngsters playing generally can also waft fast.

The very last output offers a screening-focused prediction that can assist clinicians with early-degree assessment.

B. Architecture Diagram

Fig. 1. A visual representation of some of the repetitive behaviors showed by autistic individuals

Fig. 1. shows three not unusual varieties of repetitive behaviors often found in youngsters: spinning, wherein the child rotates repeatedly in circles; shaking, wherein the kid repeatedly moves the body (or components of the frame) in a rhythmic manner, normally even as sitting or status; and obsessive behavior, wherein the child becomes overly centered on a single object or action (like again and again touching or gazing something).

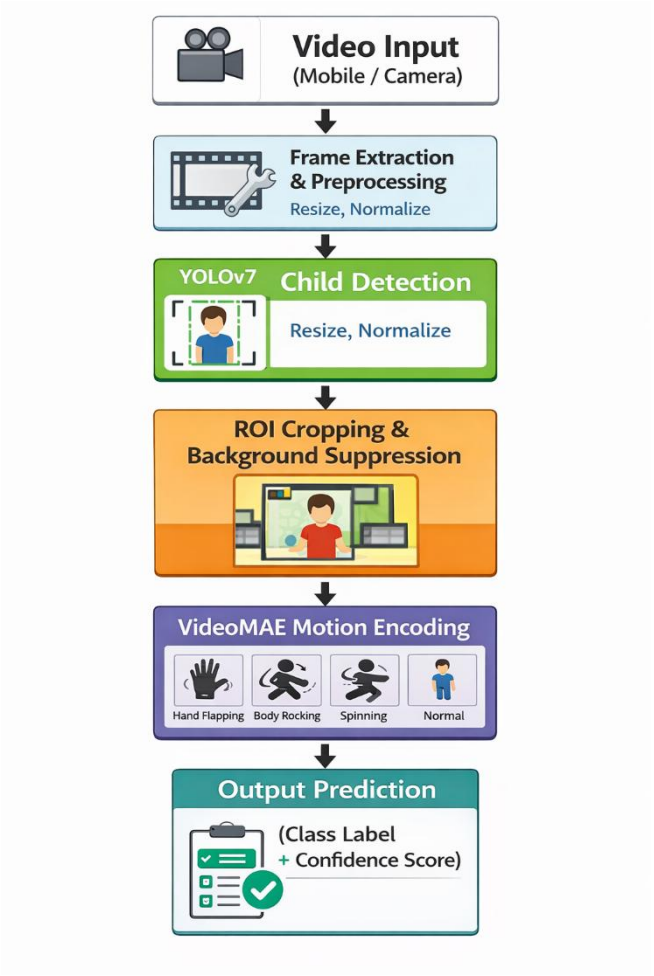

Fig. 2. Overall architecture of the proposed early autism detection system using repetitive movement.

Fig. 2. suggests the entire pipeline (step-with the aid of-step workflow) for detecting and classifying a toddler's repetitive actions from a video. Then, the device plays body Extraction & Preprocessing, which means the video is damaged into multiple frames (images), and every body is cleaned and standardized the usage of operations like resizing and normalization. This facilitates the version handle movies of various excellent, lighting fixtures, and backbone in a regular manner.

After preprocessing, the pipeline uses YOLOv7 for child Detection, which mechanically finds the child in every frame via drawing a bounding container round them. subsequent comes ROI Cropping & background Suppression, wherein only the crucial region (the kid) is cropped out and needless historical past info are reduced. that is critical due to the fact the historical past can confuse the model and reduce accuracy.

Then the cropped video segment is handed to VideoMAE motion Encoding, which learns the child's motion patterns across frames and converts them into meaningful movement features. eventually, the machine offers an Output Prediction, which incorporates the anticipated class label (example: hand flapping, body rocking, spinning, or regular) in conjunction with a confidence rating showing how certain the version is about its prediction.

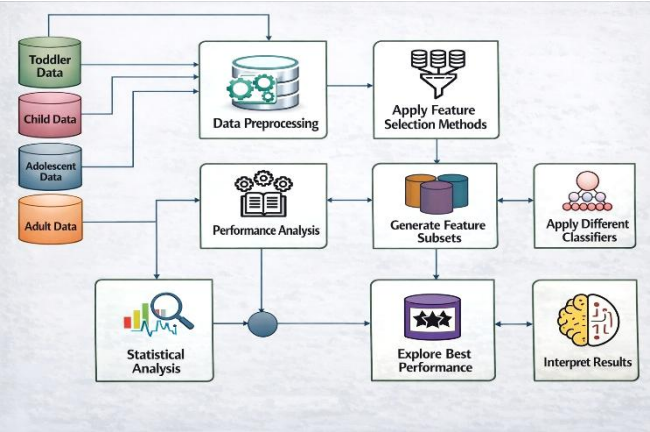

Fig. 3. Efficient Machine Learning Models for Early Stage Detection of Autism Spectrum Disorders

Fig.3. shows a machine studying workflow for studying datasets collected from exclusive age groups: little one statistics, baby facts, Adolescent records, and person records. these types of datasets are first combined and despatched into statistics Preprocessing, in which the uncooked information is cleaned, formatted, and organized (for example: doing away with missing values, normalizing values, and converting the facts into a usable form).

After preprocessing, the pipeline applies feature choice techniques, which means it selects only the most crucial features (columns/variables) from the dataset in order that the model specializes in meaningful patterns in place of needless or noisy information. subsequent, the system generates function subsets, which means it creates distinct combos of selected features to test which set gives the exceptional overall performance.

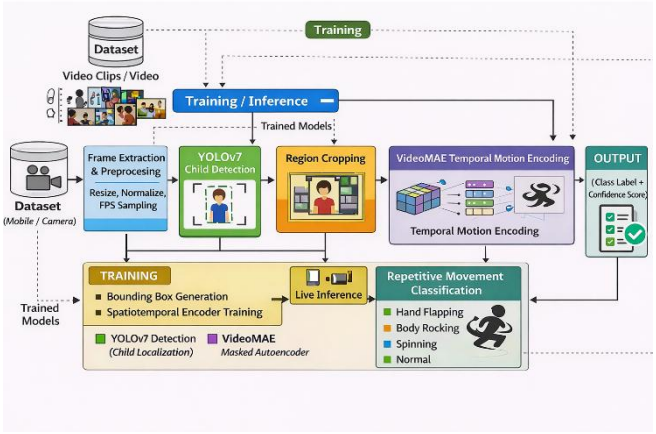

Fig. 4. Detailed architecture of the proposed system, consisting of preprocessing, child detection, temporal motion encoding and repetitive movement classification.

Fig. 4. shows a machine learning workflow for analyzing datasets collected from different age groups: Toddler Data, Child Data, Adolescent Data, and Adult Data. All these datasets are first combined and sent into Data Preprocessing, where the raw data is cleaned, formatted, and prepared (for example: removing missing values, normalizing values, and converting the data into a usable form).

The pipeline is modular, that means every block has a clean position and may be progressed independently. for example, YOLOv7 focuses simplest on in which the kid is, while VideoMAE focuses on how the kid is moving over the years. In actual-international usage, this layout helps the version paintings higher in herbal environments consisting of homes, faculties, or clinics, where lighting fixtures and heritage situations are not managed.

For Example, if the version predicts "Hand Flapping – 0.92", it way the detected movement sample suits hand flapping with excessive fact. however if the self belief is low (like zero.55), it is able to imply uncertain motion, partial visibility, or combined behaviors.

IV. METHODOLOGY AND SYSTEM WORKFLOW

To certainly summarize the entire workflow of the proposed early autism detection framework, table I presents the module-clever breakdown of the device. The desk highlights the principle processing levels, the deep gaining knowledge of fashions utilized in every stage, the input/output flow, and the purpose of every block. This based view enables in expertise how the pipeline transitions from uncooked video enter to the very last repetitive motion prediction, at the same time as making sure that heritage distractions are minimized and temporal movement styles are effectively captured.

A. Data and Tables

This table offers a clear module-wise precis of the whole device workflow. It suggests each processing level, the technique/model used, the enter furnished, and the output produced at each step and it helps us recognize how the system converts a uncooked video clip right into a very last prediction of repetitive movement, making sure smooth processing, correct classification, and significant output.

S. No	SYSTEM MODULES			
	Module	***Technique***	***Input***	***Input***
1	Video Acquisition	Mobile Or Camera Recording	Raw clip	Video stream
2	Frame Extraction	Fixed FPS sampling	Video stream	Frames
3	Preprocessing	Resize, Normalize	Frames	Enhanced frames
4	Child Detection	YOLOv7	Enhanced frames	Child bounding box
5	ROI Cropping	Region extraction	Frames + bounding box	Cropped child frames
6	Motion Encoding	VideoMAE	Cropped frame sequence	Temporal motion features
7	Classification	Softmax / MLP head	Motion features	Gesture class
8	Output	Confidence scoring	Predicted class	Label + probability

Table. I. MODULE-WISE DESCRIPTION OF THE SYSTEM

As shown in Table I, the proposed approach combines object detection and temporal motion encoding to gain correct and dependable repetitive movement category from short videos. First, the system extracts frames from the raw video and applies preprocessing steps consisting of resizing and normalization to improve visible consistency. Then, YOLOv7 is used to stumble on and localize the kid in each body, permitting the version to cognizance simplest at the relevant issue rather than background distractions.

These extracted features are then fed right into a classifier (inclusive of a softmax or MLP head) to expect the repetitive movement class.

RESULTS AND DISCUSSION

This system was tested using video samples with repetitive actions like hand flapping, body rocking, and spinning. The model had a high classification rate for detecting repetitive actions based on temporal motion patterns from video clips.

The combination of YOLO-based child localization and VideoMAE-based temporal motion encoding enhanced the accuracy of classification by minimizing the impact of background elements and considering only motions of interest. In addition, a two-step procedure helped in minimizing interference from irrelevant environmental factors like lighting, clutter, and camera motion [5], [7].

The classification accuracy of the system is quite high (90+%), demonstrating that spatiotemporal learning can detect repetitive behavioral patterns efficiently. Masked autoencoder architecture of VideoMAE allowed the model to recognize the continuity of motions despite partial occlusions and missing segments in the video [7].

Nevertheless, several issues should be addressed:

- Normal child activities like walking could be mistakenly classified as repetitive behavior.
- Predictions with low confidence levels happened if there were partial visibility or inconsistent motions.
- Video quality (lighting, angle, resolution) had a slight influence on recognition efficiency.
- Even with these drawbacks, the system is still scalable and applicable in any environment without controlled conditions [13].

CONCLUSION

In this study, an idea has been presented regarding an image-based autism screening process based on the analysis of repetitive movement by means of short video sequences.

The system proves to be successful in showing that:

- Repetitive behavior can be identified using videos.
- Temporal patterns can be learned using deep learning models. It lessens the need for manual observation.
- As a whole, the proposed system provides a cost-effective and scalable method of autism screening that is very easy to implement.

REFERENCES

[1] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders (DSM-5), 5th ed. Washington, DC, USA: American Psychiatric Publishing, 2013.

[2] World Health Organization, International Classification of Diseases 11th Revision (ICD-11). Geneva, Switzerland: WHO, 2019.

[3] L. Zwaigenbaum, S. Bryson, and A. Garon, “Early identification of autism spectrum disorders,” Behavioral Brain Research, vol. 251, pp. 133–146, Oct. 2013.

[4] C. Lord, M. Rutter, P. C. DiLavore, and S. Risi, Autism Diagnostic Observation Schedule (ADOS-2), 2nd ed. Torrance, CA, USA: Western Psychological Services, 2012.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, real-time object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 779–788.

[6] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” arXiv preprint arXiv:2207.02696, 2022.

[7] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, “VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 5998–6008.

[9] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2014, pp. 568–576.

[10] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the Kinetics dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 6299–6308.

[11] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Santiago, Chile, 2015, pp. 4489–4497.

[12] P. S. Rajagopalan, A. Dhall, and R. Goecke, “Self-stimulatory behaviours in the wild for autism diagnosis

[13] M. Tariq, S. L. Daniels, J. N. Schwartz, S. Washington, H. Kalantarian, and D. P. Wall, “Mobile detection of autism through machine learning on home video:

[14] H. Kalantarian, S. Jedoui, M. Washington, H. Tariq, and D. P. Wall, “Labeling of autism videos by crowd workers: A feasibility study,” Journal of Medical Internet Research, vol. 18, no. 6, pp. 1–12, Jun. 2016.

[15] A. Dosovitskiy, L. Beyer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby