**ACM-ASC Internship 2024**

**Milestone II : Phase I**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**TAG 3 : AI and Disabilities Studies**

**Group ID: DIS10**

**Project Mentor: Dr. Namitha K**

**Project Lead: Girish S**

**Title: AI for Neurodiversity**

**Team Members**

| Girish S | AM.EN.U4AIE22044 |
| --- | --- |
| Anuvind MP | AM.EN.U4AIE22010 |
| R S Harish Kumar | AM.EN.U4AIE22042 |
| Harishankar Binu Nair | AM.EN.U4AIE22023 |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Courses**

| **Course** | **Status** | **Links** |
| --- | --- | --- |
| Lecture Collection - stanford Convolutional Neural Networks for Visual Recognition (Spring 2017) | In progress | https://www.youtube.com/playlist?list=PL3FW7Lu3i5JvHM8ljYj-zLfQRF3EO8sYv |
| OpenCV Course - freecodecamp | In progress | https://www.youtube.com/watch?v=P4Z8\_qe2Cu0 |
| AI for everyone - Andrew Ng | In progress | https://www.coursera.org/learn/ai-for-everyone/home/welcome |
| HarvardX: CS50's Introduction to Artificial Intelligence with Python | Not started | https://authn.edx.org/register?course\_id=course-v1%3AHarvardX%2BCS50AI%2B1T2020&enrollment\_action=enroll&email\_opt\_in=false |
| Stanford CS224N: Natural Language Processing with Deep Learning | Winter 2021 | Not started | https://www.youtube.com/playlist?list=PLoROMvodv4rMFqRtEuo6SGjY4XbRIVRd4 |

**Literature Review**

| **Research Papers** | **Read-By** | **Status** |
| --- | --- | --- |
| Vision-based activity recognition in children with autism-related behaviors - DOI: [10.1016/j.heliyon.2023.e16763](https://www.cell.com/heliyon/pdf/S2405-8440(23)03970-1.pdf) | Anuvind, Harishankar,  Harish | Under review |
| Activity Recognition with Moving Cameras and Few Training Examples: Applications for Detection of Autism-Related Headbanging - DOI: [10.1145/3411763.3451701](https://dl.acm.org/doi/abs/10.1145/3411763.3451701) | Girish,  Harishankar | Under review |
| Application of Skeleton Data and Long Short-Term Memory in Action Recognition of Children with Autism Spectrum Disorder - DOI: [10.3390/s21020411](https://www.mdpi.com/1424-8220/21/2/411) | Girish,  Anuvind,  Harish | Under review |
| Social Recognition of Joint Attention Cycles in  Children With Autism Spectrum Disorders - DOI: [10.1109/TBME.2023.3296489](https://ieeexplore.ieee.org/abstract/document/10185592?casa_token=kZ1_LMI2F60AAAAA:94ThDxe6mANPmXCL2ohJg0g0fKQi2KsTPrP3gV2eIhC3P6dqR2fI7V1R9M0L18WAqFhXfNknkjqw) | Girish,  Harishankar | Under review |
| Automated Detection Approaches to Autism Spectrum Disorder Based on Human Activity Analysis: A Review - DOI: [10.1007/s12559-021-09895-w](https://link.springer.com/article/10.1007/s12559-021-09895-w) | Girish,  Anuvind,  Harish | Under review |

**Vision-based activity recognition in children with autism-related behaviors**

-> ASD diagnosis, preprocessed data, collected data, human detection, background removal, temporal convolutional models, light weight, Inflated 3D convnet and Multi Stage Temporal Convolutional Network, Weighted F1 score = 0.83, 3 activities, ESNet backbone weighted F1 score =0.71, uncontrolled Environment

-> [“ASD is a neurodevelopmental disorder characterized by a set of social communication deficits, self-harm, or persistent repetition of actions”](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibABB7B79D860AFD1314264C986AF4606Cs1)

-> “ASD related biomarkers such as those determined through functional magnetic resonance imaging (fMRI). [Facial expressions](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibD2F6BC878DD136CE1116D13D2DF5BCA9s1), [eye gaze, and motor control/movement patterns](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibF685F61EDB75CC4742D43DD621AEA962s1)”

-> dataset from SSBD from [rajagopalan et al](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibA1DD584B1FB446E21CB1B438E6012CB9s1).

-> 75 videos from yt, rajagopalan et al provides no other novel methods other than the “new Self-Stimulatory Behavior Dataset (SSBD) for ASD”

-> additionally added videos from online platforms => not mentioned sources

-> Feature extraction (CNN model) => action classification (TCN)

-> dataset transformed into 168 short clips after removing noisy data and adding new data

-> [168 clip dataset](https://github.com/Samwei1/autism-related-behavior/blob/main/url_list.pdf)

-> “used compressed feature representation for real time throughput”?

-> two crucial steps:

i) a feature extractor that extracts frame-wise action-related semantic features

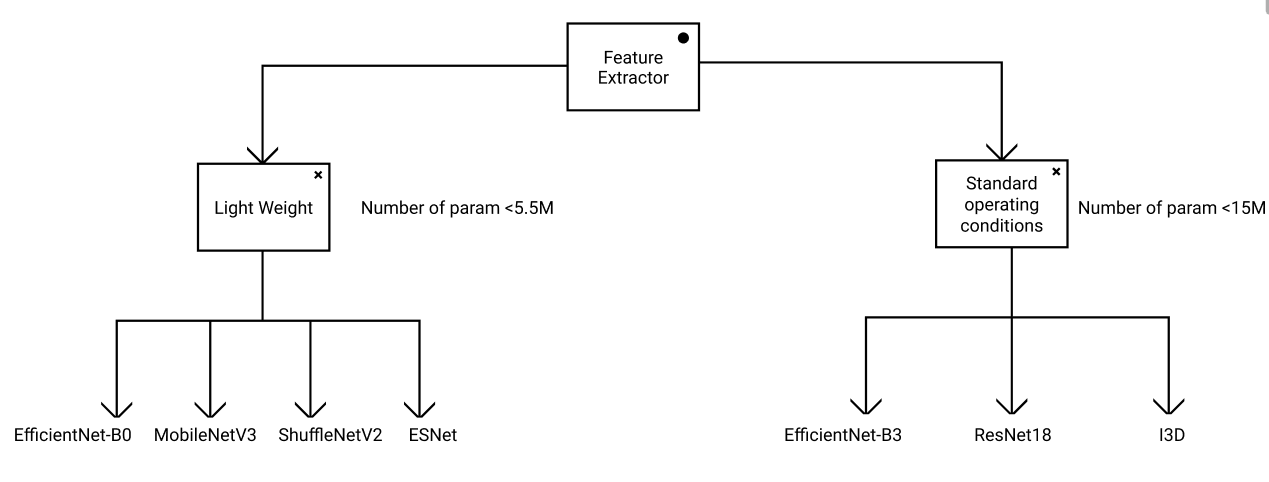
ii) an action recognition model which recognizes autism-related behaviors by modeling the temporal evolution of the extracted frame-wise features.

-> Comparing 7 feature extractors and 4 learning strategies

| Feature Extractor Backbone | Pros | Cons |
| --- | --- | --- |
| EfficientNet | - Computationally efficient compared to ResNet, I3D, etc. - Robust features due to the compound scaling mechanism | - Not suitable for mobile applications. |
| MobileNet | - Computationally efficient. - Does not require an impressive amount of data to train. - Suitable for mobile and embedded vision applications. - Low computational resources are required. | - Limited model capacity to learn large-scale data. - Produce less accurate results compared to EfficientNet, ResNet and I3D |
| ShuffleNet | - Computationally more efficient than MobileNet. - Does not require an impressive amount of data to train. - Suitable for mobile and embedded vision applications. - Low computational resources are required. - More accurate than MobileNet. | - Limited model capacity to learn large scale complex data. |
| ESNet | - Computationally more efficient than MobileNet and ShuffleNet. - Real-time performance. - Suitable for mobile and embedded vision applications. - Does not require an impressive amount of data to train. - Low computational resources are required. | - Limited model capacity to learn large scale complex data. |
| ResNet | - Produce more accurate results than the shallow networks - Suitable for learning large-scale complex data. - The residual connections overcome the vanishing gradient problem. | - Requires more memory and computational resources, which makes it less suitable for mobile applications. |
| I3D | - More suitable for video-based applications to capture spatio-temporal features. - Suitable for learning large-scale complex data. - Produce more accurate results than shallow networks such as MobileNet etc. | - Requires more memory and computational resources, which makes it less suitable for mobile applications. |

-> We select feature extractors appropriate for two different operating conditions:   
 i) a light-weight environment condition suitable for an embedded device

ii) a standard operating condition (i.e. conventional models) suitable for a powerful workstation or server.

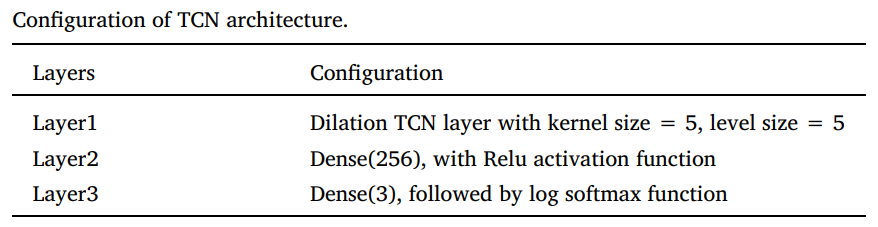


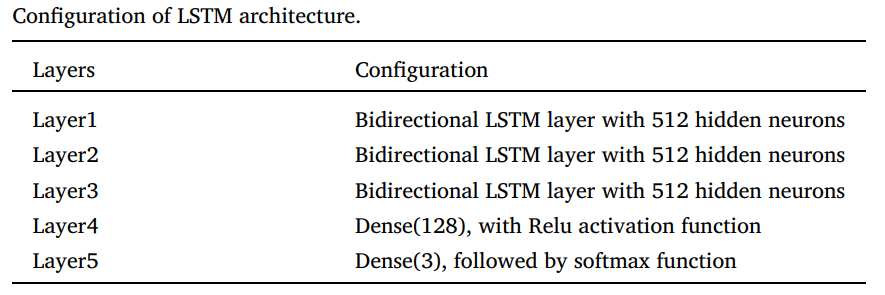
-> more complex networks  
 => ResNet101\_vd 44M

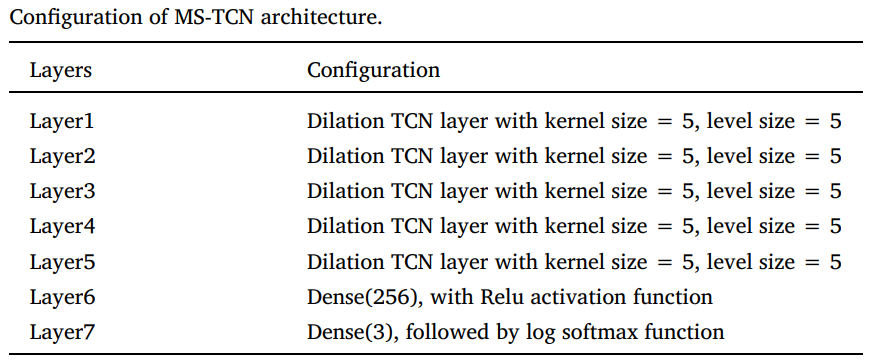
=> HRNet\_W48 78M

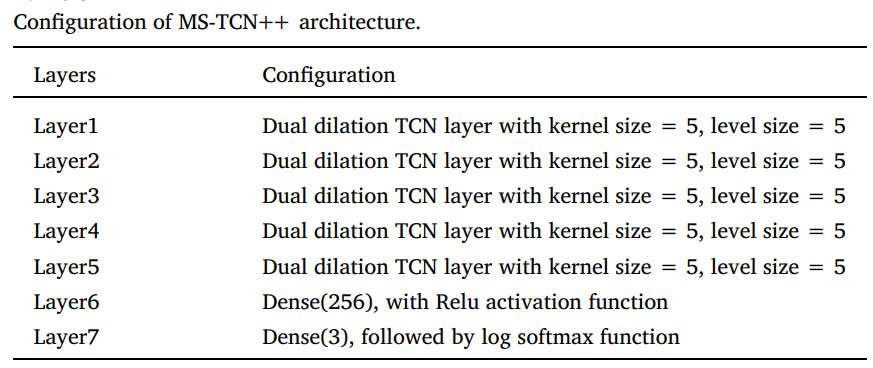
=> SwinTransformer 99M

| Action Recognition Method | Pros | Cons |
| --- | --- | --- |
| LSTM | - Has gates that control the flow of information in and out of the cell. - Low memory requirements. - Faster training. | - May suffer from the vanishing gradients when the data sequence is very long. - May overfit if not properly regularized. |
| TCN | - Able to perform temporal mapping of very long sequences well due to its multiple layers of dilated convolutions (large receptive field). - Can handle variable-length sequences. - Does not suffer from vanishing/exploding gradients. - Able to make more accurate predictions. - Faster training. | - Higher memory requirements compared to LSTM |
| MS TCN | - Achieves better results than TCN as it operates on full temporal resolution (no pooling layers are used). - Able to perform temporal mapping of very long sequences well due to its multiple layers of dilated convolutions (large receptive field). - Can handle variable-length sequences. - Does not suffer from vanishing/exploding gradients. - Multi-stage architecture provides an improved receptive field. - Faster training | - Higher memory requirements compared to LSTM |
| MS TCN++ | - Similar capabilities to MSTCN, however, MSTCN++ achieves better results than MSTCN due to the dual dilated layer. | - Higher memory requirements compared to LSTM |

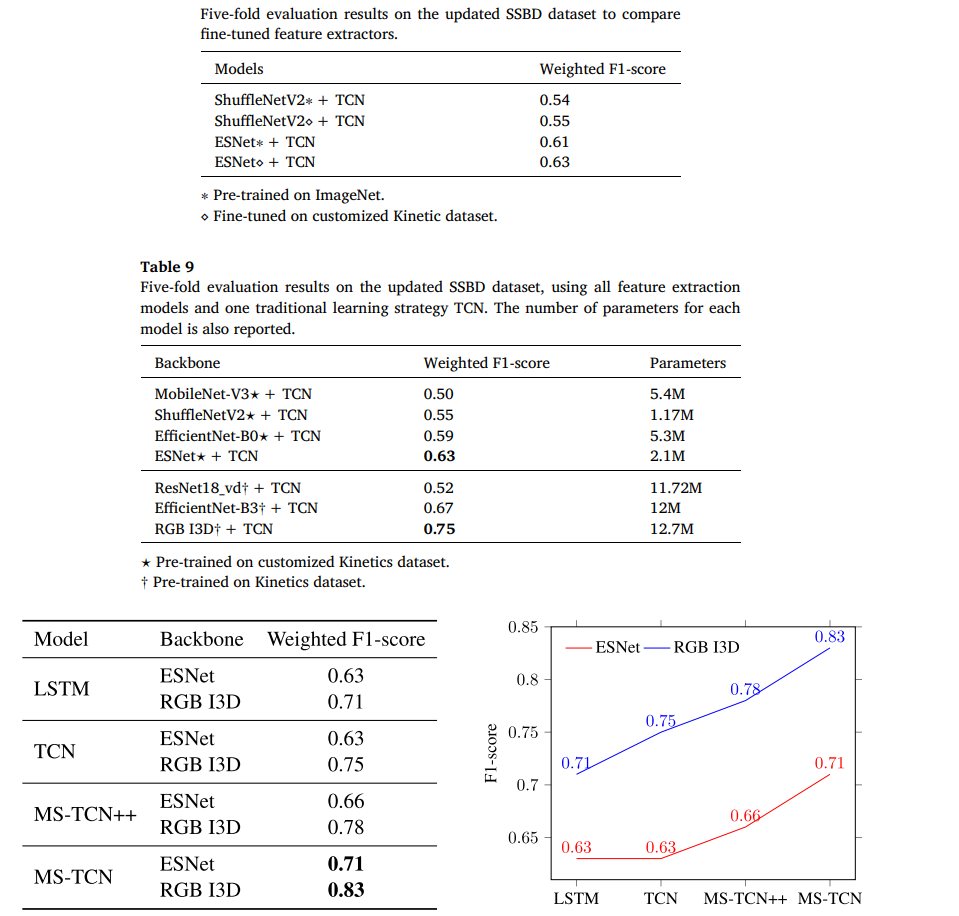








-> result



-> The experimental procedures involving human subjects described in this paper were approved by the CSIRO Health and Medical Human Research Ethics Committee (CHMHREC). The CHMHREC is an NHMRC Registered Human Research Ethics Committee (EC00187). CSIRO Ethics ID 2022\_004\_LR

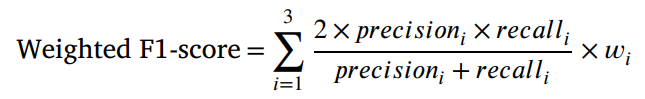
-> The SSBD dataset contains three typical action classes: [arm flapping, headbanging, and spinning](http://refhub.elsevier.com/S2405-8440(23)03970-1/bib4D9D9730B08D8B0C61D9383EE9EEA541s1)

-> pre processing of dataset by cropping the subject alone and removing noisy data from the database

-> pre processing done by a [Mask-RCNN](http://refhub.elsevier.com/S2405-8440(23)03970-1/bib08C68D5A15B6099E3974E116F67B7866s1) pretrained on [COCO](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibC01EB5E415C8DE0E0940FB894AA903ECs1) and implemented through [Detectron2](https://github.com/facebookresearch/detectron2), then resized to a fixed size using the bounding box size

-> 5-fold validation and “Leave-One-Group-Out” setting

-> mean Weighted F1-score for 5-fold validation



-> 200 epochs, batch size=16, learning rate=e^-3, first moment decay rate=0.9, second moment decay rate=0.999

**Activity Recognition with Moving Cameras and Few Training Examples: Applications for Detection of Autism-Related Headbanging**

-> ASD Diagnosis, preprocessed data, collected data, OpenPose realtime multi-person pose estimation, estimated skeletal pose, extracting key points in the head region for accuracy, convolutional neural network (CNN) using the Keras [10] Python library with a Tensorfow [1] backend, a long short-term memory (LSTM) [28] neural network, 3-fold cross validation, 90.77% across the 3 cross-validation folds. The individual F1-scores per fold are 83.3%, 89.0%, and 100.0%.

* -> [“ASD is a neurodevelopmental disorder characterized by a set of social communication deficits, self-harm, or persistent repetition of actions”](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibABB7B79D860AFD1314264C986AF4606Cs1)
* -> “ASD related biomarkers such as those determined through functional magnetic resonance imaging (fMRI). [Facial expressions](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibD2F6BC878DD136CE1116D13D2DF5BCA9s1), [eye gaze, and motor control/movement patterns](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibF685F61EDB75CC4742D43DD621AEA962s1)”
* To minimize overfitting, rotation at a random interval between -45 and 45 degrees and zooming in with a random zoom factor between 1.0 and 2.0.
* dataset from SSBD from [rajagopalan et al](http://refhub.elsevier.com/S2405-8440(23)03970-1/bibA1DD584B1FB446E21CB1B438E6012CB9s1).
* 27 video clips containing headbanging video clips containing “normal” head motions.
* OpenPose realtime multi-person pose estimation to track skeletal keypoints in each frame
* To account for the body part occlusion issue, which would inject unnecessary noise which would confuse the classifer
* extracting key points in the head region.
* We implement a time-distributed convolutional neural network (CNN) using the Keras [10] Python library with a Tensorfow [1] backend.
* We train using Adam optimization [34] with an initial learning rate of 0.0001.
* We perform 3-fold cross validation, ensuring that no child who appeared in the train set would appear in the test set for all folds.
* To minimize overftting and increase generalization, we apply the following data augmentations to each frame: rotation at a random interval between -45 and 45 degrees and zooming in with a random zoom factor between 1.0 and 2.0.

**Pros And Cons of Dense Optical Flow:**

Pros : Dense optical fow computes fow for all points in a frame, resulting in “flow vectors” with a magnitude and direction.

Cons : extra non-relevant movement patterns add much noise to the dense optical fow image, making it a non-ideal representation.

**Pros And Cons of Lucas Kanade Optical Flow:**

Pros : Lucas-Kanade optical fow, in contrast to dense optical fow, computes fow for a sparse number of points pre-defned by the user [3], for example detected edges or corners.

Cons:The Lucas-Kanade method brings its own set of limitations by detecting movement outside of the body to an extent that is more dramatic than dense optical fow. This technique is particularly sensitive to background movement detection associated with slight shifts in the camera.

**Pros And Cons of Pose Estimation:**

Pros: Pose estimation is a technique which is more robust to camera movement compared to optical flow. . OpenPose uses a CNN which is trained to predict part afnity felds, or fow felds representing relationships between body parts, and confdence maps which encode body part locations. Unlike optical fow, OpenPose predicts each frame independently of the surrounding frames.

Cons: There are some clear limitations to using unmodifed pose estimation. The noisy skeleton problem is a documented issue which arises when body parts are self-occluded. We observe that body part occlusion is a frequent occurrence in unstructured home videos, making unmodifed pose estimation a non-ideal feature representation.

**Application of Skeleton Data and Long Short-Term Memory in Action Recognition of Children with Autism Spectrum Disorder**

-> asd diagnosis, skeletal data + LSTM

-> OpenPose algorithm for skeletal data

-> 4 denoising techniques applied to skeletal data before passing to stage 2

-> stage 2 = tracking humanoids based on temporal data

-> stage 3 = LSTM classification

-> OpenPose + denoising > tracking > classification

-> [characterized by persistent deficits in social communication and interaction as well as restricted and repetitive behaviors](https://sci-hub.live/10.1016/b978-0-12-809324-5.05530-9)

-> First, we scale the coordinates of the key points outputted by OpenPose to the same units

-> Second, we remove the five joints on the head

-> Third, we discard frames without skeleton data or missing important joints.

-> Finally, we use the relative joint positions in adjacent frames to fill in the unrecognized joint positions

-> head joints provides less information on the activity classification, thus they removed the head joints from the maps

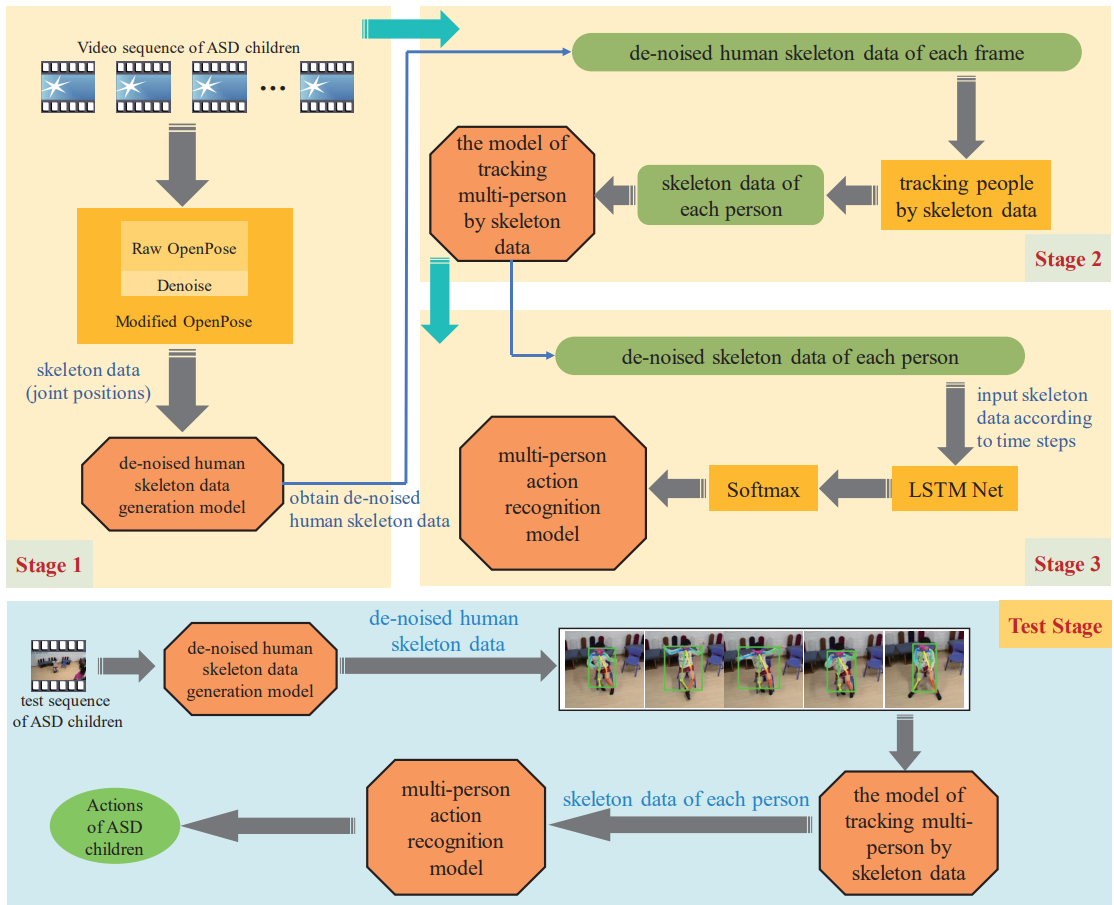
-> frames with major joints missing will be discarded

-> other joints missing will be filled in by the temporal frames adjacent to them

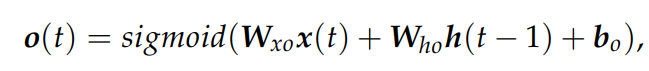
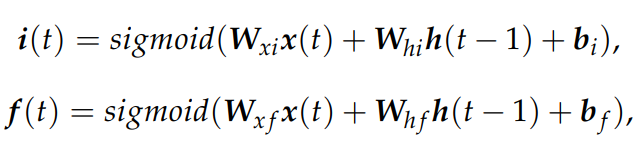
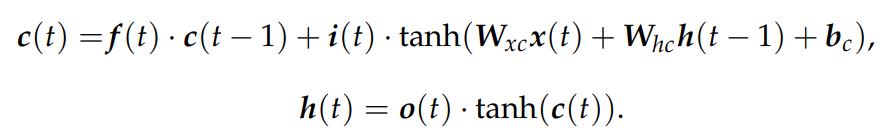
-> facilitates tracking of multiple skeletons

-> DeepSORT and SORT are computationally expensive

-> a new algorithm is developed which finds the distance of the joint i of the jth person id and calculates the relative position to the frame’s center, this location is compared with adjacent temporal frames to identify if it is the same person for tracking the skeleton





-> collected data, 5 actions, sit stand squat shake body shake hands, 1062 sequences

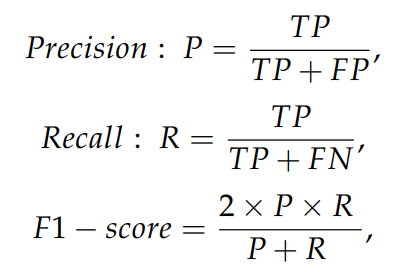
-> normalized to 656x368 -> sent to OpenPose

-> for training the model, only one subject should taken in case of multiple subjects in frame, the subject closest to the center of the frame was taken

-> 70-30 split, 32x13x2, 32 frames, 13 skeletal key points, 2 x-y coordinates

-> hidden states = 128, learning rate =0.001, time steps =32, prob\_threshold=0.9

-> mentioned [AlphaPose](https://ieeexplore.ieee.org/abstract/document/9954214/) is also a good method to implement



-> Compared other methods

-> DNN, SVM, decision tree, RF

-> F1 score is used for performance metric

Action Original De-noised

Sit 0.734 0.8868

Stand 0.9279 0.9896

Squat 0.7757 0.8925

Shake body 0.8604 0.9415

Shake hand 0.9094 0.9803

**Automated Detection Approaches to Autism Spectrum Disorder Based on Human Activity Analysis: A Review**

**-Datasets Used-**

**Skeleton Dataset :**

[Rihawi et al](https://ieeexplore.ieee.org/abstract/document/8078544?casa_token=omwAbF501e0AAAAA:QQmTyOr5DqGqc2a3c4yX5Zg-AYK24Pi5ASYaLAjbBnkarUt4BagYN_YnVTohfbxjRFIkhQ0U8rU5). developed the first publicly available 3D dataset named ‘3D-AD’ based on ASD subjects using the Kinect-v2 camera. This dataset includes depth maps, which have been captured at 33 frames per second. The sequence of skeleton joint features was collected for ten different actions, e.g., hands-on the face, hands back, hand moving front of the face, headbanging (or rocking back and forth), tapping ears, flicking, hands stimming, toe walking, playing with a toy, and walking in circles. This research reported Dynamic Time Wrapping (DTW) distance as a distinguishable feature to detect ASD and TD.

[Al-Jubouri et al.](https://www.researchgate.net/profile/Ahmed-Abdulrahman-6/publication/344044462_GENERATING_3D_DATASET_OF_GAIT_AND_FULL_BODY_MOVEMENT_OF_CHILDREN_WITH_AUTISM_SPECTRUM_DISORDERS_COLLECTED_BY_KINECT_V2_CAMERA/links/5f4f91b1458515e96d22fca7/GENERATING-3D-DATASET-OF-GAIT-AND-FULL-BODY-MOVEMENT-OF-CHILDREN-WITH-AUTISM-SPECTRUM-DISORDERS-COLLECTED-BY-KINECT-V2-CAMERA.pdf) have published a publicly available 3D skeleton-based gait dataset of 50 children with Autism and 50 typically developing (TD) child. In the video data captured in Kinect v2, the participant walked along a line at their normal speed. The skeletal data included the joint position of 25 three-dimensional joint positions, sixteen angles and on gait cycle from 2 or more feature extracted from the data. Diferent data augmentation techniques: Jittering, Scaling, translation, Flipping and slicing were used to increase the diversity of data.

**Video Dataset :** [Zunino et al.](https://ieeexplore.ieee.org/abstract/document/8545095?casa_token=t7wb3xJdNH0AAAAA:w8fXgt9CFsY3EVGvJgpDViXN0ATkhhOcjPIBnbRt-oyhcXY5lCGjD_eMsK7AwbRIl0CrXqeg-Dvv) developed the only available video dataset to analyze the action style of an autistic child. The dataset includes activities such as the task of placing, picking, passing, and grasping a bottle of a particular size by an autistic child. A Vicon VUE video camera with resolution: 1280 x 720 pixels and a frame rate of 100 frames/sec was used to devise the whole experimentation. The study includes 20 TD and 20 ASD children as subjects, whose state of autism was confirmed by the DSM-5 method. All participants were in between 7 to 12 years of age.

**Gaze Dataset :** [Duan et al.](https://dl.acm.org/doi/abs/10.1145/3304109.3325818?casa_token=-Loud73uj8sAAAAA:5qGdk-Q5S0HJDnFeh1SuclTNHe3IjSo7LpmuBgcOH-kii6X0Hj6MJOW6GhX37r0-_igcD77z86c5) developed an eye movement dataset named ‘Saliency4ASD’ from 14 ASD and 14 typically developed children. All the participants were in between 5 to 12 years of age. During the eye-tracking process, the participants viewed 300 images, including natural scenes, animals, the human body, and objects. This dataset provided fixation points, fixation maps, heat maps, and scan path data of the participants.

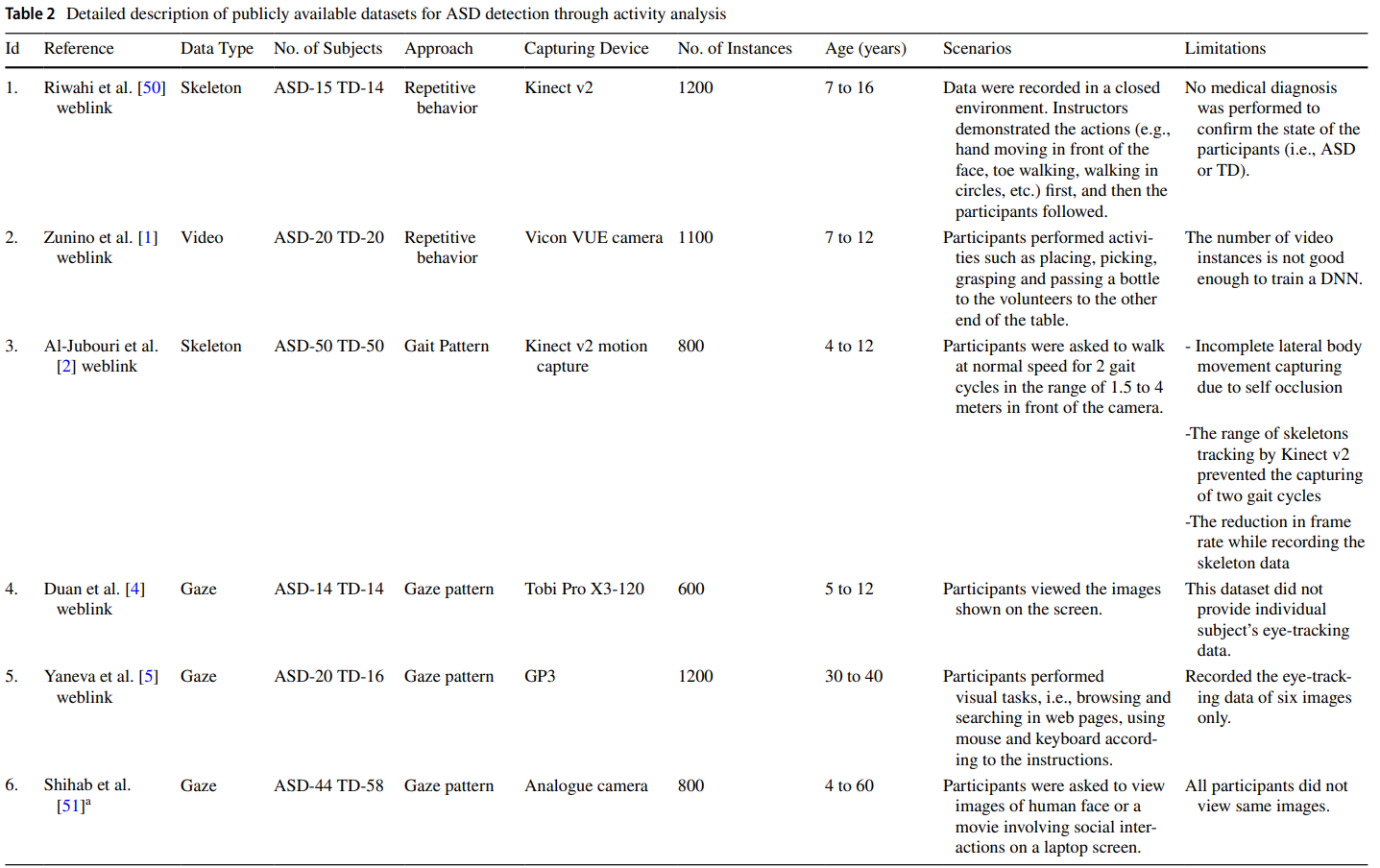
[Yaneva et al.](https://dl.acm.org/doi/abs/10.1145/3192714.3192819?casa_token=Ss628HTZPogAAAAA:YppwDpMAn18-lXBb1_GpHeMMBO3XDBuCDyBJYMDDqm4dhA2G85eSDIXK3LU9Hdvj7SsYygAsjwPl) also developed a gaze dataset, which included autistic adults instead of children. The participants were in between 30 to 40 years of age. The experiment included 30 participants, where 15 were diagnosed with high functioning autism or Asperger’s syndrome, while the rest were non-autistic individuals. In the data collection process, the participants were asked to perform visual tasks such as web browsing and searching using a mouse and keyboard according to the given instructions. The study provided fixation time on the area of interest in images, duration of searching task in a web page, and the number of fixation points. This work utilized this dataset to classify ASD individuals and healthy people applying simple logistic regression

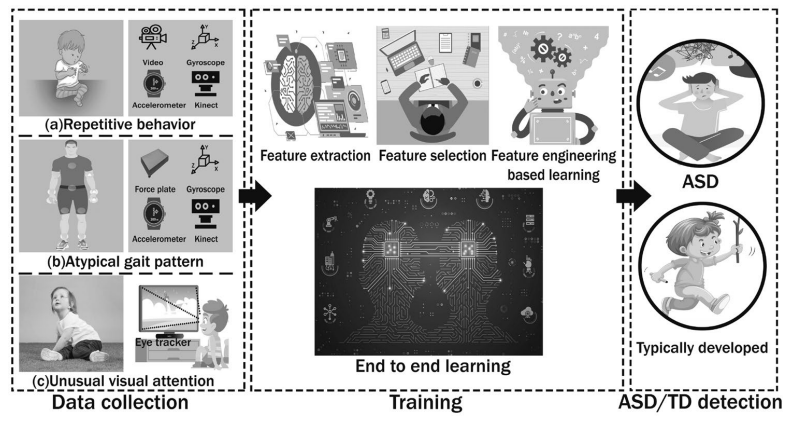
[Shihab et al.](https://downloads.hindawi.com/archive/2020/3407907.pdf) provided a gaze dataset that comprises face-scanning data of adults and children diagnosed with ASD. The participants were in between 4 to 60 years of age. Participants were asked to view human face images and movie clips involving social interactions on the laptop screen while the eye-tracking data were recorded using two analog cameras placed in front of the laptop. This research studied the difference in the face-scanning pattern of ASD and TD individuals and classified ASD and TD using the PCA.

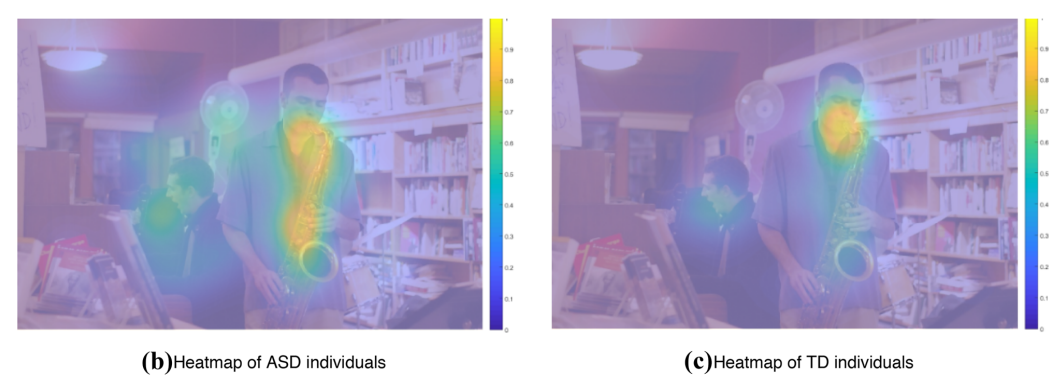
**Performance Metric:**

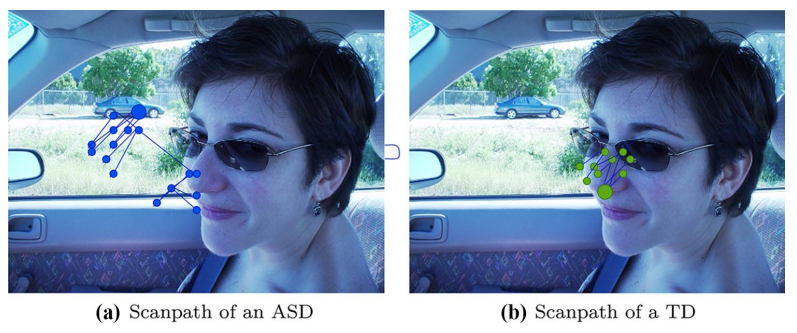
the efficiency of a method is assessed with accuracy (i.e., ratio of the correctly labeled samples to a total number of samples), recall or sensitivity (i.e., the proportion of actual positive cases that got predicted as positive, in other words, true positive), specificity (i.e., the proportion of actual negatives, which got predicted as the negative or true negative), precision (i.e., the number of positive class predictions that actually belong to the positive class) and F1 score (i.e., the harmonic mean of the precision and recall.).

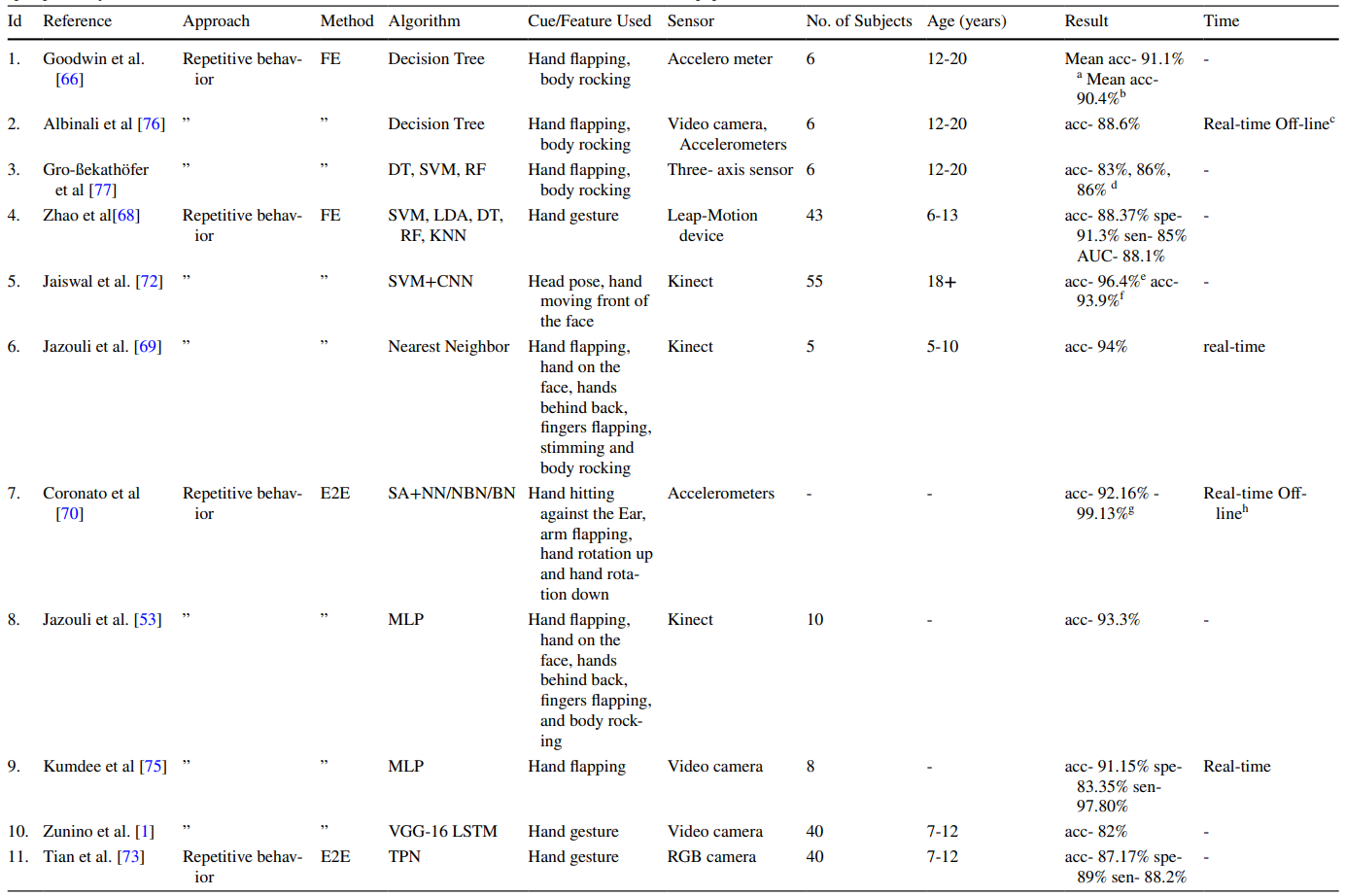
visual or graphical tools like receiver operating characteristic (ROC) curve and area under the curve (AUC) are used in measuring the performance of models.

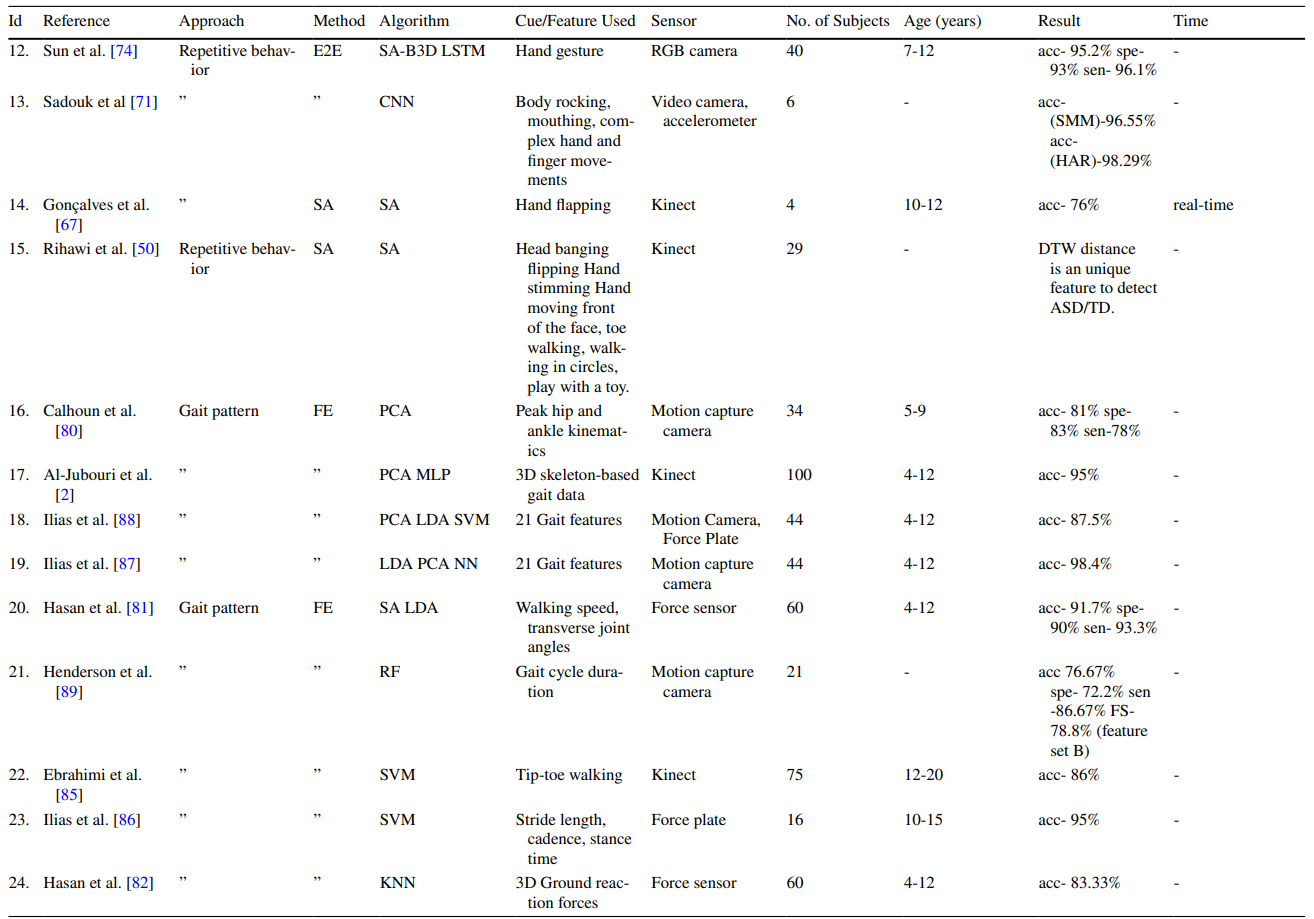


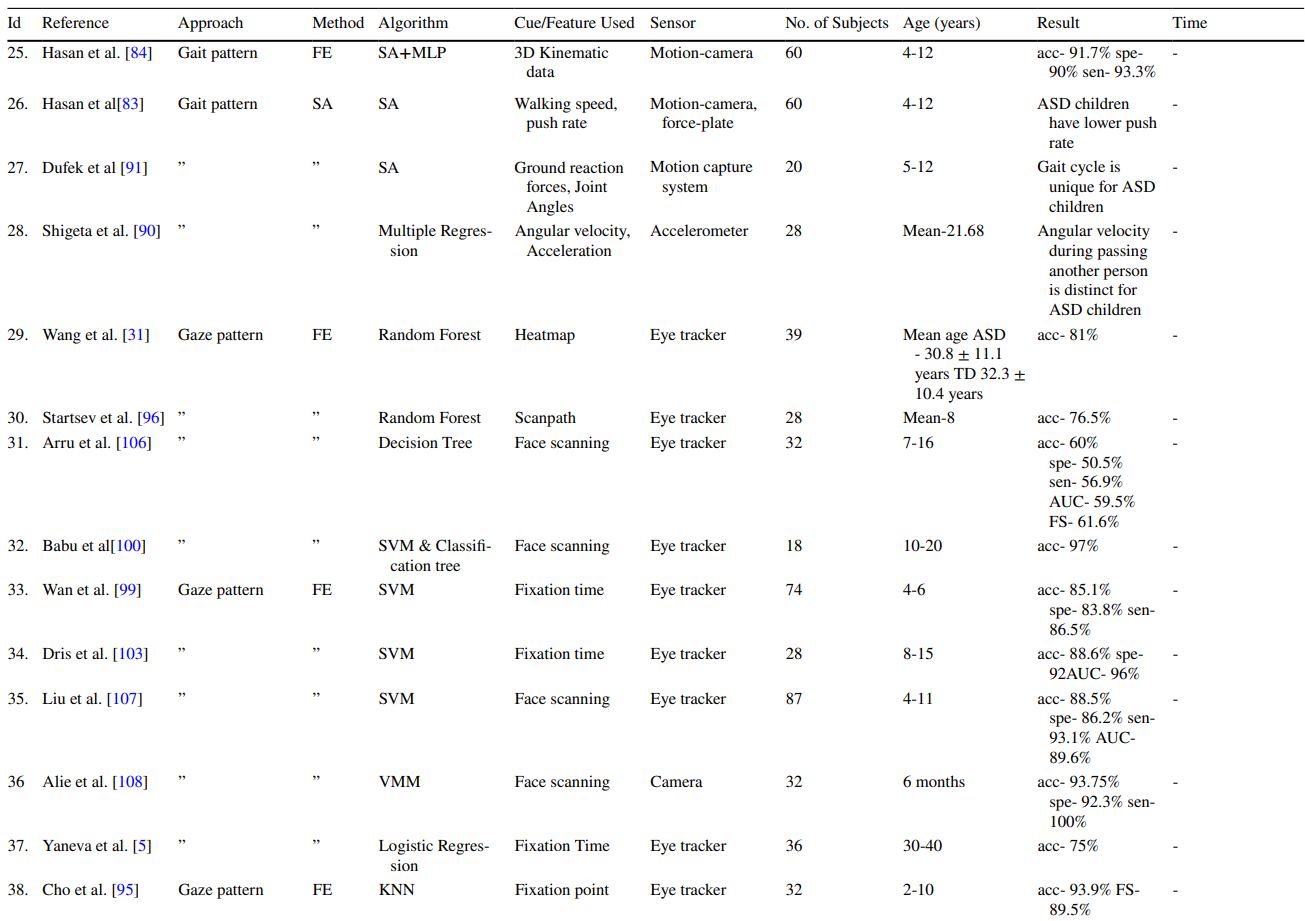


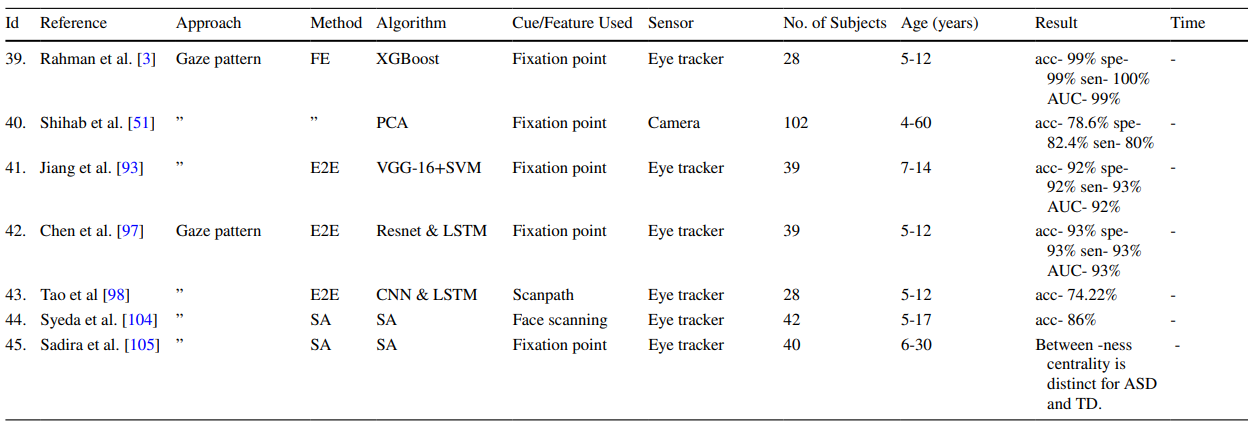


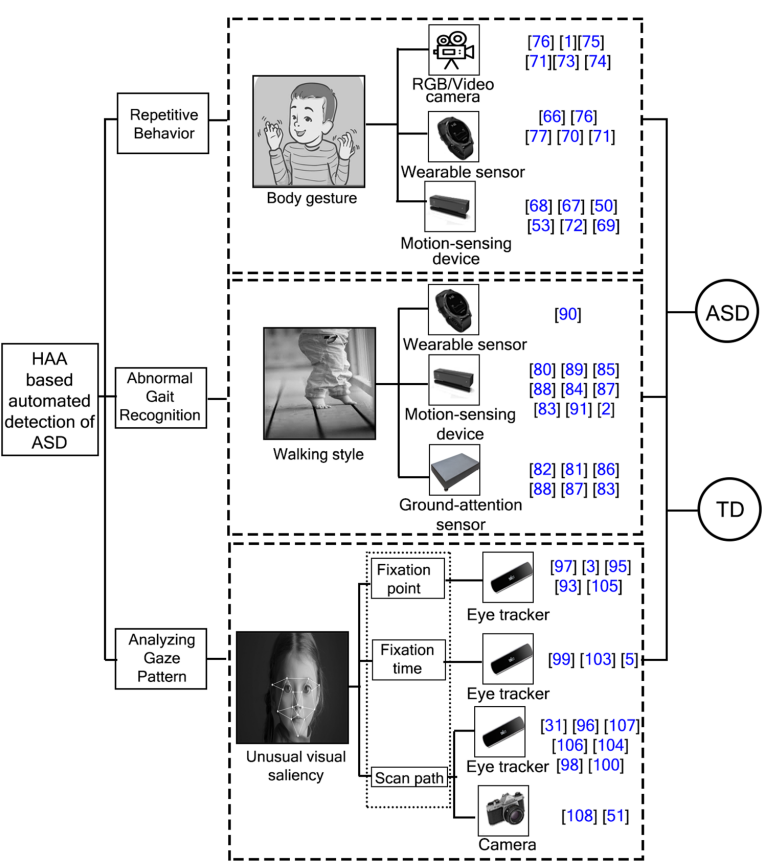




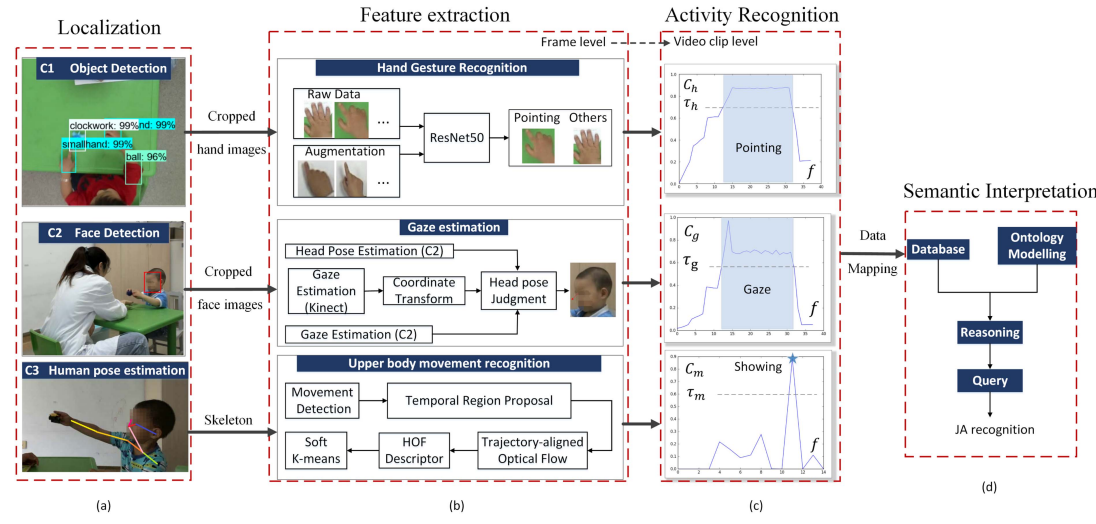








**Social Recognition of Joint Attention Cycles in Children With Autism Spectrum Disorders**



-> automatic recognition of both responding to JA (RJA) and initiating JA (IJA) in children with ASD.

-> multi-modal framework combining vision and human behavior analysis for joint attention assessment.

**Dataset:**

-> The dataset consists of recordings of free play sessions between children (ASD and typically-developing) and adults in a controlled environment.

-> The children were aged 4-6 years old.

-> Ground truth labels for JA cycles were provided by human coders.

-> Included children with ASD, mental retardation (MR), developmental language disorder (DLD), and typically developed children (TD).

-> 7 children with ASD, 5 with MR, 5 with DLD, and 3 TD children.

The framework employs a multi-stage approach:

* Localization: Detects the positions of the child and adult in the video frames.
* Feature Extraction: Extracts features related to gaze direction, head pose, hand gestures, and object interactions.
* Activity Recognition: Utilizes a rule-based approach to identify three key JA activities:
  + Attention estimation (e.g., following gaze direction)
  + Spontaneous pointing
  + Showing actions
* JA Cycle Recognition: Combines the identified activities into temporal sequences to recognize JA cycles.

-> Gaze direction and head pose: estimated from facial landmarks extracted using a pre-trained deep learning model (MTCNN).

-> Hand gestures and object interactions: detected using open-source libraries (OpenPose, YOLOv5).

**Methodology:**

Leverages localization, feature extraction, and activity recognition.

Links critical activities in joint attention (JA) through semantic interpretation.

Performance Metric:

Intra-class coefficient of 0.959.

Demonstrates recognition reliability.

-> Results:

-> Significant differences in JA performance between ASD and TD groups.

-> Significant correlations found between RJA task response latency and IJA times, as well as between social gaze accumulation and JA performance.

-> Limitations Addressed:

-> Small sample sizes acknowledged.

-> Need for improvement in feature extraction methods and statistical analysis highlighted.

-> Potential Applications:

-> Highlighted potential in clinical diagnosis and intervention.

-> Future considerations include increasing sample size, refining technical methods, and conducting further statistical analyses.

-> Contribution:

-> Provides a comprehensive framework for recognizing JA in children with ASD, MR, DLD, and TD.

-> Offers potential for improving clinical diagnosis of autism and enhancing joint attention skills in affected children.