Detection of Autism Spectrum Disorder Using Deep Learning

Thesis by

Manjeet Singh

In Partial Fulfillment of the Requirements for the
Degree of
Bachelors of Technology
in
Computer Science & Engineering
with Specialization
Data Science and Artificial Intelligence

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY RANCHI
Ranchi, Jharkhand

2025
Defended 3rd Devember 2024

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, **Dr. Nidhi Khushwaha** *(Faculty, Department of Computer Science and Engineering, Indian Institute of Information Technology Ranchi)* for her invaluable guidance, encouragement, and support throughout the course of this project. Her expertise and insights were instrumental in shaping this research and ensuring its successful completion.

Additionally, I extend my heartfelt appreciation to all those who contributed directly or indirectly to this work. Your support has been deeply appreciated.

ABSTRACT

Timely autism diagnosis is crucial for effective intervention, yet current processes are often slow and inaccessible. This study explores the use of machine learning for detecting abnormal movements, such as hand flapping, head banging and spinning from unstructured home videos to aid early autism diagnosis. Utilizing the Self-Stimulatory Behavior Dataset (SSBD) [7], we developed privacy-preserving models by converting video frames into hand landmark coordinates using MediaPipe [8] and extracting convolutional features with ResNet[2], Vision Transformer (ViT) [1], and Swin Transformer [6] architectures. These models were trained and evaluated, with ResNet[2] and Swin Transformer [6] achieving the highest performance. Our findings demonstrate the feasibility of integrating state-of-the-art deep learning models into accessible diagnostic tools, offering a scalable approach to detecting autism-related behaviors.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Number *Page*

Chapter 1

INTRODUCTION

1.1 Motivation

Autism Spectrum Disorder (ASD) is a developmental condition that is often diagnosed late, delaying critical interventions. Early detection of autism-related behaviors, such as self-stimulatory actions (e.g., hand flapping, head banging or spinning), is essential, as it allows for timely identification that can significantly improve long-term outcomes for children. Traditional diagnostic methods typically require in-person assessments by clinicians, which can be time-consuming, expensive, and out of reach for many families, particularly in impoverished regions.

With advancements in machine learning and computer vision, there is an emerging opportunity to address these challenges by automating the detection of autism-related behaviors. Video-based analysis can offer a non-invasive, scalable solution, enabling continuous monitoring of children in natural settings such as their homes. However, current methods of detecting these behaviors from videos have limitations, particularly regarding privacy concerns, model generalization, and computational efficiency.

The motivation behind this research is to create a more accessible and privacy-conscious tool for early autism detection. By utilizing deep learning models to analyze unstructured home videos, this study seeks to bridge the gap between the need for early intervention and the accessibility of diagnostic tools, providing an efficient and scalable solution for autism detection.

1.2 Objectives

This study aims to explore the use of machine learning techniques for the automated detection of abnormal hand movements, such as hand flapping, head banging and spinning, which are commonly observed in children with autism. Specifically, the objective is to evaluate the feasibility of using deep learning models, including ResNet[2], Vision Transformer (ViT) [1], and Swin Transformer [6], for detecting these behaviors from unstructured home videos. The specific goals of this research are:

1. **Develop a Privacy-Preserving Framework:** To ensure that sensitive per-

sonal information is not exposed, the study proposes converting video frames into hand landmark coordinates using MediaPipe [8], a privacy-preserving approach that avoids the use of facial or body images.

2. **Explore Advanced Deep Learning Models:** This research will compare the effectiveness of MediaPipe [8], MobileNet [5], ResNet [2], and Transformer models with lstms [3] in detecting hand flapping behaviors, focusing on their ability to generalize across diverse datasets and environments.

3. **Build a Scalable Solution:** The aim is to develop a solution that can be implemented on mobile devices, making the model scalable and easily accessible to a larger population, including those in rural or underserved areas.

4. **Evaluate and Compare Performance:** The models will be evaluated using metrics such as accuracy, precision, recall, and F1 score, and compared with traditional methods to assess their potential for real-world deployment.

Chapter 2

LITERATURE SURVEY

Recent studies have investigated the use of machine learning for autism diagnosis, particularly through the analysis of behavioral patterns. Various approaches, including wearable devices and mobile applications, have been employed to capture self-stimulatory behaviors such as hand flapping, head banging, and repetitive motions. Computer vision techniques, particularly pose estimation and motion tracking, have also been explored to detect these behaviors from video data. Existing methods such as the use of convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) [3] have shown promising results. However, these methods often rely on fixed and predefined feature extraction techniques, limiting their ability to generalize across diverse datasets and environments. Some studies, including those using MediaPipe [8] for hand landmark detection and MobileNet [5]for feature extraction, have achieved moderate success in detecting abnormal hand movements. This literature highlights the need for more robust, flexible, and privacy-preserving methods to analyze unstructured video data for autism diagnosis.

Chapter 3

RESEARCH GAP

Although previous research has explored the detection of autism-related behaviors through machine learning, there remain several gaps:

- **Limited Privacy Preservation:** Many existing methods rely on full-face or body images, raising concerns about privacy and the potential for misuse of personal data.
- **Model Generalization:** Existing models often fail to generalize across different age groups, environments, or video qualities, limiting their practical use in real-world scenarios.
- **Scalability:** Current models tend to be computationally expensive, making them difficult to deploy on mobile devices or in low-resource settings.
- **Lack of Advanced Architectures:** While previous models have used traditional CNNs, more recent architectures like Vision Transformers (ViT) [1] and Swin Transformers [6], which have shown superior performance in image recognition tasks, have not been extensively explored for this problem.

Chapter 4

DESIGN AND METHODOLOGY

This research aims to develop a deep learning-based system capable of detecting abnormal hand movements, specifically hand flapping, as potential indicators of autism from unstructured home video data. The proposed approach combines privacy-preserving data preprocessing with state-of-the-art deep learning models (MobileNet [5], ResNet [2], Vision Transformer [1], and Swin Transformer [6]) to create a scalable, efficient, and non-invasive system for early autism detection. The system will leverage MediaPipe [8] for hand landmark extraction, followed by advanced deep learning models to classify the presence of hand flapping behaviors in video sequences.

4.1 Dataset Selection Process

The effectiveness of any machine learning model largely depends on the quality and relevance of the data used for training and evaluation. In this study, we selected the Self-Stimulatory Behavior Dataset (SSBD) [7], which contains annotated video clips of children exhibiting various self-stimulatory behaviors, including hand flapping, head banging, and spinning. This dataset was chosen because:

1. **Relevance to Autism Behaviors:** The SSBD [7] is specifically designed to capture behaviors that are common indicators of autism spectrum disorder (ASD), such as hand flapping, head banging and spinning which are frequently used in clinical assessments for autism.

2. **Diversity in Data:** The SSBD [7] includes videos of children of varying ages, genders, and developmental stages, providing a diverse set of examples. This diversity is essential for training models that can generalize well to a wide range of real-world scenarios.

3. **Annotated Labels:** The dataset provides detailed annotations of when self-stimulatory behaviors occurring within the videos, making it easier to label positive (hand flapping,head banging or spinning) and control (non-hand flapping) video clips. This allows for accurate training and validation of the model.

4. **Open-Source and Accessible:** The SSBD [7] is publicly available, making it an ideal choice for research.

Data Preprocessing:

- **Video Extraction:** From the SSBD [7], we extracted clips specifically containing hand flapping behaviors, as well as control clips with no self-stimulatory movements. The videos vary in length, with hand flapping events typically lasting between 2 to 5 seconds.

- **Data Augmentation:** To increase the robustness of the model, we performed data augmentation by flipping, rotating, and adjusting the lighting of the video clips. This helps prevent overfitting and improves the model's generalization ability.

- **Split into Training, Validation, and Test Sets:** The dataset will be divided into cross-validation sets to ensure that the model is trained and evaluated on different subsets of the data. Each fold will consist of a balanced set of hand flapping and control clips to prevent class imbalance from affecting the model's performance.

4.2 Privacy-Preserving Data Preprocessing with MediaPipe [8]

To ensure privacy and safeguard sensitive information, the study will avoid the use of full-body or facial data, which may raise privacy concerns. Instead, we will employ MediaPipe [8], an open-source framework developed by Google, for real-time hand tracking and landmark detection. MediaPipe [8] uses a pre-trained model to extract the (x, y, z) coordinates of 21 hand landmarks, which represent key points on the hands, such as the fingertips, palms, and wrist.

Process Overview:

1. **Hand Landmark Detection:** MediaPipe's [8] hand model detects and outputs the coordinates of 21 hand landmarks for each frame in the video. These landmarks include the positions of the five fingers, wrist, and palm center.

2. **Feature Vector Creation:** The raw video frames are converted into a sequence of feature vectors, where each vector contains the (x, y, z) coordinates of the detected landmarks. This reduces the dimensionality of the input while

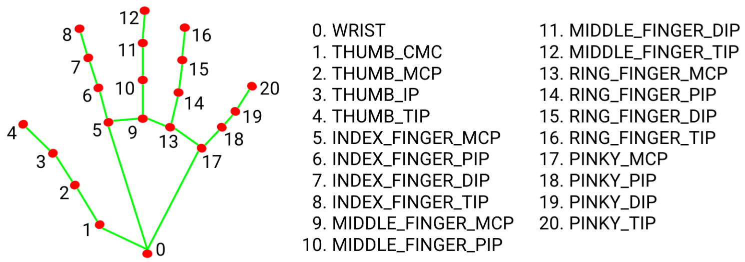


Figure 4.1: Hand Landmarks Extracted From MediaPipe

preserving the essential spatial information necessary for behavior classification.

3. **Privacy Preservation:** This approach ensures that only hand movement data is used for classification, thus eliminating the need for full-body or facial images and maintaining the privacy of the individuals in the video.

The landmark-based features extracted by MediaPipe [8] serve as the input data for the deep learning models, allowing the system to analyze hand movements without compromising privacy.

4.3 Deep Learning Model Selection

Four state-of-the-art deep learning architectures are used to process the hand landmark data and classify hand movements. These models—ResNet [2], Vision Transformer (ViT) [1], Swin Transformer [6], and MobileNet [5]—are chosen for their superior performance in feature extraction from visual data, particularly for image and video analysis tasks.

1. **MobileNet**

   - **MobileNet** is a lightweight convolutional neural network (CNN) designed specifically for mobile and embedded applications. It utilizes depthwise separable convolutions to reduce the computational cost and memory footprint of the network without significantly sacrificing performance.
   - **Application:** MobileNet [5] will be used to extract compact yet effective features from the sequence of hand landmarks. Its efficiency makes it an ideal choice for deployment on mobile devices, ensuring that the model can be used in real-time on low-resource platforms.

2. **ResNet (Residual Networks)**

   - **ResNet** is a deep convolutional neural network (CNN) designed to overcome the limitations of training very deep networks by introducing residual connections. These connections allow the model to learn residuals (differences) between layers, making it easier to train deeper networks.
   - **Application:** In this research, ResNet[2] will be used to extract high-level features from the sequence of hand landmarks, enabling it to learn complex patterns such as hand flapping and other self-stimulatory behaviors. Its ability to capture detailed hierarchical features makes it ideal for this task.

3. **Vision Transformer (ViT)**

   - **Vision Transformer** represents a shift from traditional CNNs by using transformer models, which have proven to be highly effective in capturing long-range dependencies in data. In ViT[1], an image is divided into fixed-size patches, and each patch is treated as a sequence of tokens for processing by the transformer.
   - **Application:** ViT[1] will be used to model the spatial relationships between hand landmarks across video frames, using its self-attention mechanism to focus on important regions of the video that might indicate hand flapping or other movements related to autism.

4. **Swin Transformer**

   - **Swin Transformer** is a hierarchical vision transformer that uses a shifted window mechanism to capture both local and global features efficiently. Unlike ViT[1], which processes the entire image at once, Swin Transformer [6] divides the image into non-overlapping windows, making it computationally more efficient while still capturing fine-grained spatial details.
   - **Application:** Swin Transformer's [6] ability to scale efficiently to larger inputs and capture both detailed local and broader contextual features makes it ideal for video-based hand movement detection.

Each of these models will be tested on sequences of hand landmark data extracted by MediaPipe [8], with the goal of determining which model best captures the nuances of hand flapping behaviors and provides accurate classification.

4.4 Model Architecture

The architecture of the proposed system involves multiple stages:

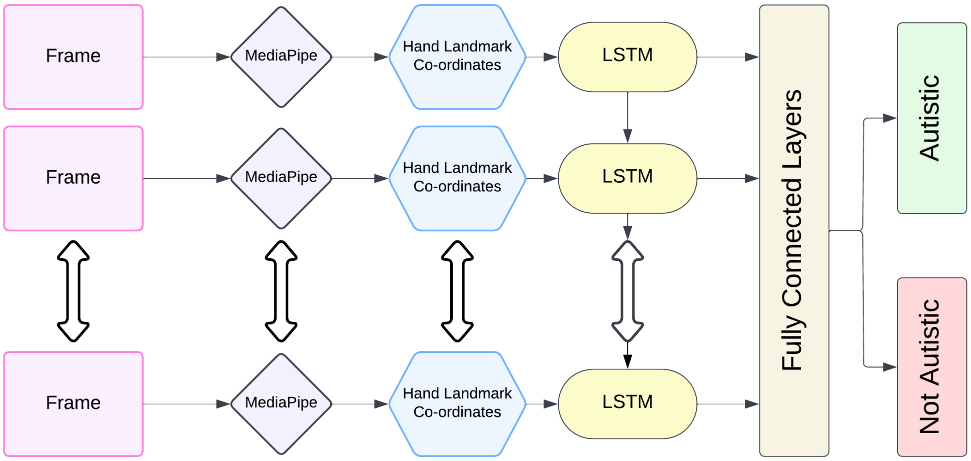


Figure 4.2: MediaPipe Model Architecture

1. **Hand Landmark Extraction with MediaPipe:** The first step in the pipeline is using MediaPipe [8] to extract the hand landmarks from each video frame. The (x, y, z) coordinates of these landmarks form the primary feature vectors that describe the hand movements.

2. **Feature Extraction by Deep Learning Models:** The hand landmark data is processed by the four deep learning models—ResNet[2], ViT [1], Swin Transformer [6], and MobileNet [6]. These models extract meaningful features from the sequential landmark data, learning to recognize patterns indicative of hand flapping behaviors.

3. **Temporal Modeling with LSTM:** Given that hand flapping and other autism-related behaviors are time-dependent, an LSTM (Long Short-Term Memory)[3] network is employed to capture the temporal dependencies between successive frames in the video. The LSTM [3] takes the features extracted by [2], ViT [1], Swin Transformer [6], and MobileNet [5], and models how hand movements evolve over time.

4. **Final Classification Layer:** The output from the LSTM [3] is passed through a fully connected layer with a sigmoid activation function to produce a binary classification—indicating whether hand flapping is present in the video sequence.

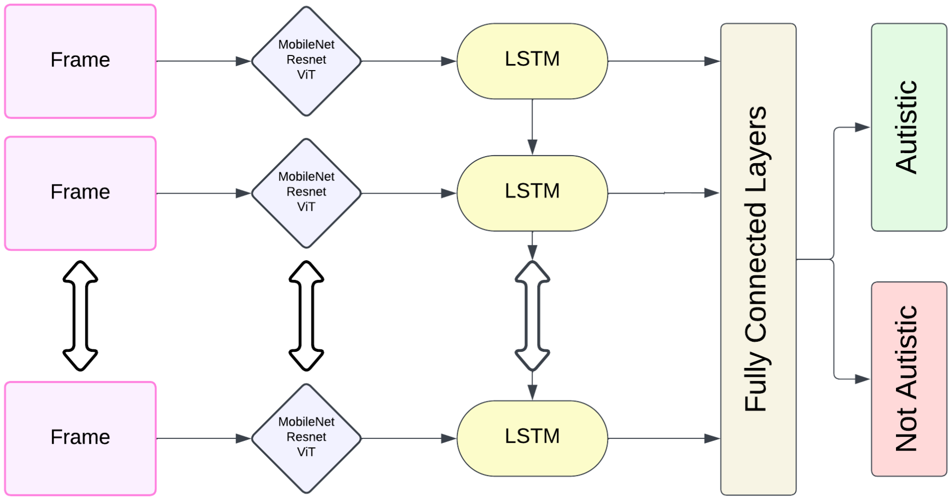


Figure 4.3: CNN Model Architecture

4.5 Model Training and Evaluation

To train the models, the Self-Stimulatory Behavior Dataset (SSBD) [7], which includes annotated video clips of hand flapping, head banging, and spinning behaviors, will be used. For this study, we will focus on hand flapping as the target behavior and extract relevant video clips from the dataset.

- **Data Splitting:** The SSBD [7] will be split into training, validation, and testing sets using 5-fold cross-validation, ensuring robust model evaluation.

- **Training Process:** The models will be trained using binary cross-entropy loss and the Adam optimizer [4]. The training will be performed on a standard CPU, and model performance will be evaluated based on accuracy, precision, recall, and F1 score.

- **Evaluation Metrics:** The models will be evaluated using metrics such as F1 score, precision, recall, and accuracy, ensuring that both false positives and false negatives are minimized.

Chapter 5

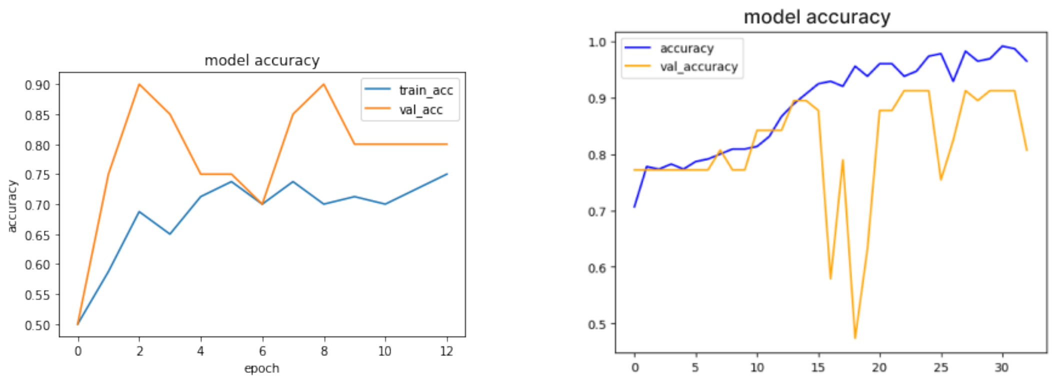
RESULTS

The models will be compared based on the following criteria:

- **Accuracy:** The proportion of correctly classified frames.
- **Precision:** The proportion of true positive hand flapping instances among all predicted positive frames.
- **Recall:** The proportion of true positive hand flapping instances among all actual positive frames.
- **F1 Score:** The harmonic mean of precision and recall, providing a balanced evaluation of model performance.

Additionally, a comparative study will be conducted to assess the performance of the deep learning models (ResNet [2], ViT[1], and Swin Transformer [6]) against traditional hand landmark-based methods. This comparison will demonstrate the benefits of using advanced models for detecting autism-related behaviors and the privacy-preserving advantage of using MediaPipe [8] for data extraction.

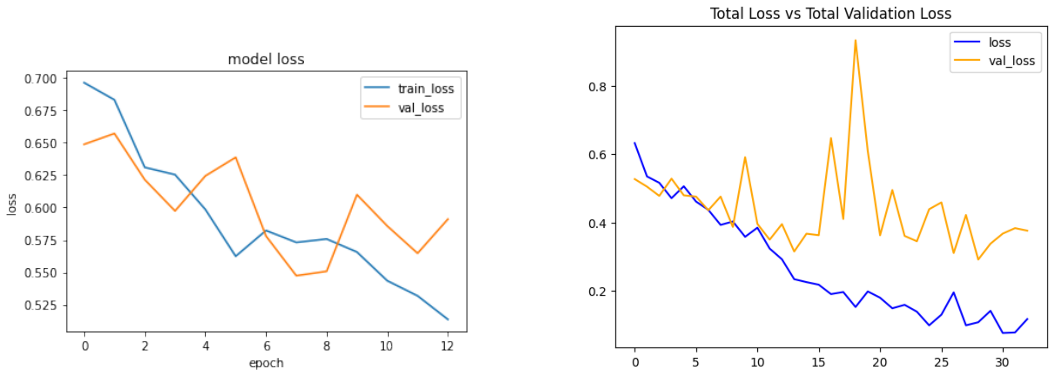
The dataset can be downloaded from **Self-Stimulatory Behaviour Dataset**.
The Code are available at **Github Repository**.



(a) MediaPipe Accuracy

(b) MobileNet Accuracy

Figure 5.1: Comparison of MediaPipe and MobileNet Accuracy



(a) MediaPipe Loss

(b) MobileNet Loss

Figure 5.2: Comparison of MediaPipe and MobileNet Loss

By integrating MediaPipe [8] for privacy-preserving landmark extraction with state-of-the-art deep learning models (ResNet [2], ViT [1], and Swin Transformer [6]), this research proposes a robust, scalable, and efficient solution for detecting autism-related behaviors from home video data. The approach ensures that sensitive personal information is protected while providing an accurate and accessible tool for early autism diagnosis.

Metric	**MediaPipe**	**MobileNet**
Accuracy	0.69	0.91
Precision	0.71	0.95
Recall	0.68	0.91
F1 Score	0.68	0.92

Table 5.1: Comparison of Metrics for MediaPipe and MobileNet

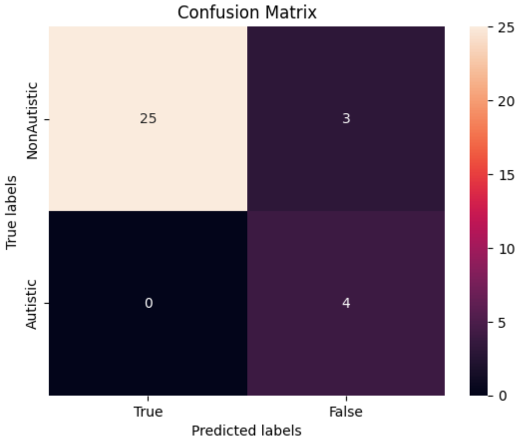


Figure 5.3: MobileNet Confusion Matrix

Chapter 6

FUTURE SCOPE

While this research focuses on detecting hand flapping, which is one of the common self-stimulatory behaviors associated with autism, there is a significant potential to expand this work to detect other autism-related behaviors, such as head banging. Head banging is another stereotypical behavior frequently exhibited by children with autism and is often used in clinical assessments to aid in diagnosis. Incorporating head banging detection will make the system more robust and versatile, enabling it to analyze a broader range of behaviors associated with autism.

6.1 Head Banging Detection with ViT, Swin Transformer, and MediaPipe

1. **ViT and Swin Transformer:** Both Vision Transformer (ViT) [1] and Swin Transformer [6] have demonstrated superior performance in extracting complex spatial and temporal features from video sequences. By extending the models used in this research to analyze head banging, we can take advantage of their ability to capture long-range dependencies and local details in visual data. These transformers will be able to recognize distinctive patterns associated with head banging, such as the motion trajectory and speed of head movement, by processing the video data in a similar fashion as they did for hand movements.

    - **Vision Transformer (ViT)** will be used to process sequences of video frames containing head movements, capturing both the local features of the head's position and global dependencies across frames.
    - **Swin Transformer** will be employed to handle the hierarchical nature of head banging, efficiently capturing both fine-grained motion features (such as rapid movements) and broader contextual relationships in the video sequence.

2. **Model Extension and Temporal Modeling:** Once the head landmarks are extracted, the sequences of head movement data will be fed into the existing LSTM (Long Short-Term Memory) network [3], allowing the model to capture the temporal aspects of head banging behaviors. The combination of ViT [1], Swin Transformer [6] will enable the system to robustly detect head banging,

distinguishing it from other motions and behaviors that might appear in the video.

6.2 Potential Challenges and Solutions

- **Motion Variability:** Head banging can occur in various forms (e.g., rapid or slow motions, different head angles), which can make detection challenging. Using a combination of ViT [1] and Swin Transformer [6] will help capture a wide variety of head movements, as these models are adept at understanding both fine-grained and broad spatial relationships.

- **Data Labeling and Annotation:** To accurately train the models for head banging detection, additional labeled data is required, specifically with labeled video clips of children exhibiting head banging behaviors. Future work will involve curating or acquiring a dataset that includes these labeled examples.

- **Real-Time Detection:** Head banging detection, especially in real-time applications, may face computational challenges. However, the use of MobileNet [5] as a lightweight alternative to ResNet [2] or ViT [1] for feature extraction can help ensure the model remains computationally efficient, enabling it to run in real-time on mobile or low-resource devices.

6.3 Conclusion

The future work on head banging detection will significantly enhance the capabilities of the current system, providing a more comprehensive tool for the detection of autism-related behaviors. By extending the work to incorporate ViT [1], Swin Transformer [6] for head banging detection, this research can contribute to the development of a scalable, efficient, and privacy-preserving tool that aids in early autism diagnosis. The ability to detect multiple behaviors, such as hand flapping and head banging, will make the system more versatile and beneficial for both clinicians and caregivers in diagnosing and monitoring children with autism spectrum disorder.

BIBLIOGRAPHY

[1] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *arXiv* (2021). arXiv: `2010.11929 [cs.CV]`. URL: `https://arxiv.org/abs/2010.11929`.

[2] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *arXiv* (2015). arXiv: `1512.03385 [cs.CV]`. URL: `https://arxiv.org/abs/1512.03385`.

[3] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-term Memory”. In: *Neural computation* 9 (Dec. 1997), pp. 1735–1780. DOI: `10.1162/neco.1997.9.8.1735`.

[4] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-term Memory”. In: *Neural computation* 9 (Dec. 1997), pp. 1735–1780. DOI: `10.1162/neco.1997.9.8.1735`.

[5] Andrew G. Howard et al. “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications”. In: *arXiv* (2017). arXiv: `1704.04861 [cs.CV]`. URL: `https://arxiv.org/abs/1704.04861`.

[6] Ze Liu et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows”. In: *arXiv* (2021). arXiv: `2103.14030 [cs.CV]`. URL: `https://arxiv.org/abs/2103.14030`.

[7] Vaibhavi Lokegaonkar et al. “Introducing SSBD+ Dataset with a Convolutional Pipeline for detecting Self-Stimulatory Behaviours in Children using raw videos”. In: *arXiv* (2023). arXiv: `2311.15072 [cs.CV]`. URL: `https://arxiv.org/abs/2311.15072`.

[8] Camillo Lugaresi et al. “MediaPipe: A Framework for Building Perception Pipelines”. In: *arXiv* (2019). arXiv: `1906.08172 [cs.DC]`. URL: `https://arxiv.org/abs/1906.08172`.