AI Vision to Care: A QuadView of Deep Learning for Detecting Harmful Stimming in Autism

Nyamtulla Shaik - Masters Defense

Date - 5/6/2025

Department of Electrical Engineering and Computer Science

UNIVERSITY OF KANSAS

KU SCHOOL OF ENGINEERING
The University of Kansas

AI Vision to Care: A QuadView of Deep Learning for Detecting Harmful Stimming in Autism

Chair:
Dr. Sumaiya Shomaji

Committee:
Dr. Bo Luo
Dr. Dongjie Wang

Introduction

Stimming

Stimming refers to repetitive actions or behaviors used to regulate sensory input or express feelings.

Actions such as

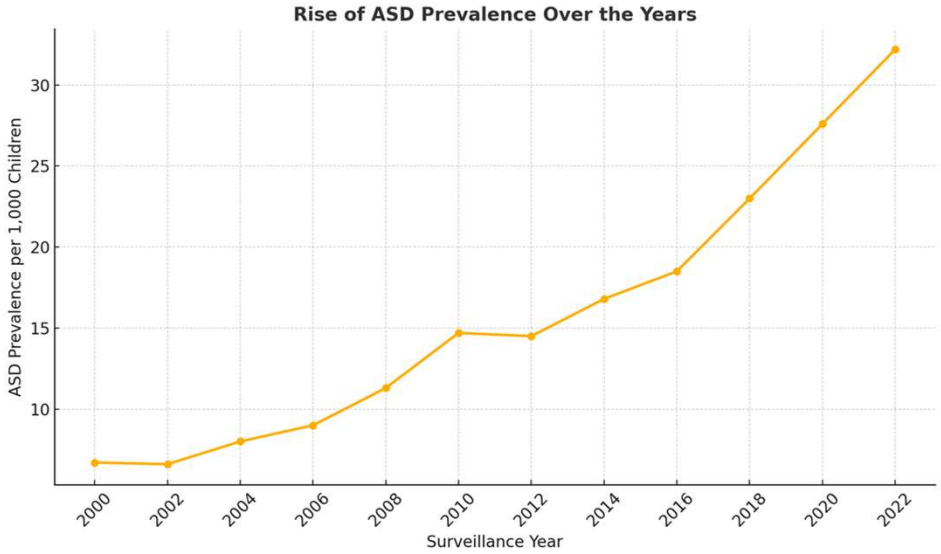
“hand flapping”
“back-and-forth rocking”
“spinning”
“repeating words, or vocalizations”

Commonly observed in children with developmental disorders like Autism.

OpenAI. (2025). *Hand flapping, spinning, and head banging stimming actions in children* [AI-generated images]. ChatGPT (DALL-E model). https://openai.com/

Significance Of Stimming

- "**Raising trend** in number of children with autism"
  - Number of children with autism *in the United States was estimated to be approximately 1 in 54 in 2016 and increased to 1 in 31 by 2022*



- "A **common characteristic** of autism spectrum disorder(ASD)"
  - *80-90% of individuals with autism engage in stimming behaviors*

- "Can also be **observed in other developmental or sensory processing disorders**"
  - *Like obsessive-compulsive disorder (OCD), Intellectual Disability (ID), Sensory Processing Disorder (SPD), Anxiety Disorders, Tourette Syndrome"*

https://www.cdc.gov/autism/data-research/index.html

Is It Harmful?

Types of Stimming

Two main categories: Harmful and Non-Harmful

- Stimming behaviors like **Head-banging**, **self-biting**, **skin-picking**, **hair-pulling**, **hitting oneself** may cause physical harm or injury to the individual.
- Stimming behaviors like **Arm-flapping**, **finger flicking**, **rocking**, **spinning objects**, **humming** are generally harmless and serve as self-soothing methods.
- Harmful stimming require intervention and support!!

Problem Statement

Detecting Harmful Stimming

How to know when a child is performing a harmful stimming?

Keep Watching the child continuously and Intervene when harmful stimming action is seen

How can we make this feasible ?

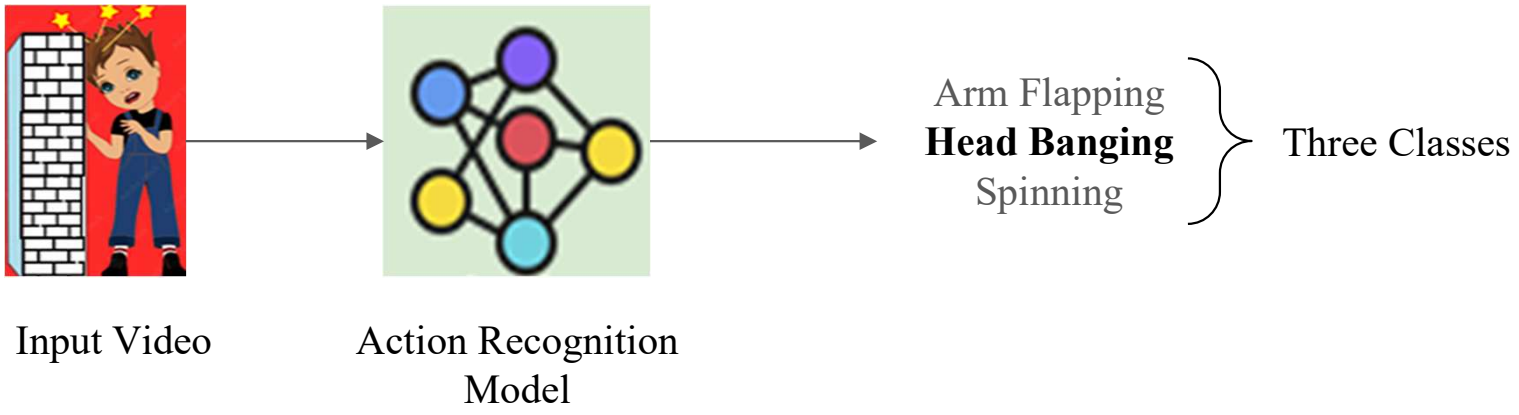
- ➔ Automatically detect stimming from live video feed and notify care givers.
- ➔ Correctly classify an action to know if it's harmful or not harmful stimming.

OpenAI. (2025). *A young child spinning in a playroom while a caretaker watches* [AI-generated image]. DALL·E. https://openai.com/dall-e

Objective

Objective

"This studies objective is to develop a deep learning-based framework for stimming action classification".



Literature Review

Existing Studies

	Name of Paper (Venue)	Year	Objective	Accuracy	Dataset	Classes
1	Rajagopalan et al. (Computer Vision and Image Understanding (Elsevier))	2013	Classify Stimming Actions	50.70%	SSBD	3
2	Zamzmi et al. (Journal of Medical Imaging (SPIE))	2019	Differentiate pain vs. stimming	84.20%	Infant pain behavior datasets	2 (ASD/No ASD)
3	Tariq et al. (PLOS ONE)	2020	ASD detection	85%	162 home videos (2-4 years old)	2 (ASD/No ASD)
4	Negin et al. (Computers in Biology and Medicine (Elsevier))	2021	Classify Stimming Actions	78%	Private dataset from ASD diagnostic centers	3

Existing Studies

	Name of Paper (Venue)	Year	Objective	Accuracy	Dataset	Classes
5	Washington et al. (Nature Scientific Reports)	2021	Headbanging detection	F1-score: 90.77%	SSBD	2 (ASD/No ASD)
6	Ali et al. (Sensors (MDPI))	2022	Classify Stimming Actions	86.04% (Activis), 75.62% (SSBD)	Activis(Clinical labs), SSBD	3
7	Wei et al. (Heliyon)	2023	Classify Stimming Actions	F1-score: 83%	SSBD	3
8	Serna-Aguilera et al. (IEEE)	2024	ASD detection	81.48%	UARK and UTSA private video datasets	2 (ASD/No ASD)

Comparing Existing Studies

	Name of Paper	Year	Stimming Classification	Public Dataset (SSBD)	3 classes
1	Rajagopalan et al.	2013	Yes	Yes	Yes
2	Zamzmi et al.	2019	No	No	No
3	Tariq et al.	2020	No	No	No
4	Negin et al.	2021	Yes	Yes	Yes
5	Washington et al.	2021	No	Yes	No
6	Ali et al.	2022	Yes	Yes	Yes
7	Wei et al.	2023	Yes	yes	Yes
8	Serna-Aguilera et al.	2024	No	No	No

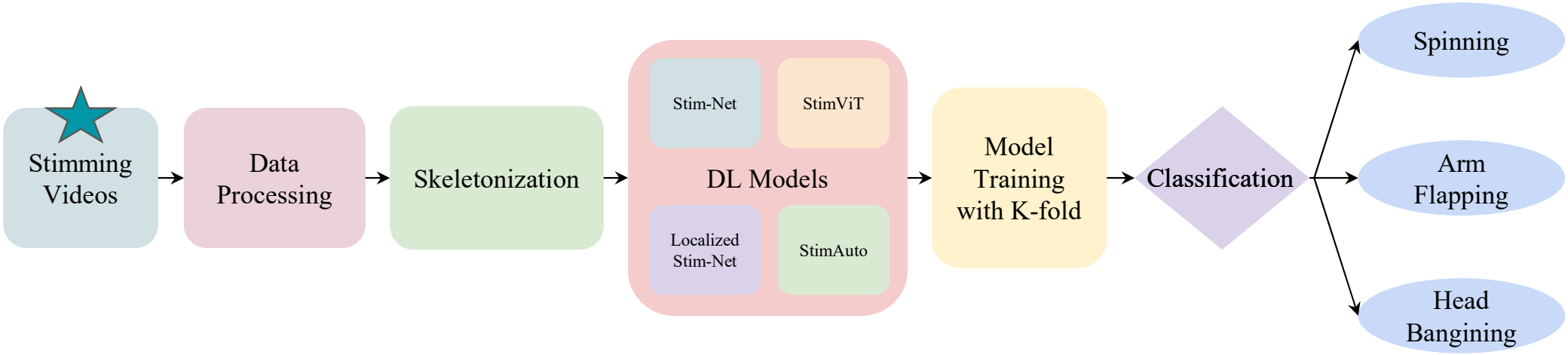
Challenges

- ➔ Limited, real-world data.
- ➔ Moving camera angles, noise, occlusions.
- ➔ Varying size of subjects and speed of actions.
- ➔ Multi-class stimming classification.
- ➔ Adapting skeletal data for various deep learning models.

Contributions

- ➔ Classifying multiple stimming actions.
- ➔ Introducing a scoring based action video data cleaning process.
- ➔ Comparison of different deep learning models.
- ➔ First work to use ViTs and Autoencoders in ASD Research.
- ➔ Proving that Skeletal-Based Deep Learning increases the performance.
- ➔ 7% increase in the accuracy compared to existing studies.

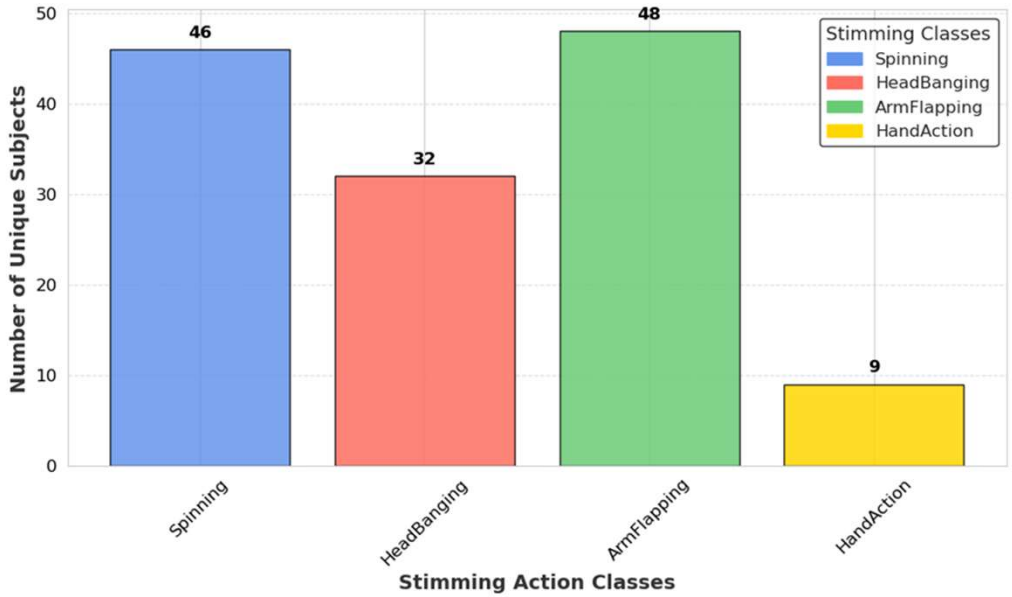
High Level Workflow



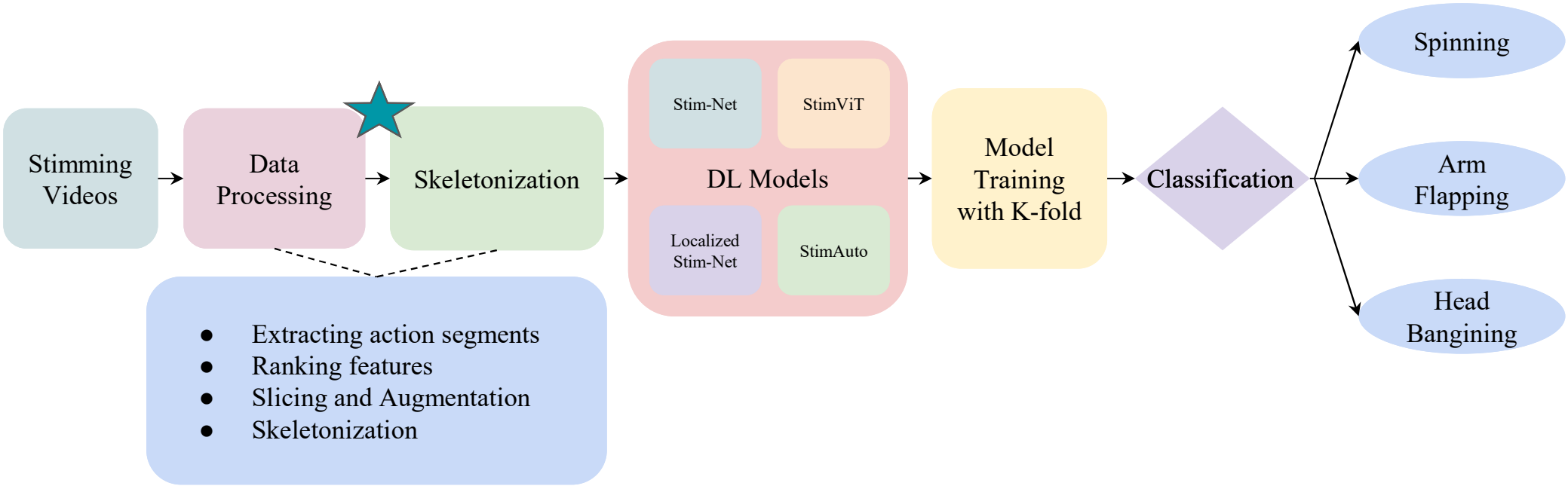
Data and Processing

Dataset

- ➔ Used the Self-Stimulatory Behavior Dataset (SSBD) introduced by Rajagopalan et al. (2013) as a base dataset.
- ➔ Collected few more videos, adding up to 135 videos total in our extended dataset initially.



High Level Workflow



Data processing

- A metadata files stores information such as subject ID, url, label, action timestamps of each video.
- The video downloader script reads URLs from metadata file and ensures proper downloading and saving the video.

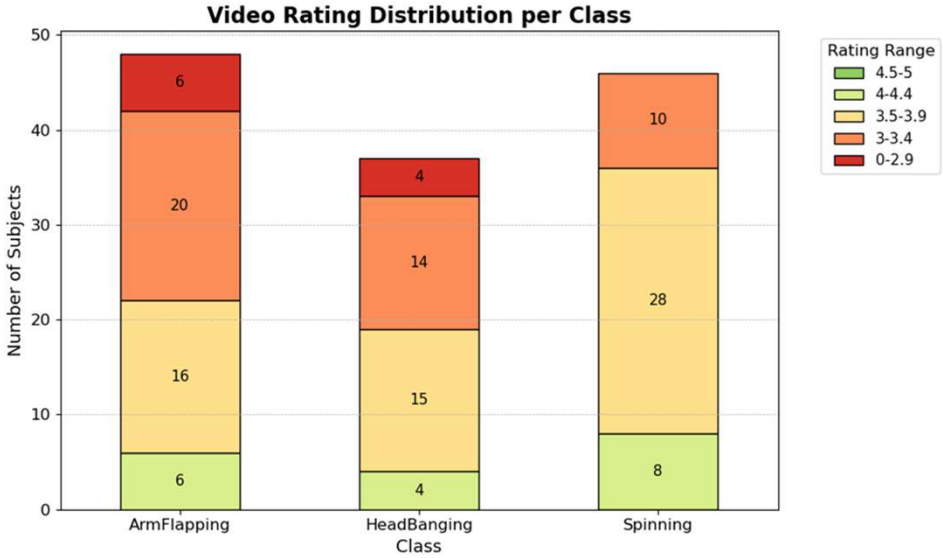
Data processing

- ➔ We strip the video files using the timestamps and extract solely the pertinent action segment.
- ➔ Each extracted action segment is then saved as a separate video file under the same subject.
- ➔ As a result of this, we end up with 144 videos as some videos have multiple action segments.

Data processing

Video Downloading → Action Segments Extractions → Manual Inspection and Cleaning → Video Slicing → Skeletonization

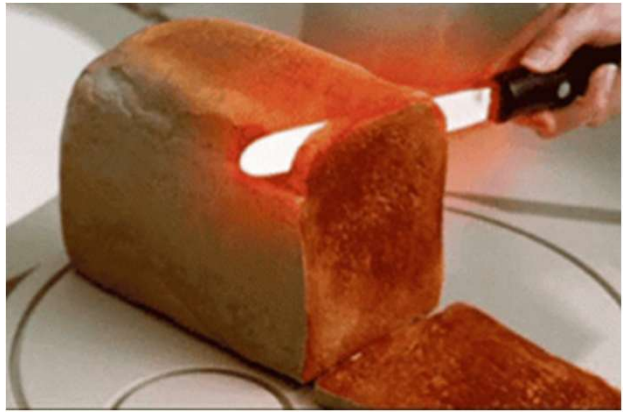
- ➔ Rated videos based on 5 features.
  - ◆ Action to label match.
  - ◆ Quality of action(proper and continues or not).
  - ◆ Bad video resolution.
  - ◆ Multiple people in the frame.
  - ◆ Footage is shaky and subject has occlusions.
- ➔ Discarded videos with poor overall rating.
- ➔ Resulted in shrinking the dataset to just 92 subjects.



Data processing

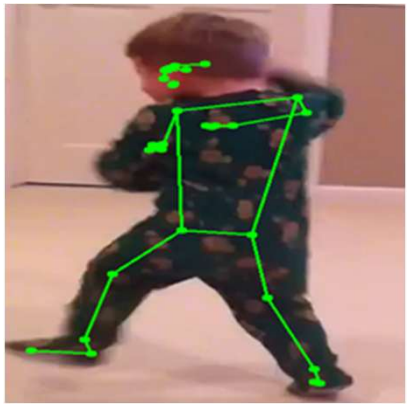
Video Downloading → Action Segments Extractions → Manual Inspection and Cleaning → Video Slicing → Skeletonization

- ➔ Videos varies in length.
- ➔ To make them uniform we cut them into 72 frame sized slices.
- ➔ These resulted in each subject having 1-4 clips.
- ➔ FInal dataset has 92 subjects and 289 video clips.

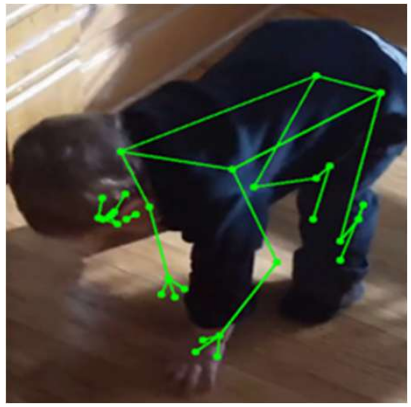


Data processing

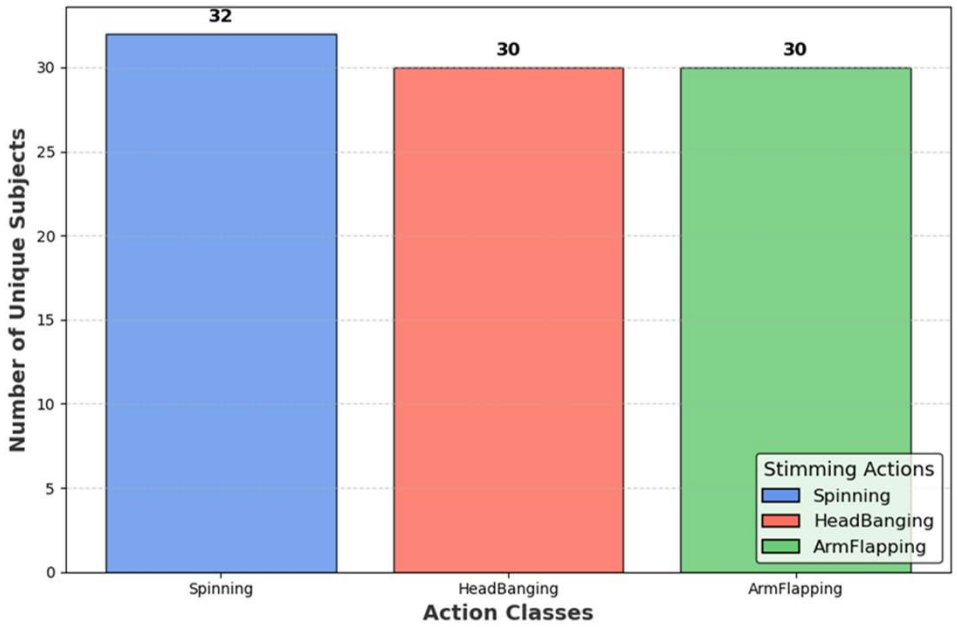
Video Downloading → Action Segments Extractions → Manual Inspection and Cleaning → Video Slicing → Skeletonization



- ➔ Used Google's open source framework MediaPipe, for extracting human pose keypoints.
- ➔ Each frame generates 33 key points in 3D space (x, y, z).
- ➔ Skeletons carry the properties needed to recognize the action.
- ➔ Skeletons 33 × 3 size is much smaller that RGB images 224 × 224 × 3 size.

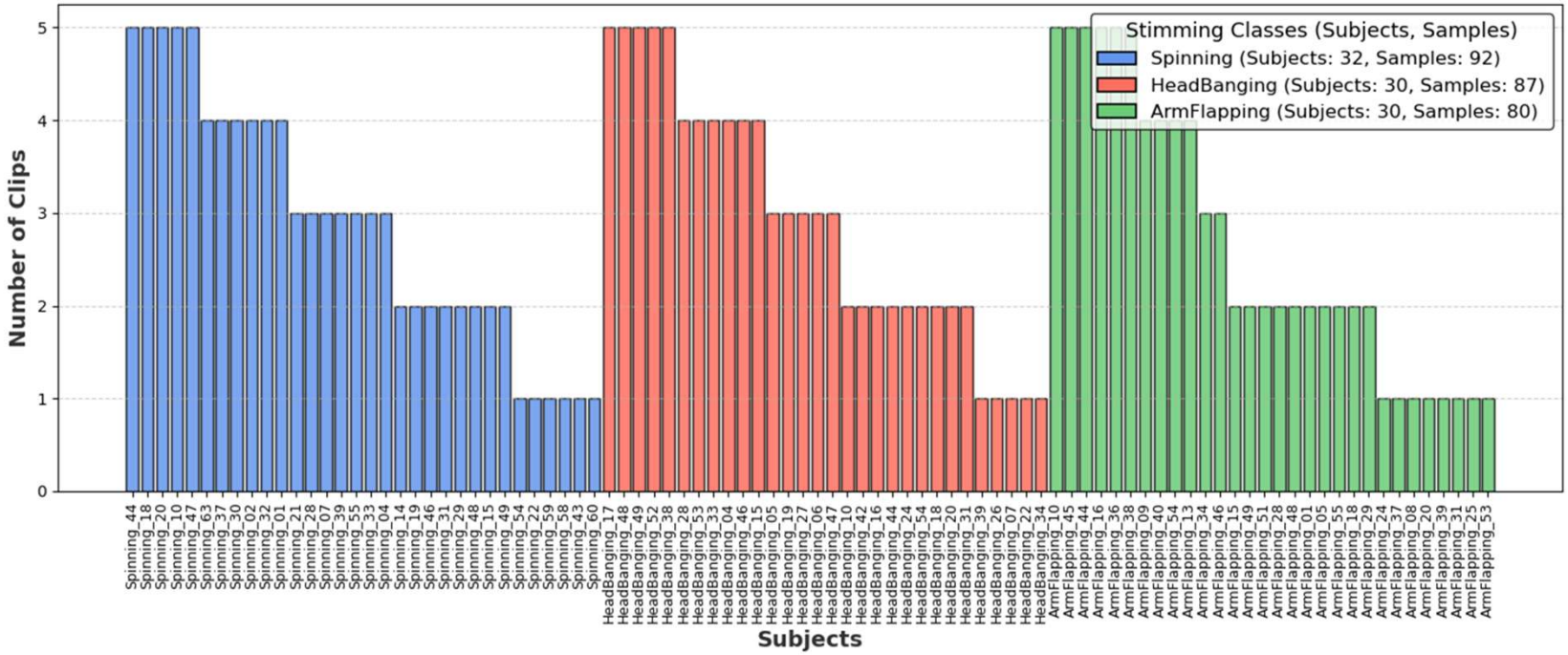


Dataset after Processing



The distribution of classes after processing

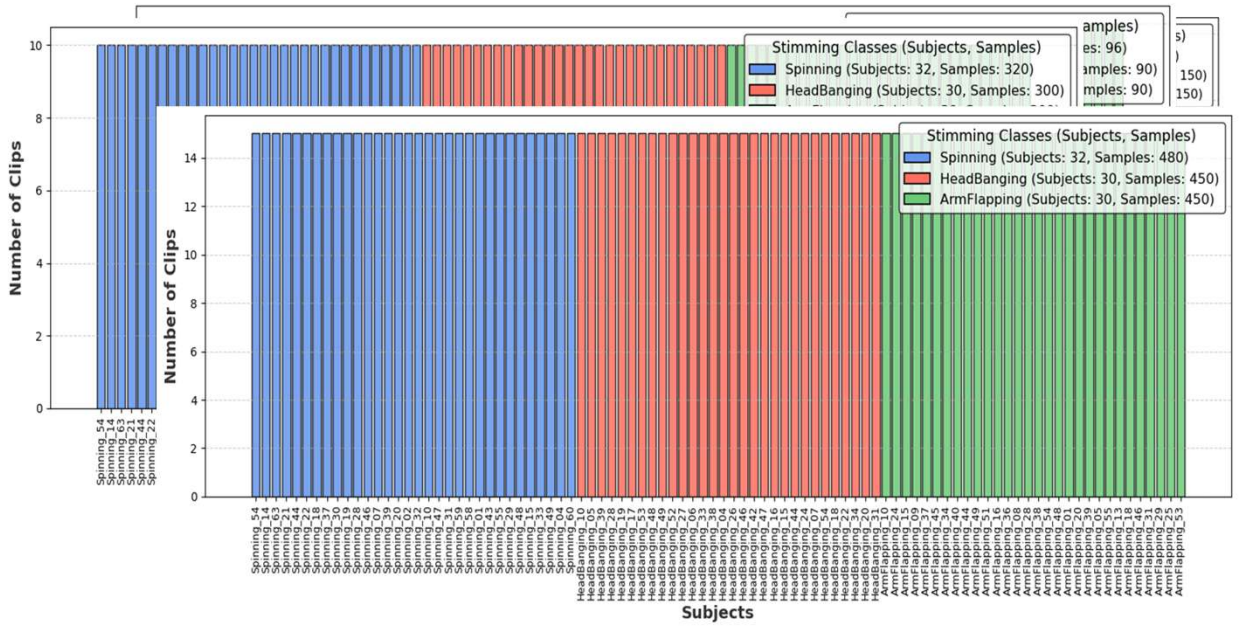
Dataset after Processing



The distribution video clips in each class

Data Augmentation

Augment dataset to achieve N number of videos for each subject

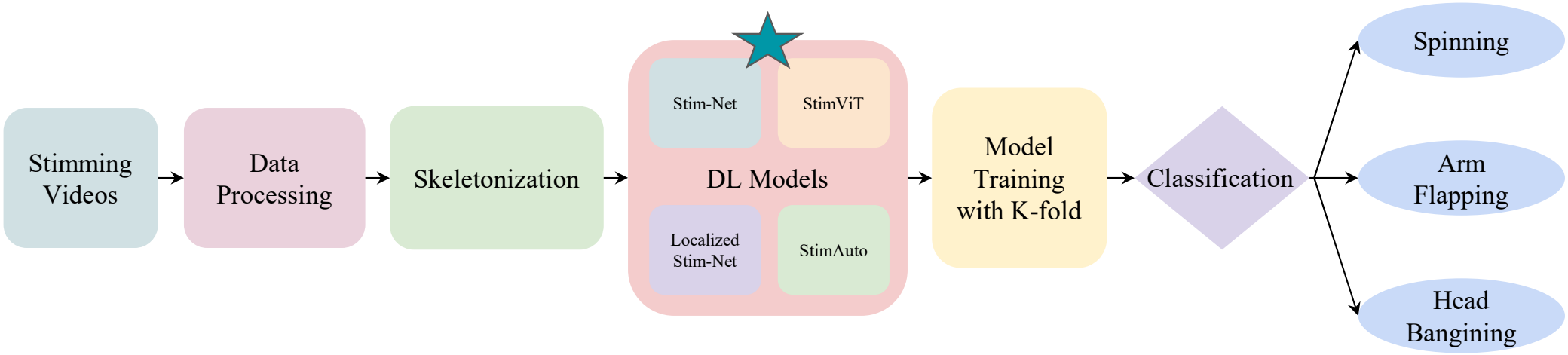


Number of Videos per Subject	Size of Dataset
1	92
3	276
5	460
10	920
15	1380

Original Video → **Scale** → **Translate** → **Rotate** → Augmented video

Architectures & Methods

High Level Workflow



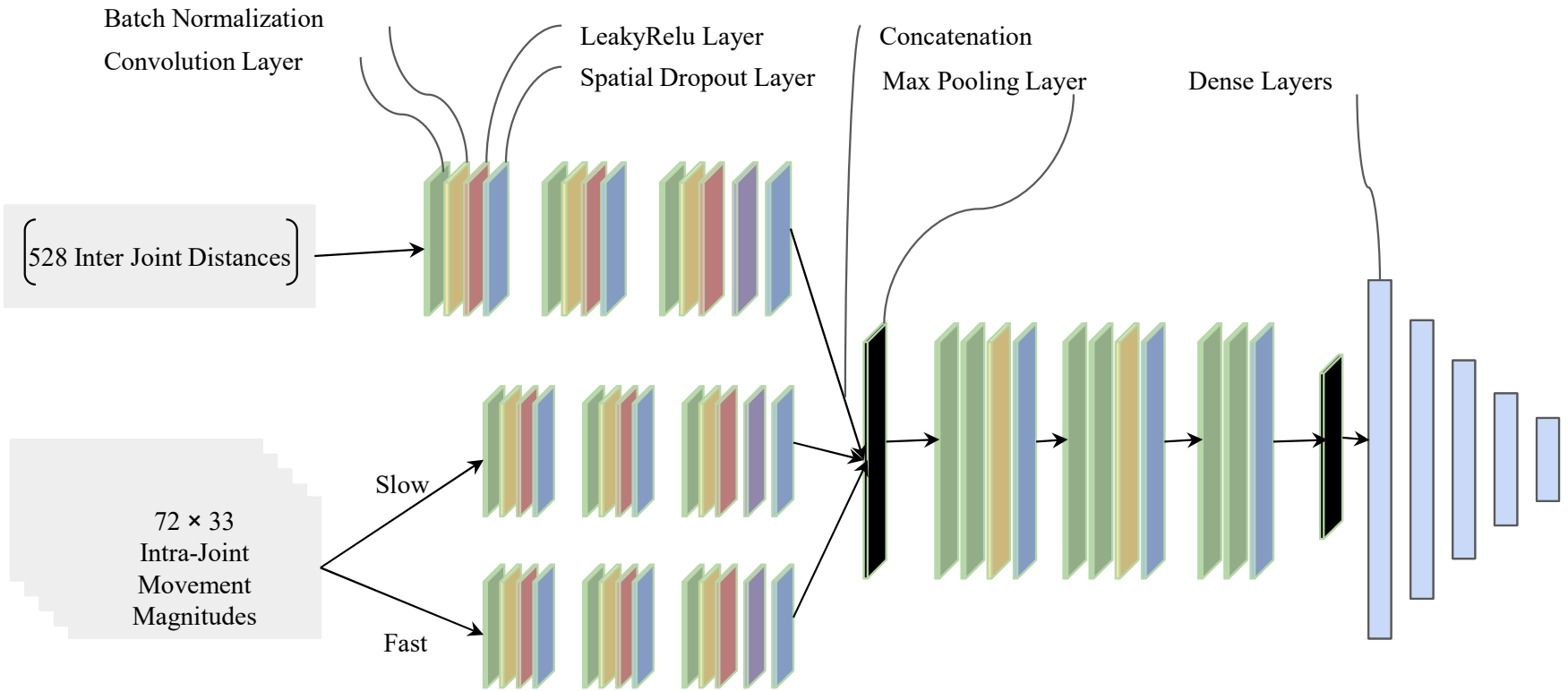
Methodologies

We have developed four models with different architectures and techniques.

1. **Stim-Net:** CNN-based Temporal Pose Differentiation Model
2. **LocalStim-Net:** CNN-based Localized Temporal Pose Differentiation Model
3. **StimAuto:** Autoencoder-Based Stimming Action Recognition Model
4. **StimViT:** Vision Transformer-Based Action Recognition Model.[10]

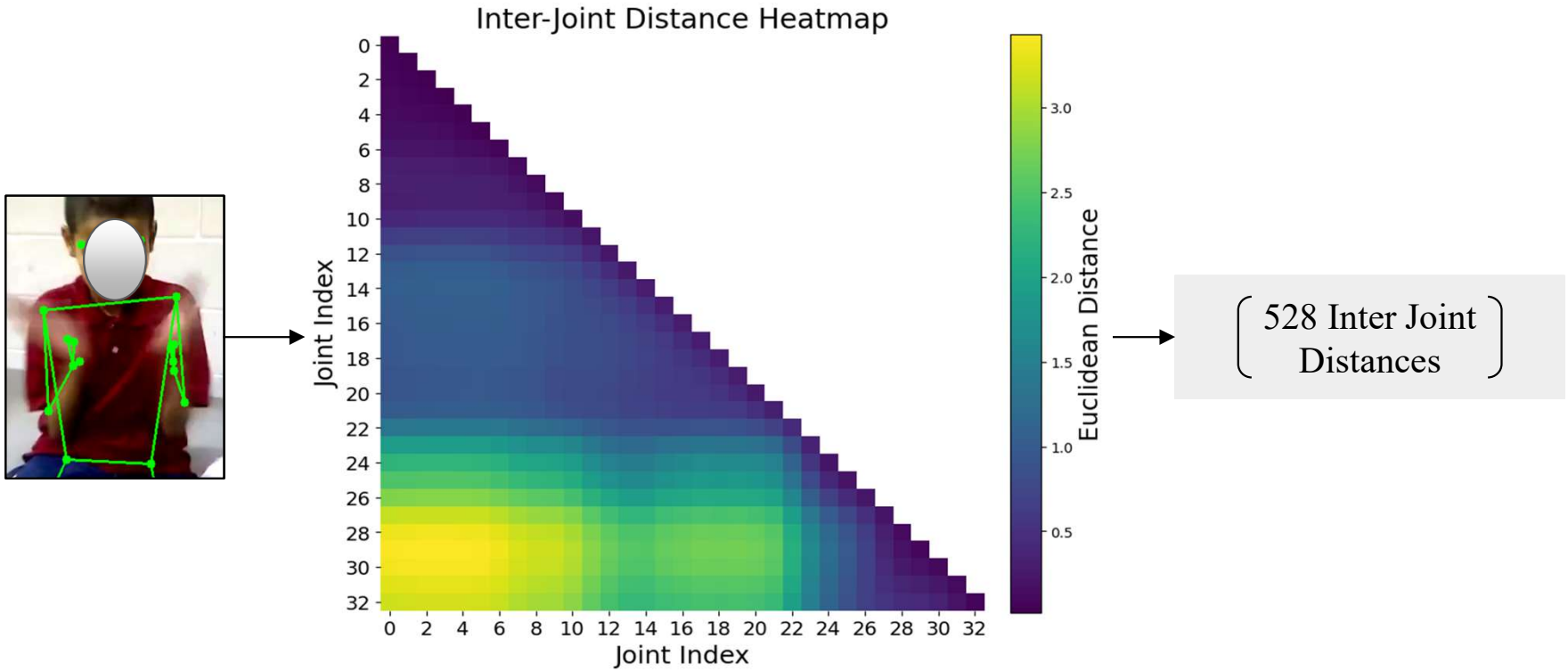
Stim-Net Model

Stim-Net: CNN-based Temporal Pose Differentiation Model (Yang et al, 2020)



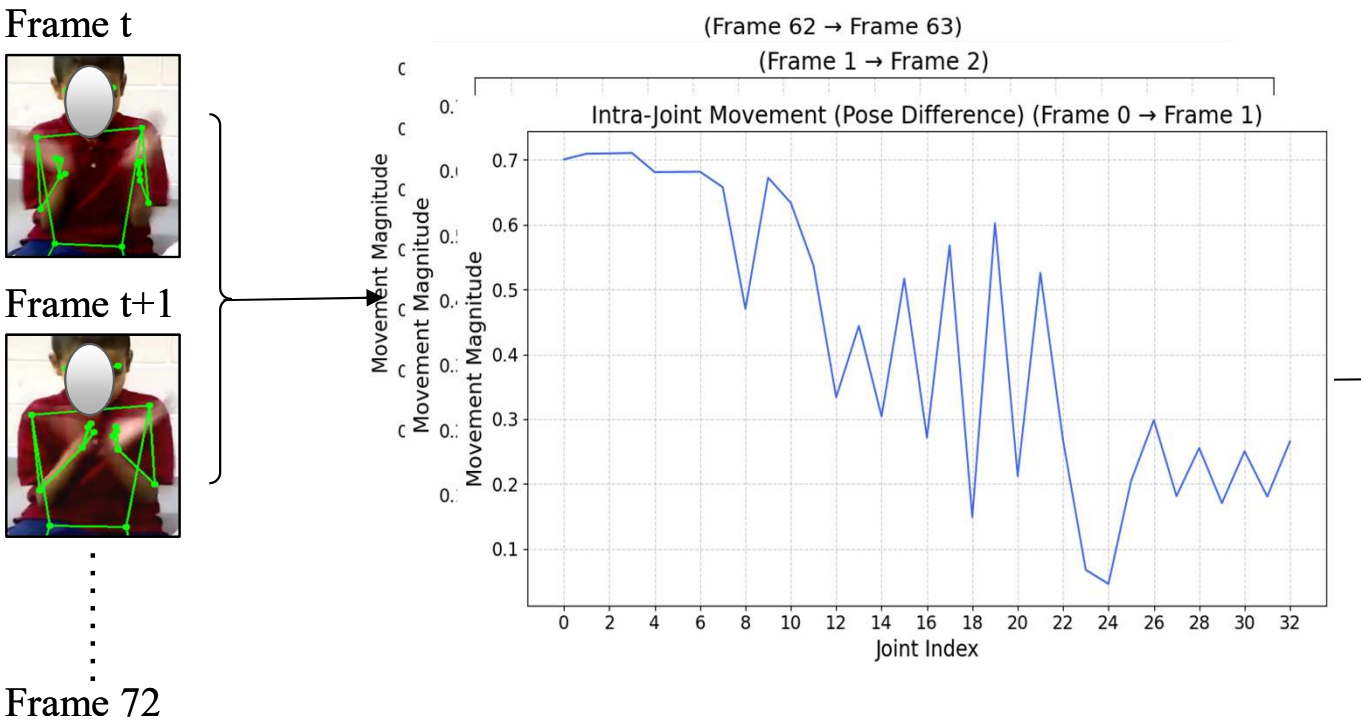
Data Preparation for Stim-Net Model

Stim-Net: CNN-based Temporal Pose Differentiation Model



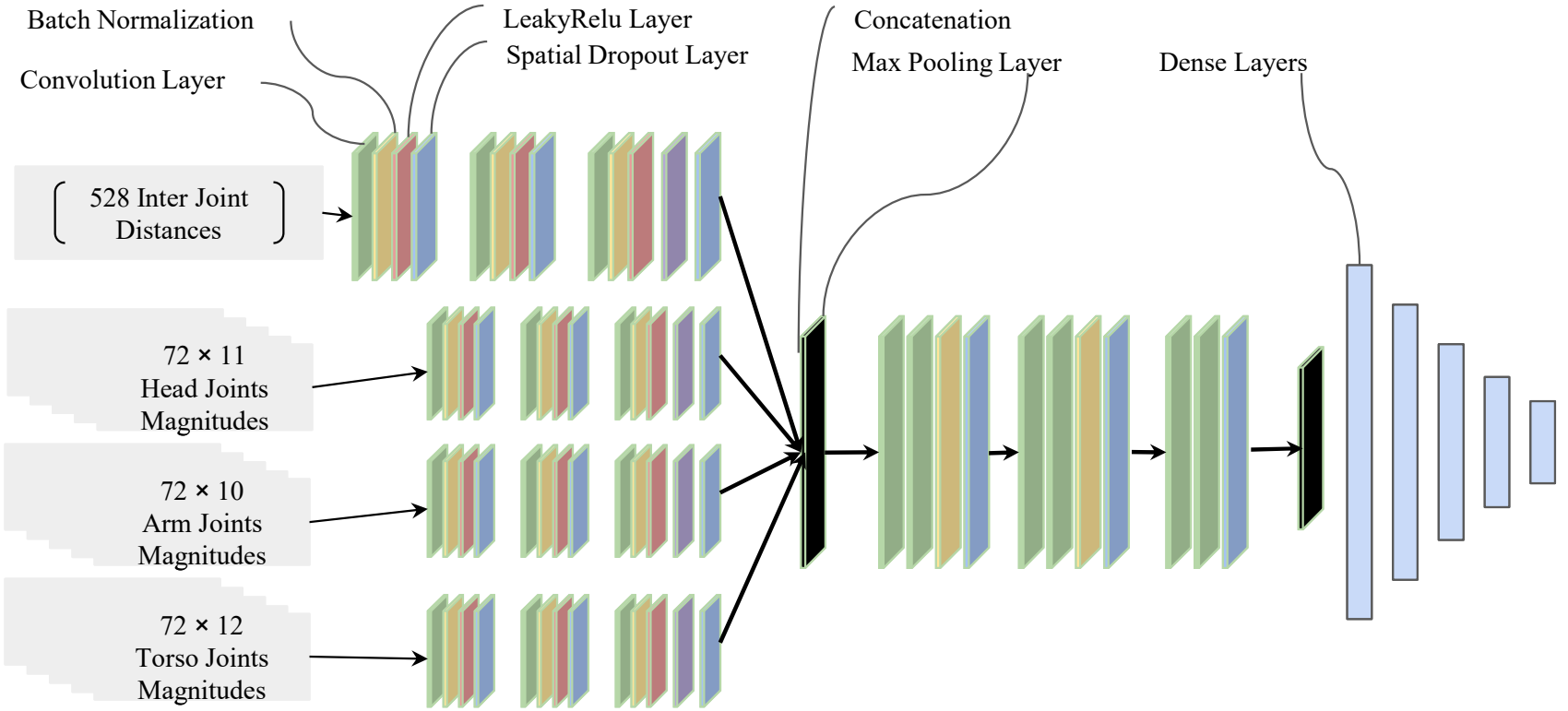
Data Preparation for Stim-Net Model

Stim-Net: CNN based Temporal Pose Differentiation Model



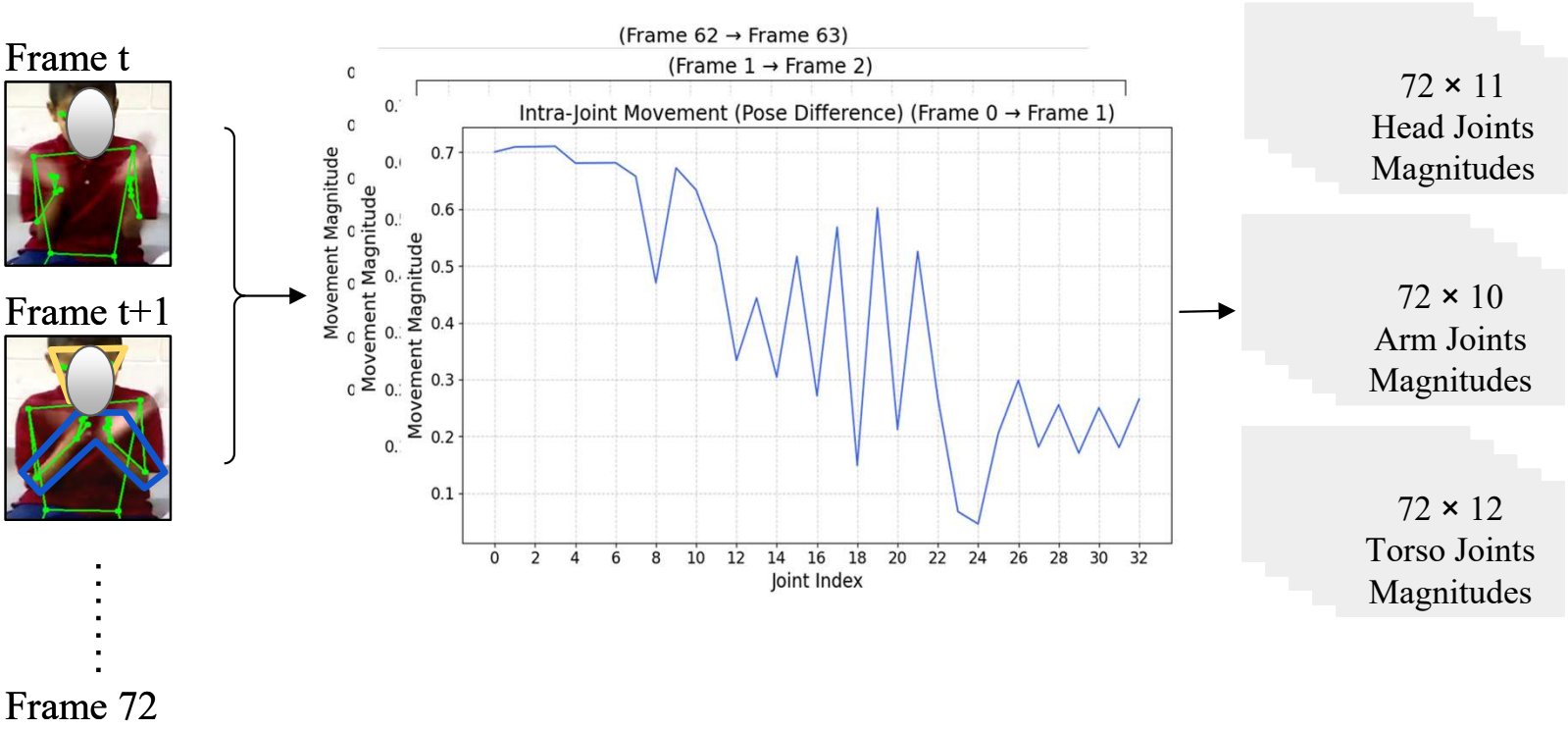
LocalStim-Net Model

LocalStim-Net: CNN-based Localized Temporal Pose Differentiation Model



Data Preparation for LocalStim-Net Model

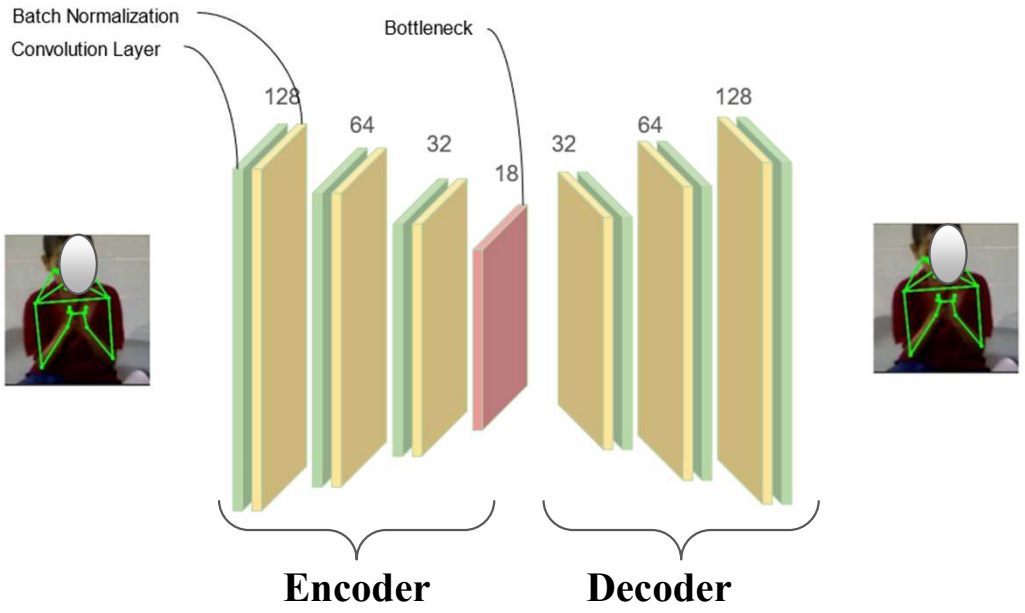
LocalStim-Net: CNN-based Localized Temporal Pose Differentiation Model



StimAuto Model

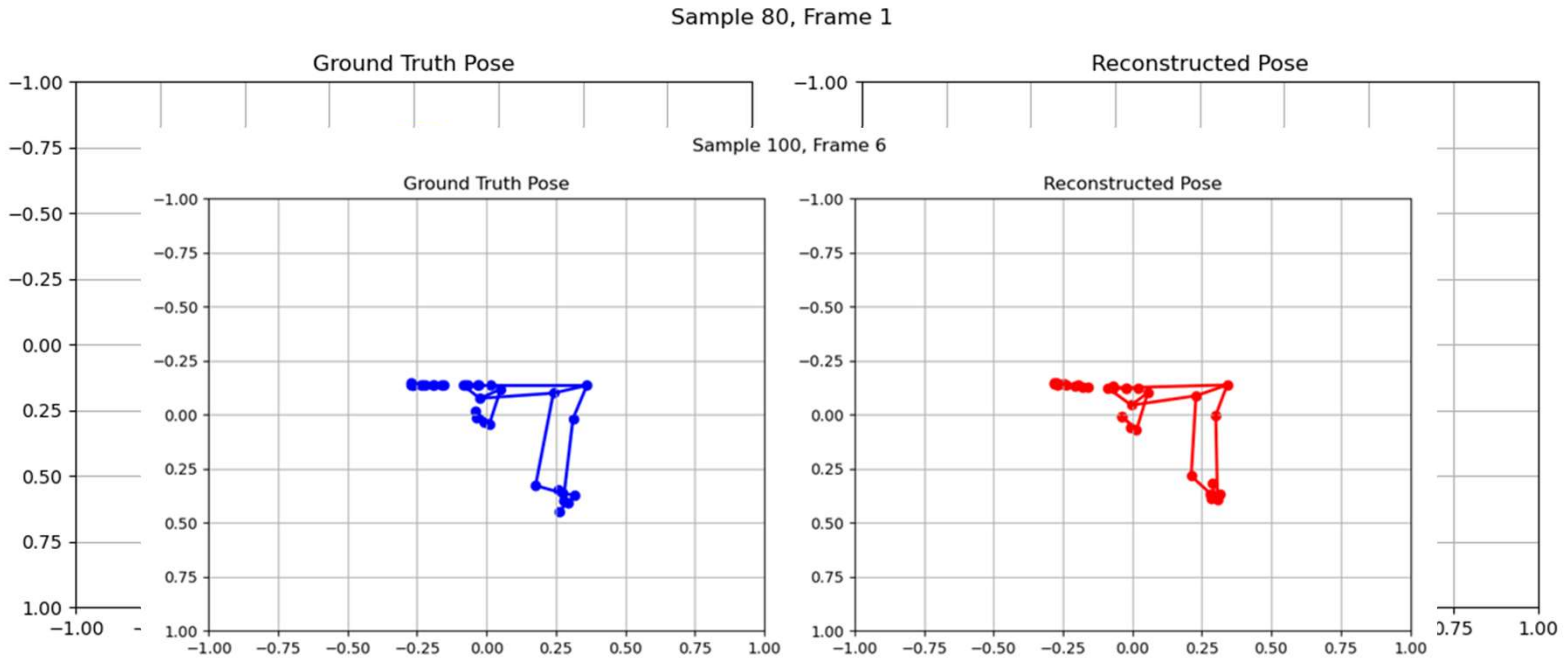
StimAuto: Autoencoder-Based Stimming Action Recognition Model

- ➔ Autoencoder was designed to learn compressed, noise-reduced representations of skeletal pose data.
- ➔ The encoder was trained with 2D convolutions and ReLU, while the decoder reconstructed input using mean squared error. (result shown in next slide)



StimAuto Auto Encoder

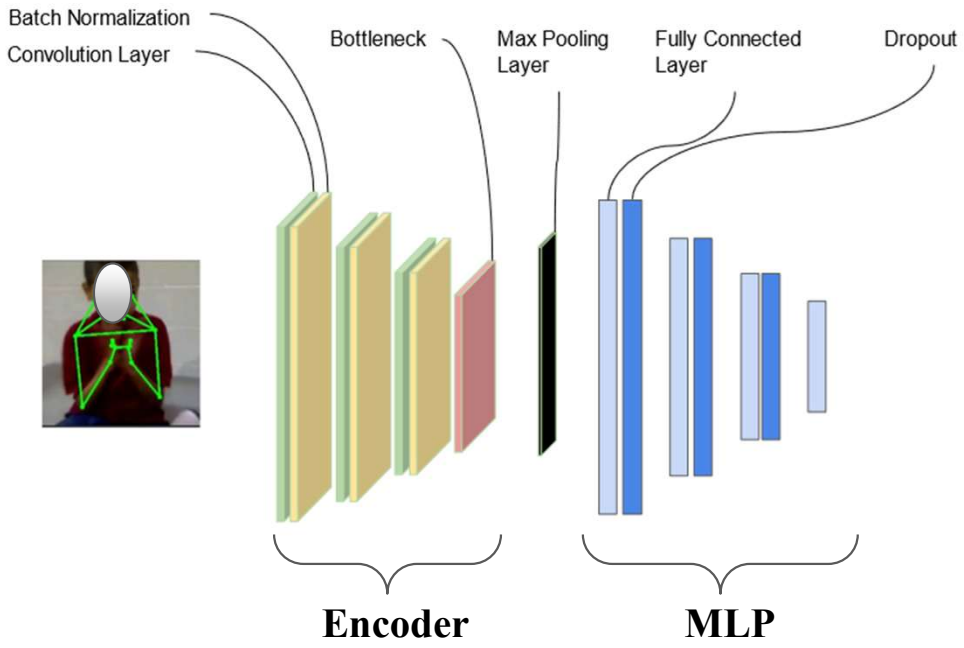
The Ground truth vs predicted skeleton by autoencoder



StimAuto Model

StimAuto: Autoencoder-Based Stimming Action Recognition Model

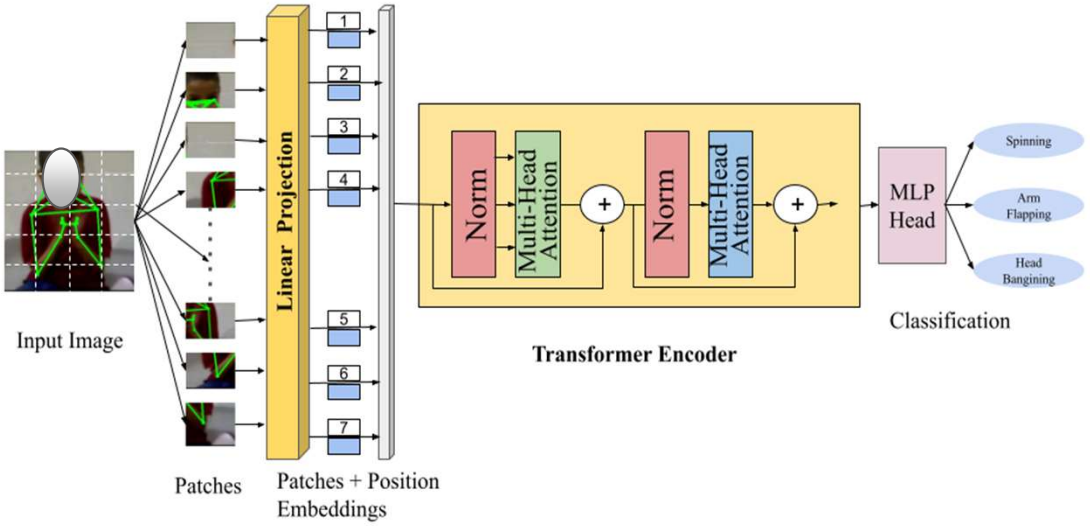
- ➔ Only the encoder was retained for feature extraction.
- ➔ Extracted features were pooled and passed to fully connected layers for classification.
- ➔ A SoftMax layer was used to predict one of three stimming action classes.



StimVit Model :Vision Transformer

StimViT: Vision Transformer Based Action Recognition Model (Dosovitskiy et al., 2020)

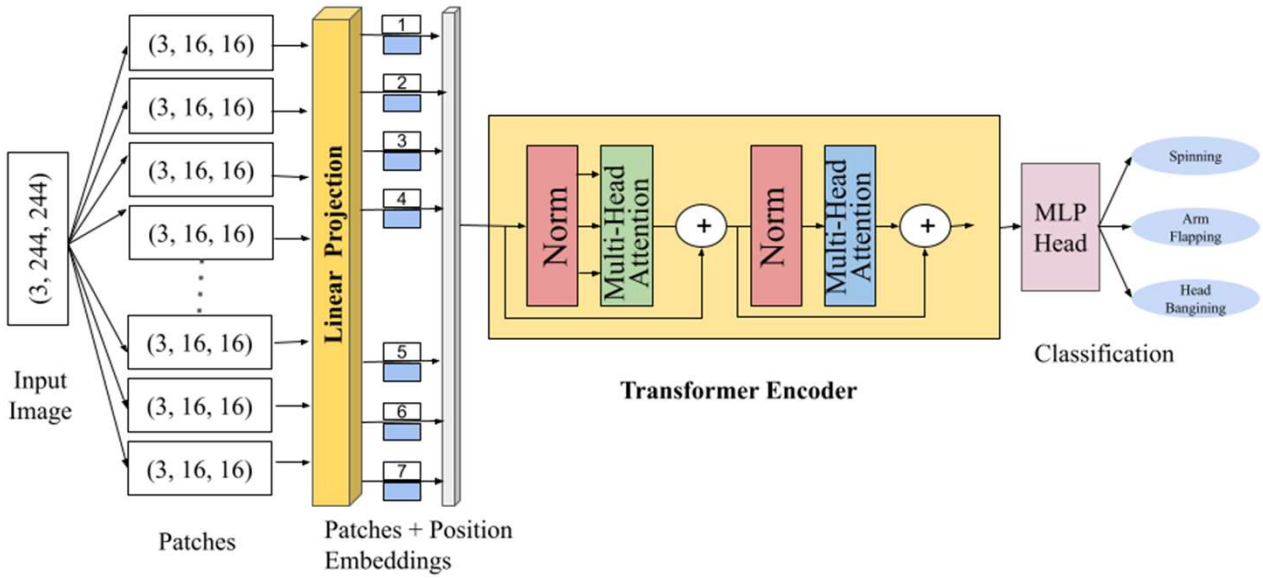
- ➔ A pre-trained Vision Transformer model originally designed for images was adapted to handle skeletal body movement data.
- ➔ The model uses self-attention, which allows it to focus on the most important frames in the sequence, instead of treating all frames equally.
- ➔ Multi-head attention means the model can look at the motion sequence in different ways at the same time, learning richer patterns (e.g., one head might focus on arm movement, another on head movement).



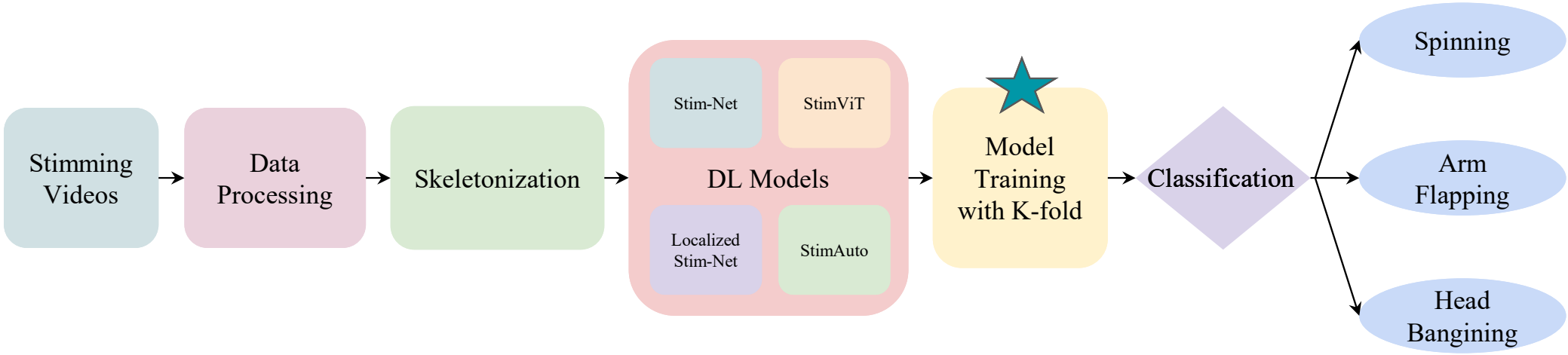
StimVit Model :Adaption to Skeletal Data

72 frames with 33 Joints (x,y,z) each frame is divided into patches of size 16
(reshape and resize (72, 33, 3) → (3,72, 33) →(3, 244, 244))

- ➔ Pose data was divided into smaller parts (patches), and each part was turned into a numeric summary (embedding).
- ➔ We fine-tuned only the higher layers of the model to learn patterns in stimming behavior while keeping the rest stable.



High Level Workflow



Training Configurations

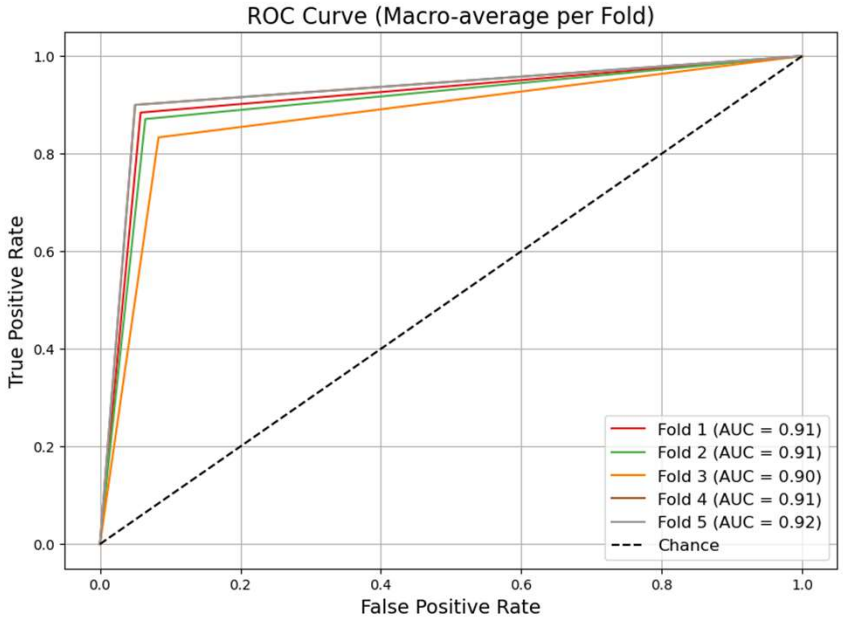
	Stim-Net	LocalStim-Net	StimAuto	StimViT
Input Type	Joint Distances	Joint Distances Arms/Head/Torso	Latent Embeddings (AE)	Flattened Joints (ViT format)
Framework	TensorFlow / Keras	TensorFlow / Keras	TensorFlow / Keras	PyTorch
Optimizer	Adam	Adam	Adam	AdamW
Learning Rate	1e-3(ReduceLROnPlateau)	1e-3(ReduceLROnPlateau)	1e-4(ReduceLROnPlateau)	1e-4 (fixed)
Loss Function	Categorical Cross Entropy	Categorical Cross Entropy	Categorical Cross Entropy	Cross Entropy Loss
Activation	LeakyReLU + Softmax	LeakyReLU + Softmax	LeakyReLU + Softmax	ReLU + Softmax
Batch Size	32	8	16	16
Epochs	300 (early stop, 20 patience)	256 (early stop, 20 patience)	256 (early stop, 20 patience)	10
Dropout	0.4	0.3	0.4	0.2

Results

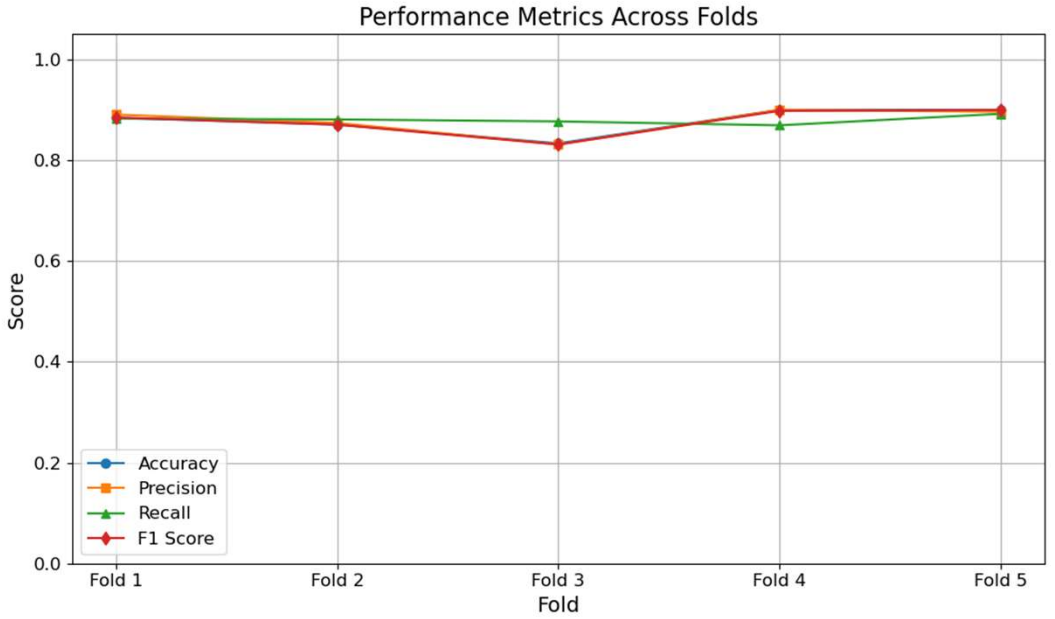
Results

- ➔ We measure the performance with plotting Confusion Matrix, Accuracy, Precision, Recall, F1-Score and ROC for all 4 models.
- ➔ The reported metrics are average of all 5 folds.

Model 1: Stim-Net Results



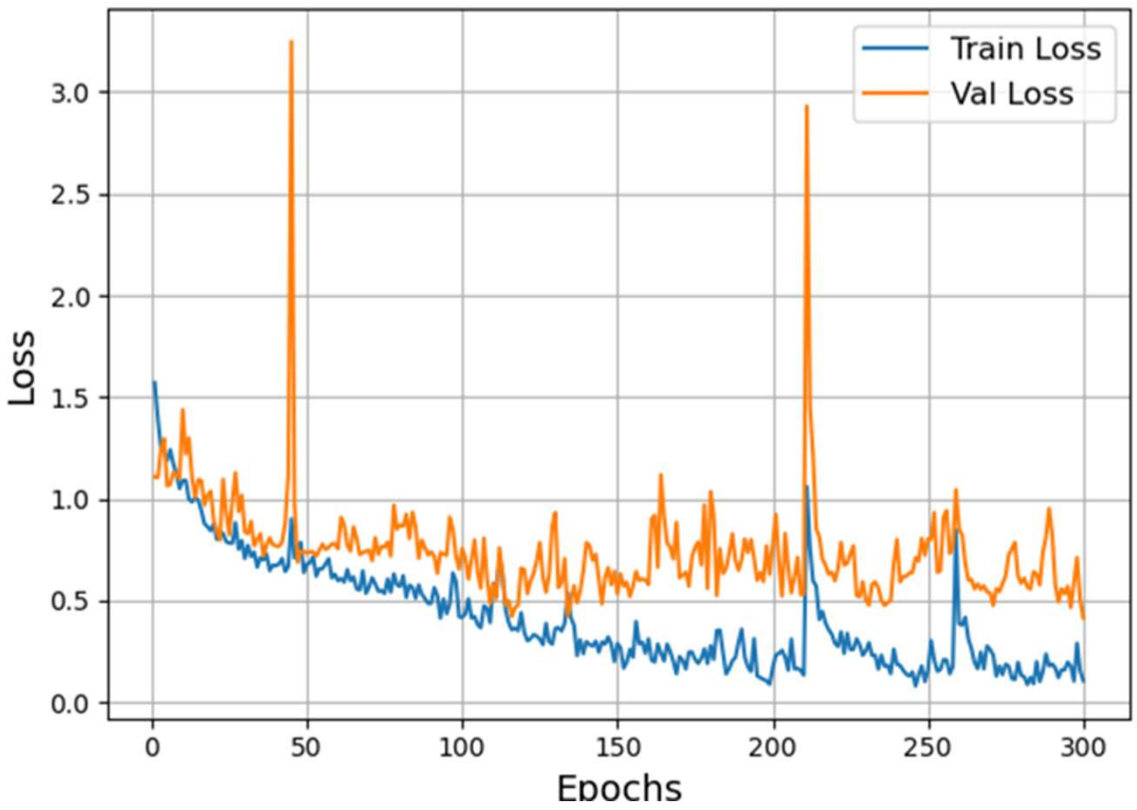
The Roc curve of 5 folds



The Accuracy across 5 folds

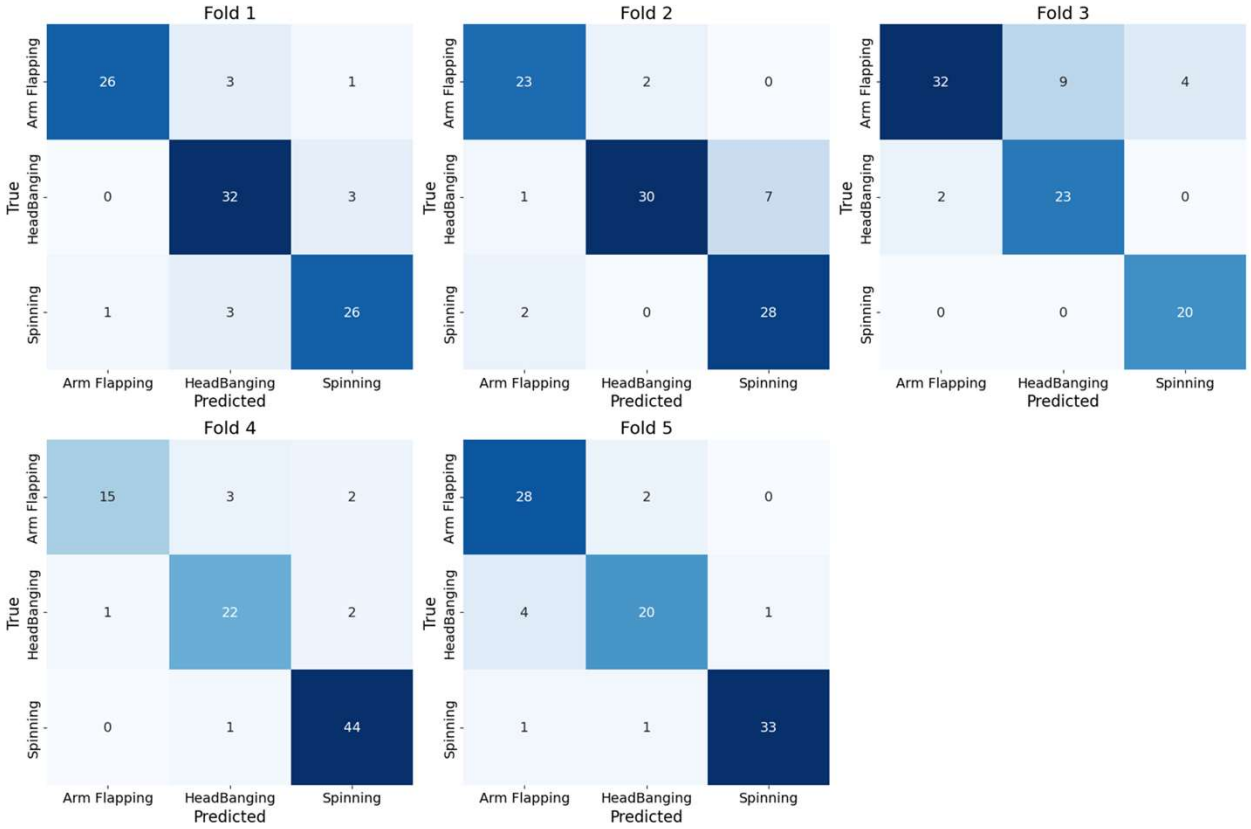
Model 1: Stim-Net Results

- Best model performance is achieved with data augmentation of 5 videos for subject.



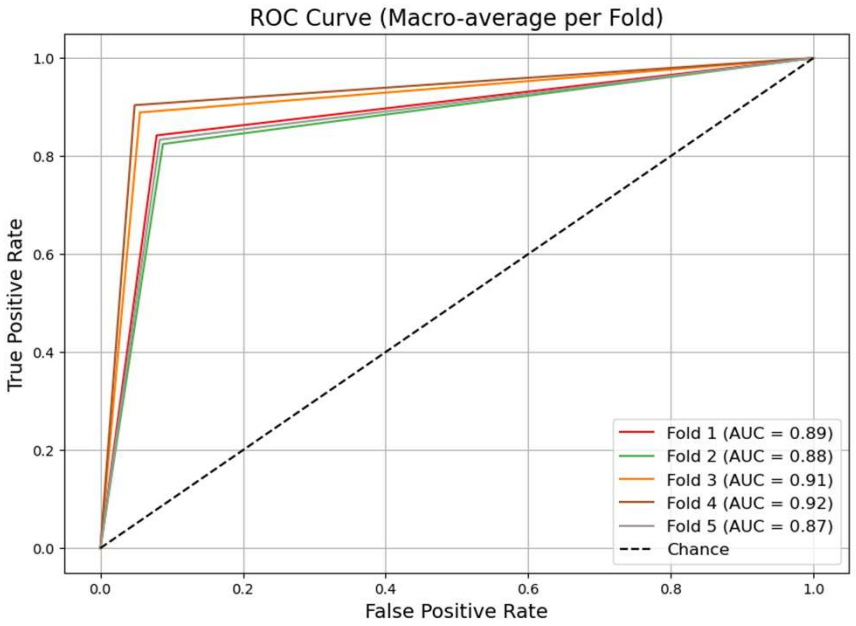
The Train vs Test accuracy of the Stim-Net

Model 1: Stim-Net Results



The confusion matrices of the 5-folds

Model 2: Localized Stim-Net Results

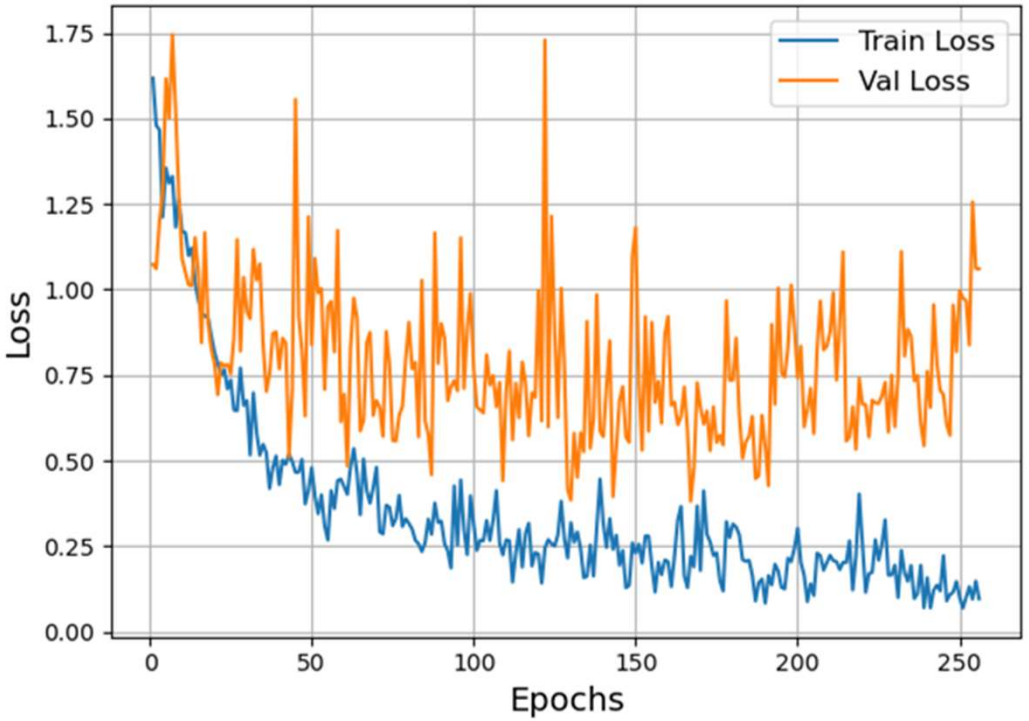


The Roc curve of 5 folds

The Accuracy across 5 folds

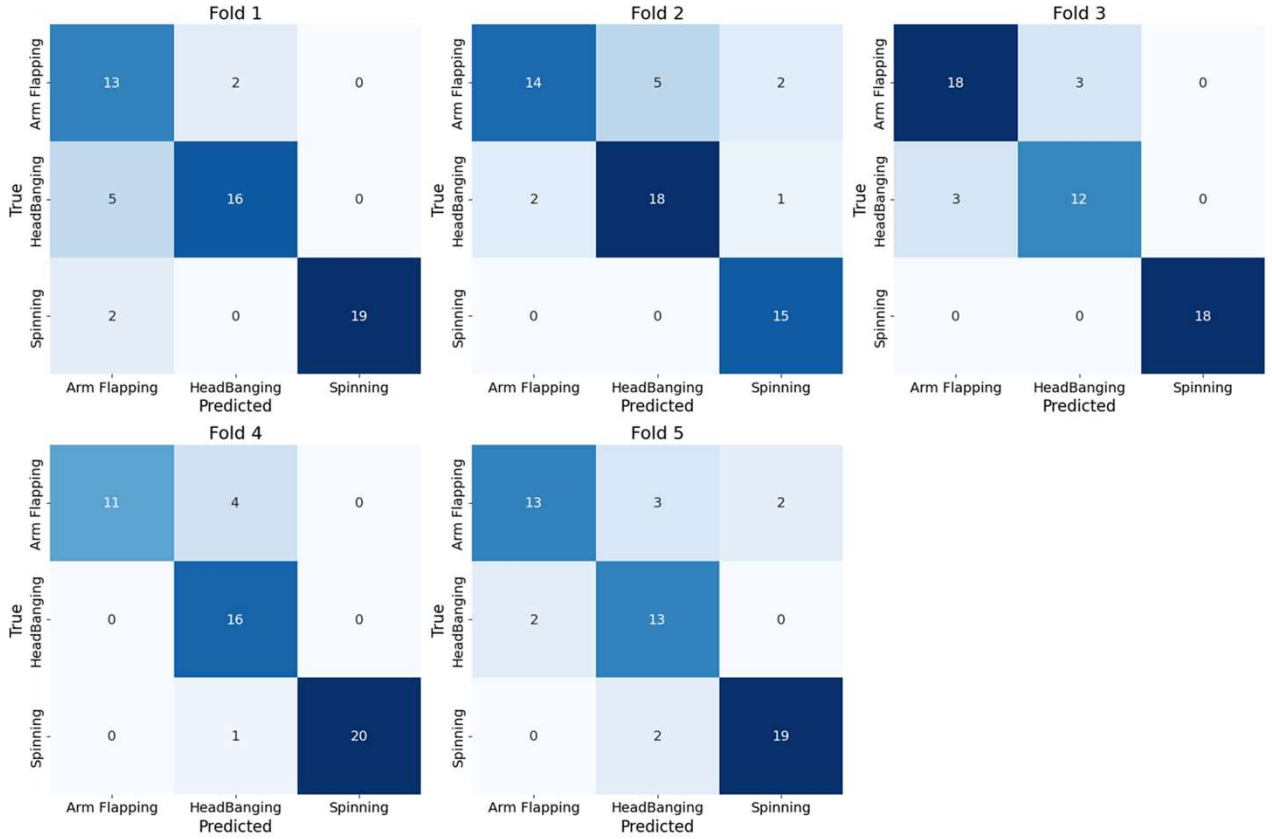
Model 2: Localized Stim-Net Results

- Best model performance is achieved with data augmentation of 3 videos for subject.



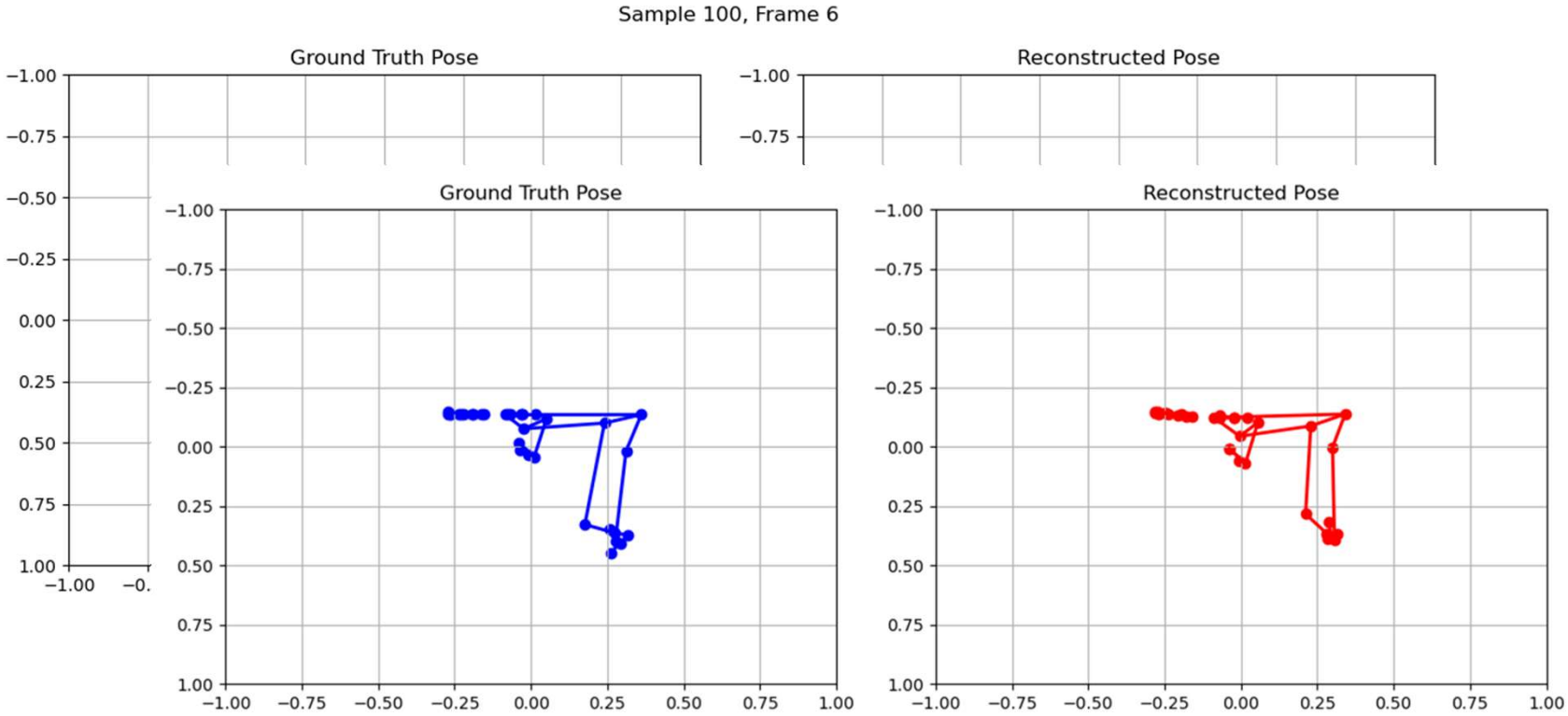
The Train vs Test Loss of the Localized Stim-Net model

Model 2: Localized Stim-Net Results



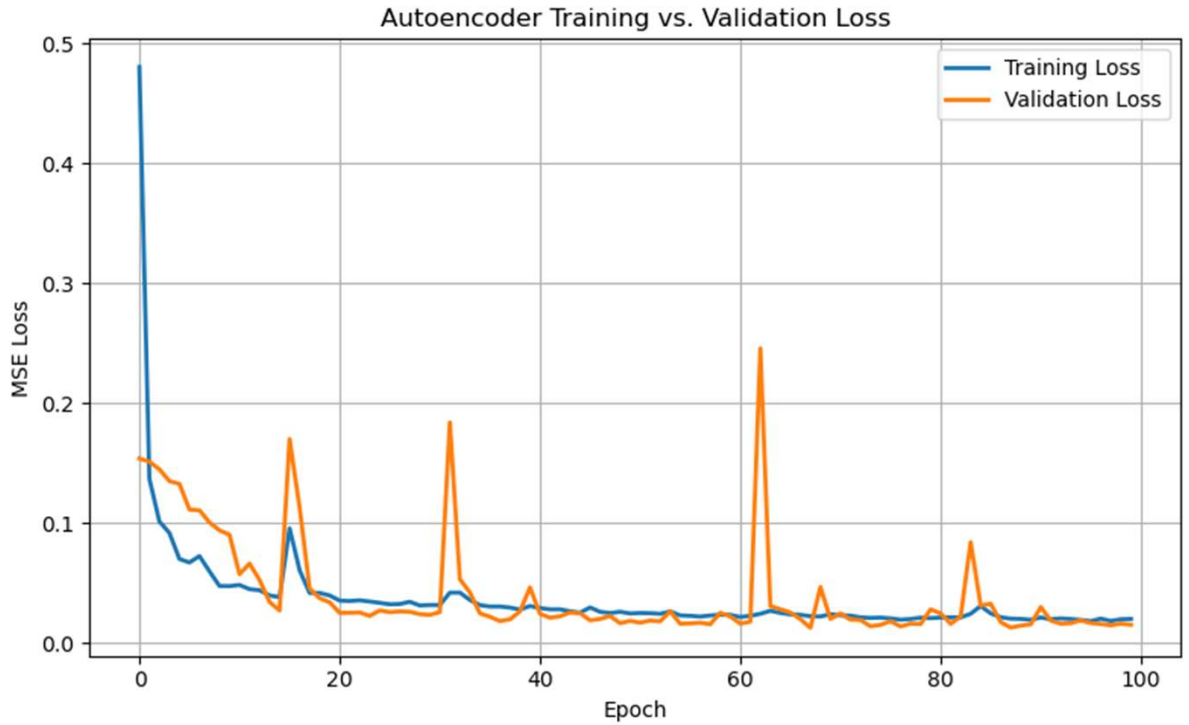
The confusion matrices of the 5-folds

Model 3: StimAuto Results



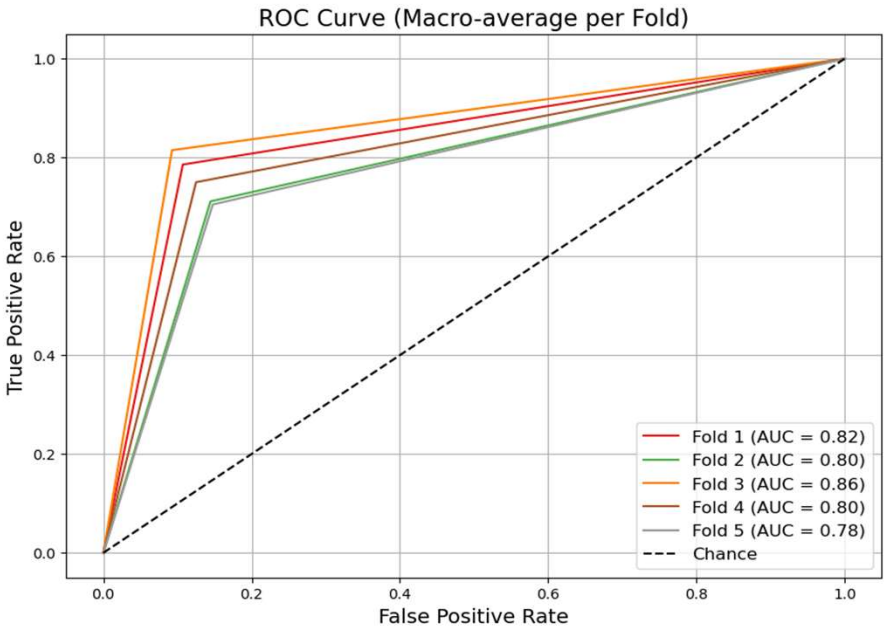
The Ground truth vs predicted skeleton by autoencoder

Model 3: StimAuto Results



The loss functions of trained autoencoder

Model 3: StimAuto Results

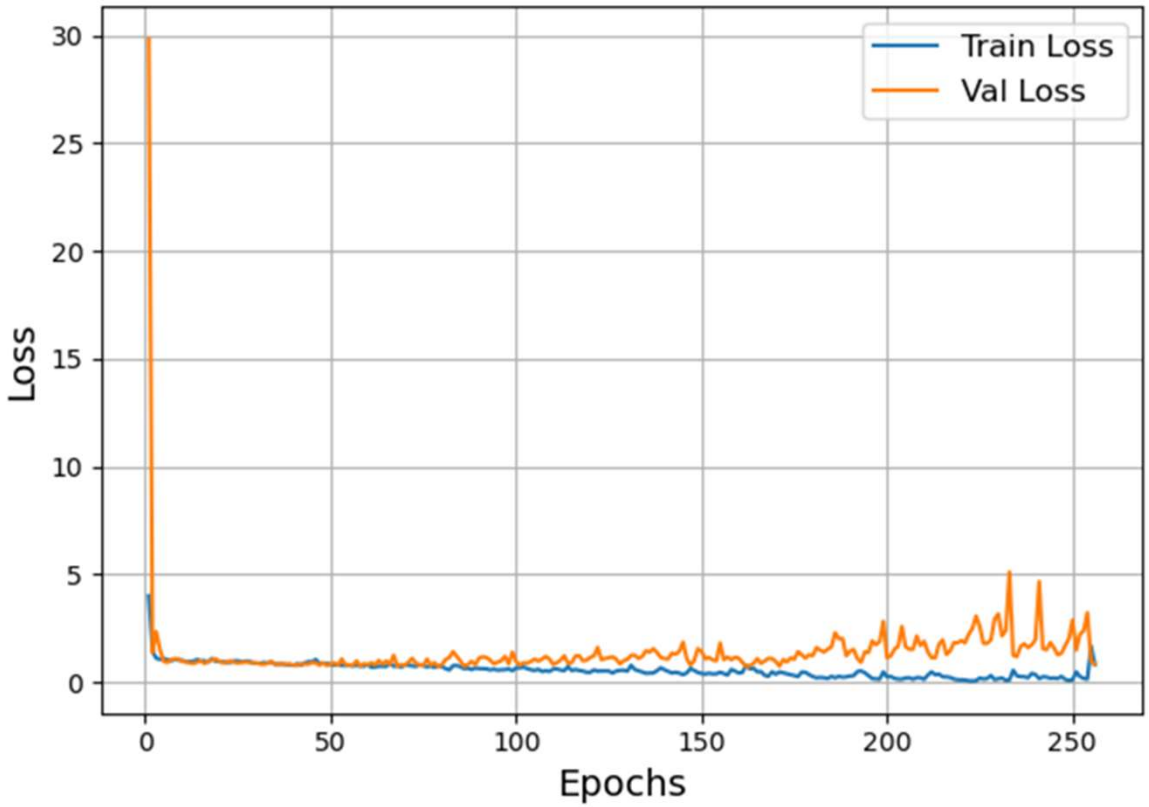


The Roc curve of 5 folds

The Accuracy across 5 folds

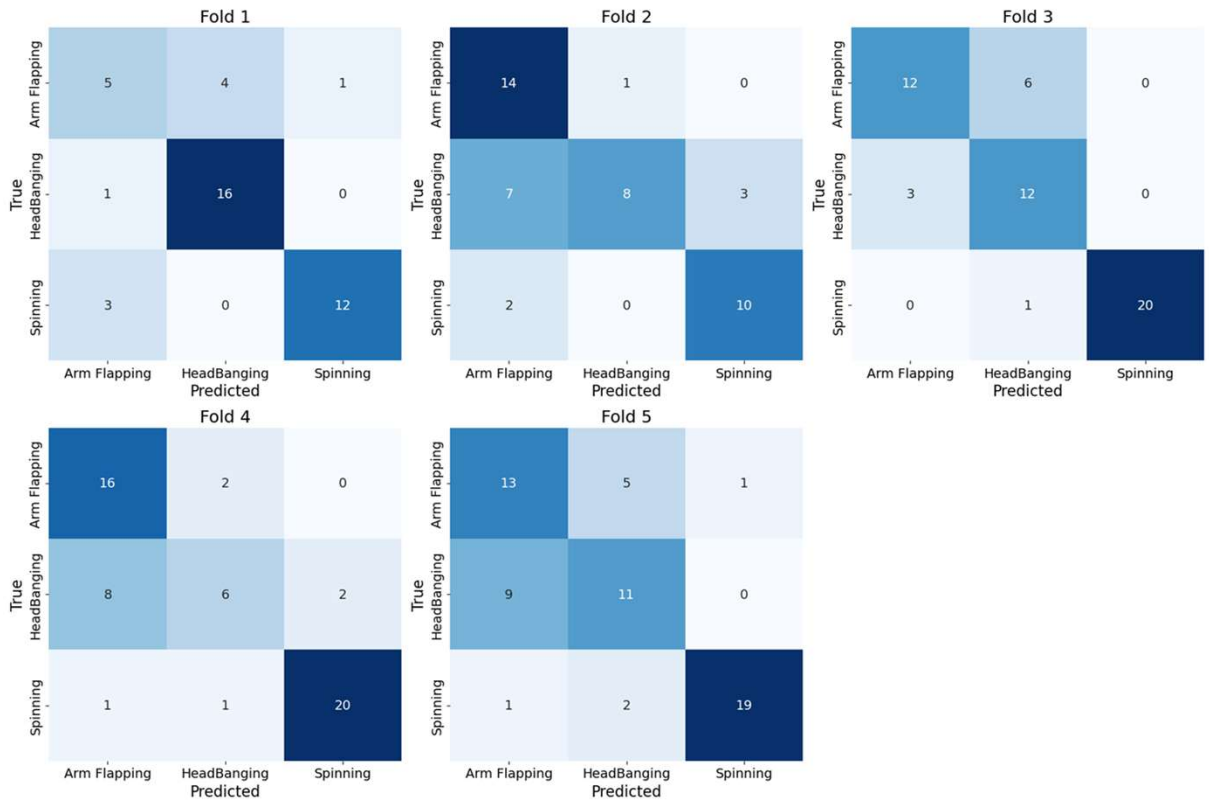
Model 3: StimAuto Results

- Best model performance is achieved with data augmentation of 10 videos for subject.



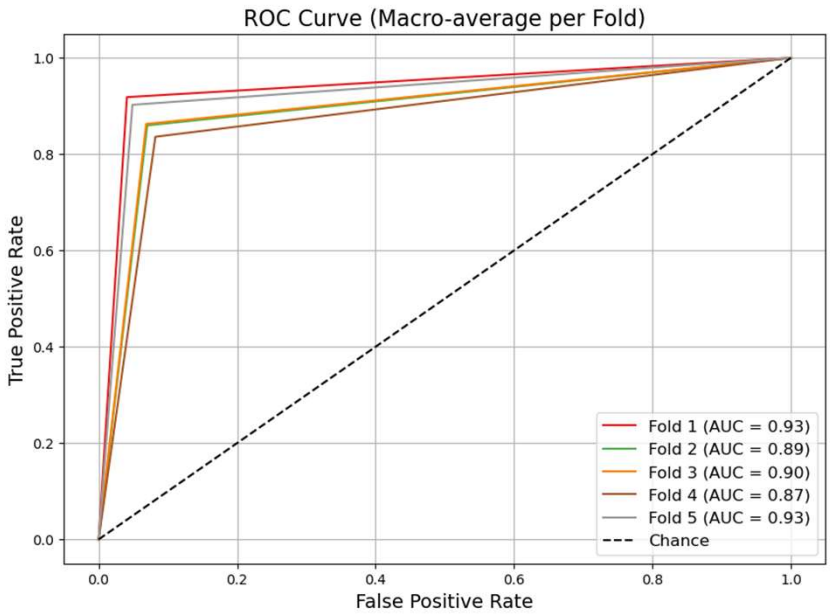
The Train vs Test Loss of StimAuto

Model 3: StimAuto Results



The confusion matrices of the 5-folds

Model 4: StimViT Results

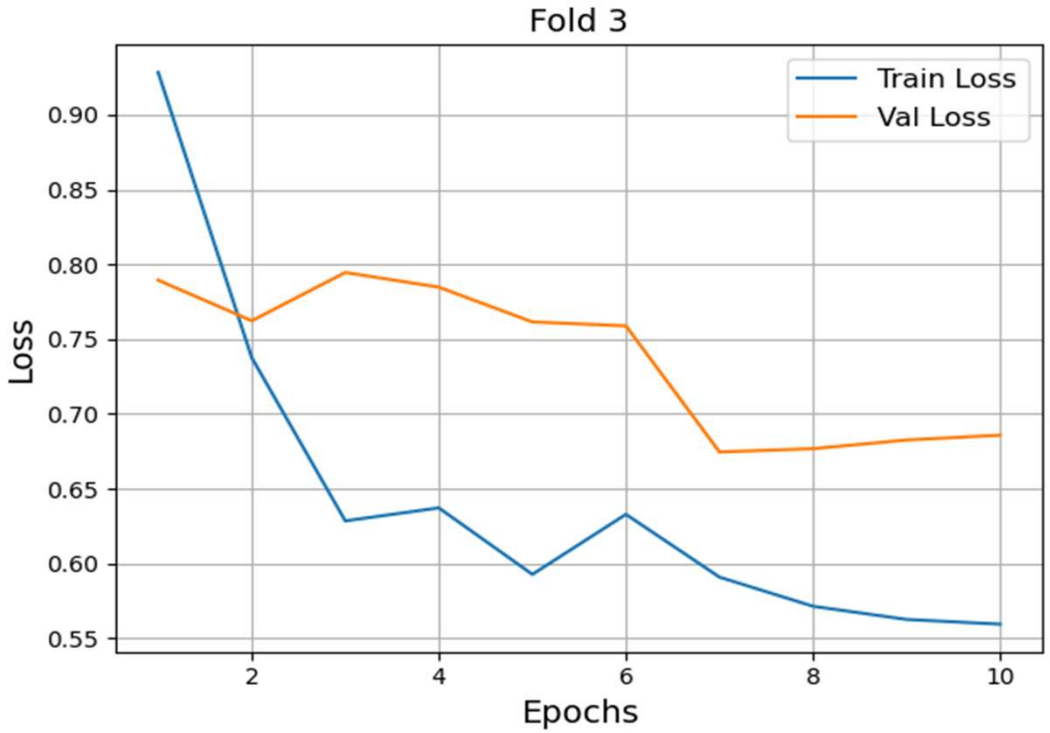


The Roc curve of 5 folds

Accuracy across 5 folds

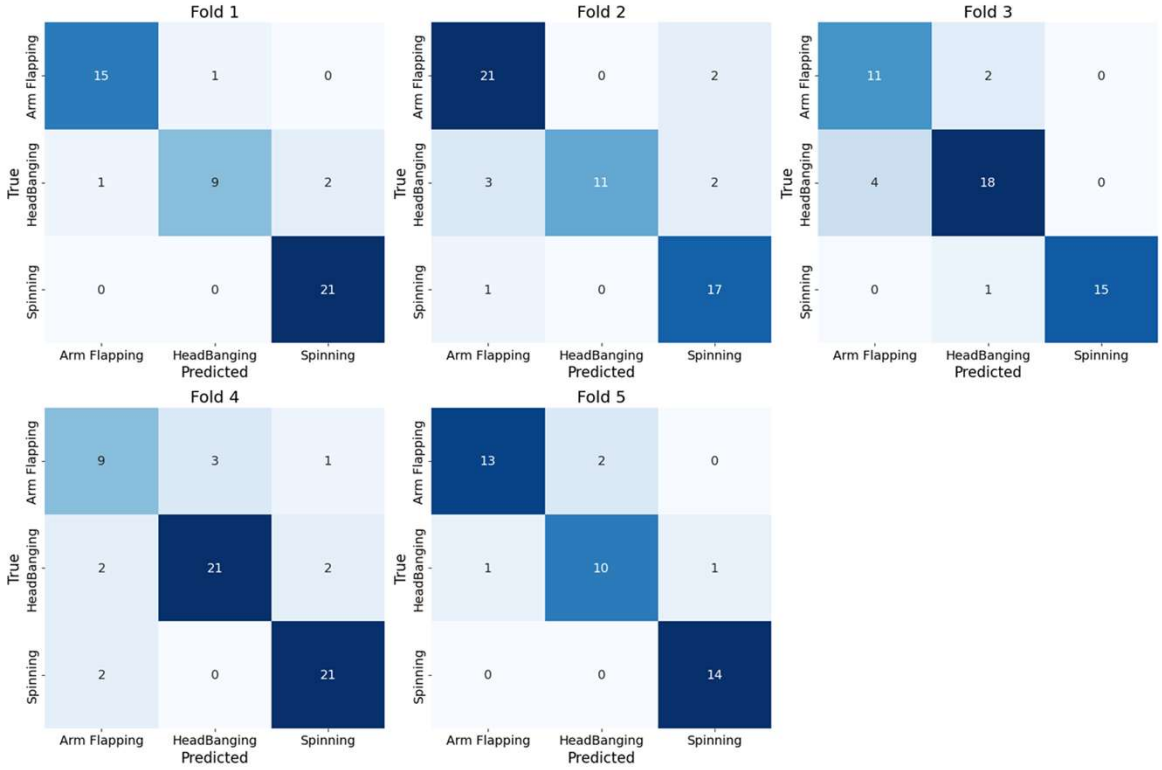
Model 4: StimViT Results

- Best model performance is achieved with data augmentation of 5 videos for subject.



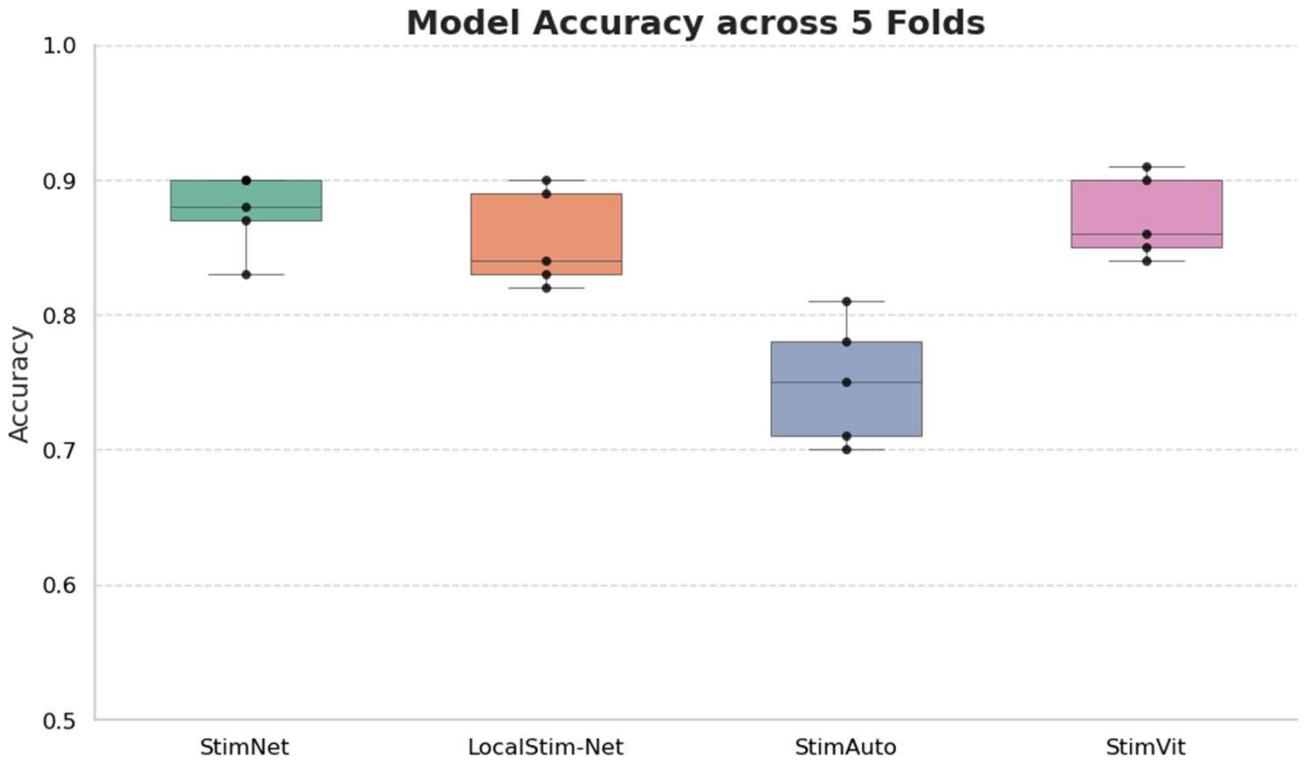
The Train vs Test Loss of the Stim-ViT model

Model 4: StimViT Results



The confusion matrices of the 5-folds

Performance Comparison of All Models



The accuracy box plot of 4 models across 5 folds

Performance Comparison of All Models

S.No	Model	Average Accuracy (5-fold)	Average F1 Score (5-fold)	Average AUC (5-fold)
1	Stim-Net	87.77%	87.69%	90.95%
2	LocalStim-Net	85.85%	85.85%	89.51%
3	StimAuto	75.28%	73.96%	76.43%
4	StimVit	87.2%	87.48%	90.41%

The mean accuracy, F1 score and Auc comparison of 4 models

Comparison of Models

Model	Architecture Type	Best Suited For
Stim-Net	CNN-based Temporal Pose Differentiation	General-purpose recognition of rhythmic, full-body stimming behaviors such as spinning and arm flapping. Fast, efficient, and suitable for real-time use.

Comparison with Existing Studies

Name of Paper	Methodology	Objective	Accuracy	Dataset	Classes
Rajagopalan et al.	SVM	Classify Stimming Actions	50.70%	SSBD	3
Negin et al.	YOLOv3 + HOF + MLP	Classify Stimming Actions	78%	ASD diagnostic centers(private)	3
Ali et al.	3D-CNN	Classify Stimming Actions	75.62%	SSBD	3
Wei et al.	LSTM, TCN, MS-TCN	Classify Stimming Actions	83%	SSBD(Extended)	3

Conclusion

- ➔ Proves that the skeletal data can significantly increase the performance.
- ➔ First to integrate Vision Transformers (ViTs) and Autoencoders into stimming action recognition, moving beyond CNN and LSTM-based approaches.
- ➔ CNNs with inter and intra joint distance features gave the best accuracy.
- ➔ Fine tuning vision transformer gives the good accuracy but there is trade of with latency and size.
- ➔ Data augmentation improved the performance and helped model convergence.

Future Direction

- ➔ Automated methods for detecting and cropping the child before extracting skeletal data.
- ➔ Expanding the dataset, particularly for the Hand Action class.
- ➔ Increasing the dataset size for underrepresented classes would enhance the model's ability to generalize across different individuals and environments.
- ➔ Using pre trained autoencoders and fine tune on Stimming data.
- ➔ Real-time application and testing.
- ➔ Ensemble with other models like facial emotion recognition for better harmful stimming detection.

Thank You!

Questions?

References

1. Rajagopalan, S. S., Dhall, A., & Goecke, R. (2013). Self-Stimulatory Behaviours in the Wild for Autism Diagnosis. *2013 IEEE International Conference on Computer Vision Workshops*, 755–761. https://doi.org/10.1109/ICCVW.2013.103
2. Zamzmi, G., Pai, C. Y., Goldgof, D. B., Kasturi, R., Ashmeade, T., & Sunshine, J. (2018). Automated video-based assessment of infants with perinatal stroke. *Journal of Medical Imaging*, *5*(2), 024501. https://doi.org/10.1117/1.JMI.5.2.024501
3. 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. https://doi.org/10.1371/journal.pmed.1002705
4. Negin, F., Ozyer, B., Agahian, S., Kacdioglu, S., & Ozyer, G. T. (2021). Vision-assisted recognition of stereotype behaviors for early diagnosis of Autism Spectrum Disorders. *Neurocomputing*, *446*, 145–155. https://doi.org/10.1016/j.neucom.2021.03.004
5. Washington, P., Kline, A., Mutlu, O. C., Leblanc, E., Hou, C., Stockham, N., Paskov, K., Chrisman, B., & Wall, D. (2021). Activity Recognition with Moving Cameras and Few Training Examples: Applications for Detection of Autism-Related Headbanging (p. 7). https://doi.org/10.1145/3411763.3451701
6. Ali, A., Negin, F., Thümmler, S., & Bremond, F. (2022). Video-based Behavior Understanding of Children for Objective Diagnosis of Autism: Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 475–484. https://doi.org/10.5220/0010839200003124
7. Wei, P., Ahmedt-Aristizabal, D., Gammulle, H., Denman, S., & Armin, M. A. (2023). Vision-based activity recognition in children with autism-related behaviors. *Heliyon*, *9*(6), e16763. https://doi.org/10.1016/j.heliyon.2023.e16763
8. Serna-Aguilera, M., Nguyen, X. B., Singh, A., Rockers, L., Park, S., Neely, L., Seo, H., & Luu, K. (2024). Video-Based Autism Detection with Deep Learning (arXiv:2402.16774). arXiv. http://arxiv.org/abs/2402.16774
9. Yang, F., Wu, Y., Sakti, S., & Nakamura, S. (2020). Make Skeleton-based Action Recognition Model Smaller, Faster and Better. Proceedings of the 1st ACM International Conference on Multimedia in Asia, 1–6. https://doi.org/10.1145/3338533.3366569
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929