International Journal of Advanced Culture Technology Vol.12 No.3 148-158 (2024)
DOI https://doi.org/10.17703/IJACT.2024.12.3.148

IJACT 24-9-18

Utilizing Deep Learning for Early Diagnosis of Autism : Detecting Self-Stimulatory Behavior

[1]Seongwoo Park, [2]Sukbeom Chang, JooHee Oh**

[1]Department of Global Entrepreneurship and Information Communication Technology, Handong Global University, Pohang, Korea
[2]Department of Global Entrepreneurship and Information Communication Technology, Handong Global University, Pohang, Korea
***Department of Management and Economics, Handong Global University, Pohang, Korea*

E-mail: jooheeoh@handong.edu

Abstract

We investigate Autism Spectrum Disorder (ASD), which is typified by deficits in social interaction, repetitive behaviors, limited vocabulary, and cognitive delays. Traditional diagnostic methodologies, reliant on expert evaluations, frequently result in deferred detection and intervention, particularly in South Korea, where there is a dearth of qualified professionals and limited public awareness. In this study, we employ advanced deep learning algorithms to enhance early ASD screening through automated video analysis. Utilizing architectures such as Convolutional Long Short-Term Memory (ConvLSTM), Long-term Recurrent Convolutional Network (LRCN), and Convolutional Neural Networks with Gated Recurrent Units (CNN+GRU), we analyze video data from platforms like YouTube and TikTok to identify stereotypic behaviors (arm flapping, head banging, spinning). Our results indicate that the LRCN model exhibited superior performance with 79.61% accuracy on the augmented platform video dataset and 79.37% on the original SSBD dataset. The ConvLSTM and CNN+GRU models also achieved higher accuracy than the original SSBD dataset. Through this research, we underscore AI's potential in early ASD detection by automating the identification of stereotypic behaviors, thereby enabling timely intervention. We also emphasize the significance of utilizing expanded datasets from social media platform videos in augmenting model accuracy and robustness, thus paving the way for more accessible diagnostic methods.

Keywords: *Autism Spectrum Disorder(ASD), Computer Vision, Stereotyped behavior, Early diagnosis, CNN, GRU*

Manuscript received: July 25, 2024 / revised: August 14, 2024 / accepted: September 5, 2024
Corresponding Author: Joo Hee Oh (E-mail: jooheeoh@handong.edu)
Tel:+82-54-260-1420
Professor, Dept. of Management and Economics, Handong Global Univ., Korea

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is characterized as a neurodevelopmental disorder encompassing a complex array of symptoms, including deficits in social interaction, repetitive and stereotyped behaviors, limited vocabulary, and cognitive developmental delays [1, 2]. Despite extensive research, the genetic etiology of ASD remains inconclusive [1, 3]. Diagnosis is typically conducted by specialists utilizing criteria from the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), or the International Classification of Diseases, Eleventh Revision (ICD-11) [4]. In South Korea, various diagnostic tools are employed, including the Autism Diagnostic Interview-Revised (ADI-R), Autism Diagnostic Observation Schedule 2 (ADOS-2), the Childhood Autism Rating Scale, Second Edition (CARS-2), and the Korean Developmental Screening Test for Infants & Children (K-DST), developed by the Ministry of Health and Welfare and the Korean Pediatric Society in 2014 [4, 5].

ASD typically manifests in early childhood, with early screening possible from a very young age. Reports indicate that symptoms may be detectable in children younger than one year old [5]. Consequently, early screening is emphasized, as early intervention facilitated by such screening positively influences the developmental trajectory of ASD symptoms [5]. However, in South Korea, ASD recognition tends to occur later compared to other disorders, primarily due to a shortage of qualified medical professionals [6]. More critically, delays often result from parents or guardians failing to recognize early symptoms of ASD in their children [5]. Additionally, a lack of awareness and proactive measures among individuals in a child's social environment, such as teachers and social workers, exacerbates this issue [7, 8].

Primary diagnostic indicators of ASD encompass various aspects such as communication abilities, social interactions, and repetitive behaviors [1]. However, communication and social interaction skills can sometimes be ambiguous when distinguishing ASD [1]. Therefore, repetitive and impulsive stereotypic behaviors are utilized as primary indicators in ASD identification. These stereotypic behaviors include body rocking, head banging, arm flapping, head shaking, jumping, and spinning [9]. These visually identifiable behaviors serve as reliable indicators for early diagnosis, enabling parents to recognize these behaviors in their children [7].

With the increasing significance of mechanically understanding human behavior, the application of artificial intelligence (AI) in behavior analysis has emerged as a prominent research area. Specifically, AI holds considerable potential in the diagnosis of ASD. Behavior recognition and analysis algorithms utilizing computer vision technologies have demonstrated the ability to identify and analyze behaviors in autistic children from video footage, classifying stereotypic behaviors such as arm flapping, head banging, and spinning [10]. The advancements in AI technologies have mitigated the previously existing ambiguities in ASD identification, facilitating more precise and rapid diagnoses. Moreover, the widespread availability of unregulated video formats on social media platforms such as YouTube, Vimeo, and Dailymotion has led to an exponential increase in video data. This proliferation of video content, combined with AI technologies, not only enhances early detection capabilities for ASD in children but also promotes continuous performance improvements and the democratization of these diagnostic methods.

This study endeavors to enhance the efficiency of early autism screening by employing computer vision technology and machine learning algorithms to analyze video data of autistic children. Specifically, it focuses on the identification of stereotypic behaviors such as arm flapping, head banging, and spinning using uncontrolled video footage sourced from social network platforms like YouTube and TikTok, as well as the Self-Stimulatory Behavior Dataset (SSBD). By addressing the challenges associated with video quality and noise through the application of deep learning algorithms, this research offers substantial implications for the automated screening of autism symptoms prior to recognition by parents and caregivers. This innovative approach empowers individuals without specialized knowledge to detect symptoms at an early stage, thereby

facilitating timely intervention and the implementation of appropriate therapeutic strategies for autistic children. Consequently, this study has the potential to significantly accelerate the process of early diagnosis and intervention.

2. LITERATURE REVIEW

Conventional methodologies for the early diagnosis of Autism Spectrum Disorder (ASD) employ various diagnostic tools and behavioral assessments, which inherently possess limitations. Primarily, traditional methods rely on direct observations and structured interviews conducted by trained professionals, which, while effective, demand considerable resources, extensive expertise, and significant time investment. Early screening instruments such as the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS) are frequently utilized for diagnosing ASD; however, their reliance on expert consultation does not expedite the initial identification process by parents [4]. Additionally, the Modified Checklist for Autism in Toddlers (M-CHAT) enables parents to observe their child's everyday behaviors and respond to specific questions to identify early signs of autism [11]. Traditional methods face limitations in immediate data access and continuous data measurement, necessitating research to enhance the speed and convenience of autism diagnosis through web and mobile platforms [7].

In a study by [6], the reliability of items in existing early screening tools was evaluated, leading to the development of a web-based system centered on verified items, allowing parents, teachers, and clinicians to identify disabilities early. Efforts persist in facilitating access to autism diagnosis via smart platform-based tools, enabling laypersons to easily evaluate autism, with results aligning significantly with medical diagnostic outcomes [7]. Research conducted by [12] utilizing smart tablet computers with touch-sensitive screens and inertial motion sensors highlighted the potential of identifying autism through smart device applications. The development of traditional screening tools and the application of smart platforms are particularly effective in environments with a shortage of medical professionals. However, the lack of awareness about screening tools and information on apps and web platforms among parents and other individuals interacting with the child poses a challenge in significantly improving the primary diagnostic speed.

Recent studies employing computer vision technology play a pivotal role in detecting the characteristic stereotyped behaviors of ASD. Approaches leveraging artificial intelligence (AI) have demonstrated substantial improvements in the speed and accuracy of traditional ASD early screening, thereby significantly contributing to the detection of early signs [2]. For instance, [2] combined VGG-16 and Long-term Recurrent Convolutional Network (LRCN) to identify and classify hand flapping in the Self-Stimulatory Behavior Dataset (SSBD), which comprises videos of autistic children recorded in uncontrolled environments. In another study, [13] employed an adjusted MobileNetV2 model and extracted features from hand landmarks detected by MediaPipe, predicting hand flapping in SSBD videos using a Long Short-Term Memory Network (LSTM). Furthermore, [14] trained a three-class classifier within a Bag Of Words (BOW) framework to classify arm flapping, head banging, and spinning behaviors using Space-Time Interest Points (STIP) and Harris3D detectors. Additionally, [1] tested various local descriptors used in the Bag-of-Visual-Words approach with Multi-layer Perceptron (MLP), Gaussian Naive Bayes (GNB), and Support Vector Machines (SVM) classifiers to identify stereotypic behaviors (arm flapping, head banging, spinning) associated with ASD. Lastly, [10] utilized EfficientNets, MobileNetV3, ShuffleNet, ESNet, ResNet, and Inflated 3D ConvNet (I3D) to extract features and recognized behaviors using Long Short-Term Memory (LSTM), Temporal Convolutional Network (TCN), Multi-Stage Temporal Convolutional Network (MS-TCN), and MS-TCN++ to identify and classify stereotypic behaviors.

These prior studies aim to identify early signs of autism by analyzing video data of autistic children, thereby

improving early screening speed and enabling timely intervention for autistic children. This study, sharing the same objective, seeks to identify stereotypic behaviors such as arm flapping, head banging, and spinning in autistic children using uncontrolled video footage. The study aims to address challenges posed by video quality and noise through the application of deep learning algorithms, holding significant implications for the potential automated screening of autism symptoms through everyday recorded videos before parents and caregivers can identify the symptoms. This approach allows individuals without specialized knowledge to recognize symptoms early, thereby promoting early intervention and appropriate therapeutic strategies for autistic children.

3. METHOD

This research employs several advanced deep learning architectures to enhance the recognition of behavioral patterns through sophisticated feature extraction and temporal continuity analysis. Firstly, the study augments traditional Long Short-Term Memory (LSTM) networks by incorporating a Convolutional Long Short-Term Memory (ConvLSTM) network. This adaptation introduces convolutional structures to both the input-to-state and state-to-state transitions, thereby rendering the architecture particularly adept at capturing spatiotemporal correlations. This capability makes ConvLSTM highly suitable for tasks that involve both spatial and temporal dependencies, such as video-based action recognition. ConvLSTM's ability to preserve spatial information through convolutional operations facilitates more precise modeling of the dynamic nature of video data, thereby enhancing prediction performance in scenarios characterized by significant temporal variation[15]. Secondly, the methodology integrates a Long-term Recurrent Convolutional Network (LRCN) approach. LRCNs synergize the strengths of Convolutional Neural Networks (CNNs) for feature extraction with those of LSTMs for sequence modeling. This approach effectively bridges the gap between spatial and temporal analysis by initially leveraging CNNs to extract spatial features from each frame, which are subsequently fed into LSTM units to capture temporal dynamics across frames. This amalgamation allows for a more nuanced understanding and prediction of behaviors over extended sequences, thereby augmenting the model's capability to manage complex temporal dependencies [16]. Finally, to further refine the recognition of behavioral patterns, an integrated deep learning architecture combining Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs) is employed. The primary objective of this architecture is to accurately classify and predict diverse behaviors captured in video footage, with particular emphasis on analyzing dynamic and complex behavior patterns within a temporal context. CNNs serve as robust visual feature extractors, identifying critical information in each frame. Concurrently, GRUs model the temporal continuity between these features, facilitating a comprehensive understanding and prediction of behaviors across entire video sequences. This integrated approach effectively addresses the inherent complexities of video data, including temporal variations, thereby transcending the limitations of simple image recognition.

This study leverages these advanced architectures—ConvLSTM, LRCN, and CNN-GRU frameworks—to identify stereotypical behaviors in children with autism, with the aim of determining the most suitable architecture for analyzing videos with low resolution and high noise levels.

3.1 ConvLSTM Approach

The ConvLSTM model architecture employed in this study is meticulously designed to capture and analyze spatiotemporal dependencies inherent in video data. Utilizing a Sequential model framework, the architecture commences with a ConvLSTM2D layer comprising 4 filters and a 3x3 kernel, utilizing 'tanh' activation and a recurrent dropout rate of 0.2. This initial layer is adept at processing input sequences with defined dimensions, accommodating the temporal sequence length and spatial dimensions of 128x128 pixels across three color channels. As shown in Figure 1, the model's structure visualizes the systematic arrangement of layers to enhance feature extraction and temporal continuity. The model incorporates additional ConvLSTM2D layers with increasing complexity, comprising 8, 14, and 16 filters, respectively. Each ConvLSTM2D layer is

followed by MaxPooling3D layers to effectively reduce spatial dimensions and TimeDistributed Dropout layers for rigorous regularization. The hierarchical stacking of these layers allows for the extraction of intricate spatiotemporal patterns, crucial for accurate behavior recognition. The final stage of the architecture involves flattening the output from the ConvLSTM layers, followed by a Dense layer with a softmax activation function, which yields probability distributions across the target classes. This comprehensive approach, integrating convolutional and recurrent layers with strategic regularization techniques, ensures robust and precise modeling of dynamic video data, particularly suited for behavior recognition tasks within the spatial constraints of 128x128 pixels.

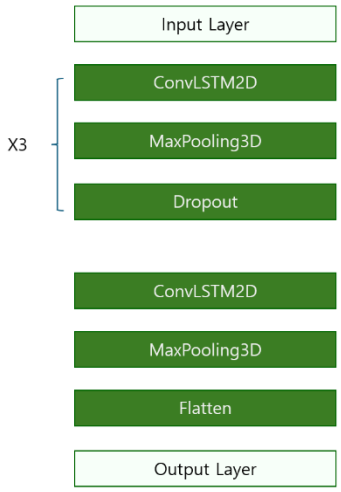

Figure 1. Model Structure(ConvLSTM)

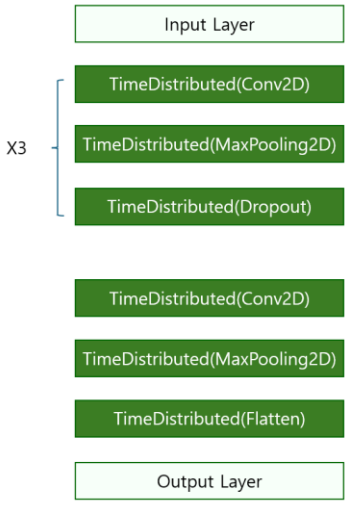

Figure 2. Model Structure(LRCN)

3.2 LRCN Approach

The LRCN (Long-term Recurrent Convolutional Network) model architecture developed in this study effectively combines convolutional neural networks (CNNs) for spatial feature extraction and long short-term memory (LSTM) networks for temporal sequence modeling. This architecture is implemented using a Sequential framework, ensuring a systematic flow of data. As depicted in Figure 2, the model's structure illustrates the layered configuration starting with a TimeDistributed Conv2D layer featuring 16 filters and a 3x3 kernel, employing 'relu' activation and 'same' padding, designed to handle input sequences with spatial dimensions of 128x128 pixels across three color channels. This layer is followed by a TimeDistributed MaxPooling2D layer (4x4) and a TimeDistributed Dropout layer (0.25) for spatial dimension reduction and regularization, respectively. Subsequent layers include additional TimeDistributed Conv2D layers with 32 and 64 filters, each followed by corresponding MaxPooling2D and Dropout layers. This hierarchical convolutional structure enables the extraction of increasingly complex features while mitigating overfitting. Post convolutional operations, a TimeDistributed Flatten layer converts 2D feature maps into 1D vectors, preparing the data for the LSTM layer. The LSTM layer, with 32 units, captures temporal dependencies, facilitating the understanding of dynamic sequences. The final Dense layer, employing a softmax activation function, produces class probability distributions. This integrated approach, combining CNNs for spatial analysis and LSTMs for temporal dynamics, is optimal for video-based behavior recognition within 128x128 pixel spatial constraints. The model summary elucidates the architecture's detailed configuration, ensuring reproducibility and clarity. By leveraging this advanced LRCN framework, the study aims to achieve robust behavior pattern recognition in video data, demonstrating the model's efficacy in handling both spatial and temporal complexities.

3.3 CNN+GRU Approach

The study utilizes the pre-trained ResNet50 architecture as a feature extractor to transform raw video frames into robust feature representations. The CNN processes image data and extracts high-level features that encapsulate essential visual information necessary for recognizing complex behaviors within video data. As illustrated in Figure 3, this processing is applied to each frame of the segmented video clips, with the CNN's hyperparameters set to an image size of 224x224 and a batch size of 64, facilitating effective model learning and convergence. The extracted feature representations are then adapted to the specific task and prepared for subsequent processing by a recurrent neural network.

To model the sequential temporal relationships between video frames and enhance prediction accuracy, the study employs Gated Recurrent Units (GRUs). The model utilizes a multi-layer GRU network, sequentially configured with layers containing 64, 32, 32, and 16 units, respectively. A dropout rate of 0.5 is applied to the intermediate layers to mitigate overfitting. This configuration effectively captures the temporal characteristics of each video clip, learning both spatial and temporal features from the high-dimensional features extracted by the CNN. The output from the final GRU layer is passed through a softmax layer to classify the video clips into predefined categories. The model is trained on a multi-class classification problem using the Adam optimizer and a cross-entropy loss function. During training, 20% of the dataset is reserved for validation to evaluate the model's generalization capability, and the best-performing model is saved. This process is crucial for preventing overfitting to the training data and enhancing the model's prediction accuracy on real-world data.

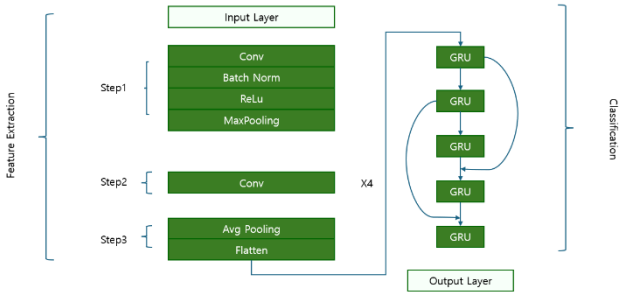

Figure 3. Model Structure(CNN+GRU)

4. DATESET/EXPERIMENT

4.1 Data

In this study, the data was constructed based on the Self-Stimulatory Behavior Dataset (SSBD) collected by [14]. The SSBD comprises video data collected under uncontrolled conditions from platforms such as YouTube, Vimeo, and Dailymotion, accompanied by corresponding annotations. The dataset is categorized into three types based on autistic behavior characteristics: Arm Flapping, Head Banging, and Spinning. Originally, the SSBD included 75 videos; however, due to accessibility issues over time and the presence of noise that made 16 videos difficult to discern, only 59 videos were used. The SSBD fundamentally utilizes annotations as its primary data format. Figure 4 illustrates an example of the SSBD annotation format, which includes information on <url>, <height>, <width>, <frames>, <persons>, <duration>, <conversation>, <behaviour>, <category>, <time>, <bodypart>, <intensity>, and <modality>. Additionally, in this study, 32 more videos were collected through social networking services such as YouTube and TikTok, expanding the dataset to a total of 91 videos. The newly collected videos comprise 8 for Arm Flapping, 9 for Head Banging, and 15 for Spinning, and these also adhere to the xml-based annotation format used in the original SSBD. The overall dataset utilizes only the tags <url>, <height>, <width>, <frames>, <duration>, <behaviour>, <category>, and <time> from the SSBD annotations, which were deemed necessary for this research. Similarly,

the annotations for the additional videos correspond to these tags. The resolution of the dataset videos ranges from a minimum of 184 x 144 pixels to a maximum of 1280 x 1280 pixels. The additional data were collected from YouTube and TikTok as follows.

The additional videos were collected from YouTube and TikTok using the following steps:

1. Data queries included terms such as 'Autism, stimming + headbanging, stimming + armflapping, stimming + spinning, autism + headbanging, autism + armflapping, and autism + spinning.'
2. The filter 'Sort by' was set to 'Upload date' to collect data from the past six years.
3. Collected videos were selected based on the following criteria:
    - A clearly identifiable single individual exhibiting Autism symptoms.
    - Identifiable behavior.

The final selected data were converted into XML format. Figure 5 presents example images for each category: 'Arm Flapping,' 'Head Banging,' and 'Spinning.' The composition of the final dataset is presented in Table 1.

Table 1. Summary of Video Counts by Stereotypic Behavior

	Armflapping	Headbanging	Spinning	Total
SSBD	25	25	25	75
Deleted Videos	4	7	5	16
Added Videos	8	9	15	32
Total	29	27	35	91

```
<video id="v_ArmFlapping_13" keyword="Flapping his arms ">
   <url>http://www.youtube.com/watch?v=3Oh_Lmehb6c</url>
   <height>360</height>
   <width>640</width>
   <frames>2586</frames>
   <persons>1</persons>
   <duration>86s</duration>
   <conversation>yes</conversation>
   <behaviours count="2" id="b_Set_01">
      <behaviour id="b_01">
         <time>0019:0022</time>
         <bodypart>hand</bodypart>
         <category>armflapping</category>
         <intensity>high</intensity>
         <modality>video</modality>
      </behaviour>
   <behaviour id="b_02">
         <time>0040:0051</time>
         <bodypart>hand</bodypart>
         <category>armflapping</category>
         <intensity>low</intensity>
         <modality>video</modality>
      </behaviour>
   </behaviours>
</video>
```

Figure 4. SSBD Annotation

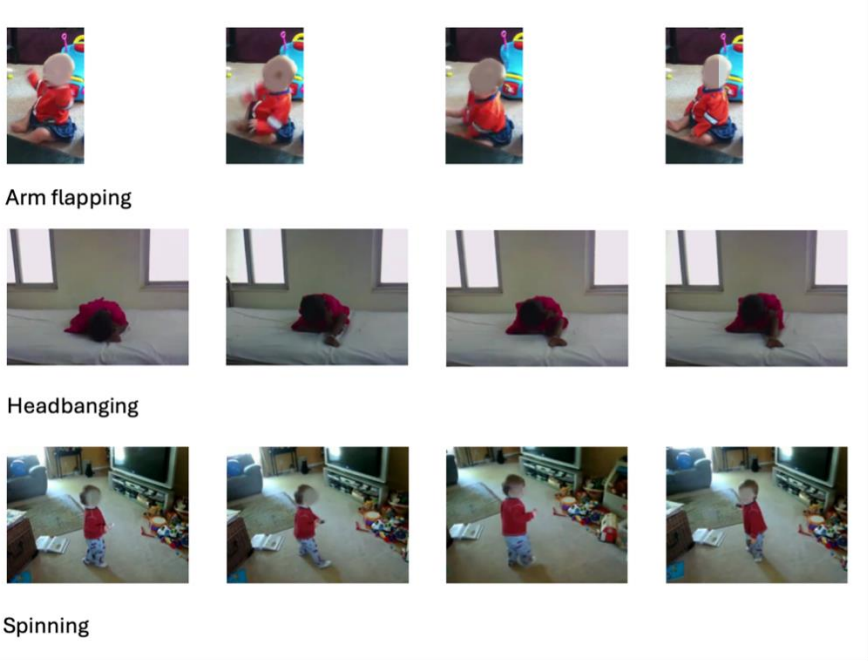

Figure 5. Example images for each category: 'Arm Flapping', 'Head Banging', and 'Spinning'

4.2 Preprocessing

The transformation of raw video frames into robust feature representations is crucial for effective processing in behavioral analysis. Initially, segments corresponding to stereotypic behaviors of autistic children, specifically Arm Flapping, Head Banging, and Spinning, were isolated from each video. Subsequently, the videos were annotated and categorized into four distinct classes: Arm Flapping, Head Banging, Spinning, and Normal. The 'Normal' category was derived from the remaining portions of the video after extracting clips exhibiting stereotyped behaviors. This process involved splitting the downloaded videos into clips based on XML annotations, categorizing the behavior, and further subdividing the clips into smaller segments if their duration exceeded five seconds. The video data were then converted into a format amenable to frame-level analysis. To enhance the preprocessing pipeline, several meticulous steps were undertaken to ensure data consistency and quality. Firstly, frames were extracted from each video at regular intervals, resized to a fixed resolution of 224x224 pixels, and normalized to ensure pixel values ranged between 0 and 1. This standardization facilitated uniform representation of each video by a consistent sequence of frames suitable for subsequent analysis. Moreover, a balanced dataset was created by selectively sampling the 'Normal' class, ensuring it constituted 70% of the total data. This approach mitigated class imbalance and enhanced the robustness of the training process. The extracted frames and their corresponding labels were subsequently converted into numpy arrays for efficient storage and processing. In the final preprocessing step, the data were stratified into training and test datasets, with 80% allocated for training and the remaining 20% reserved for testing. The distribution of the final dataset is presented in Table 2. These preprocessing steps were essential in transforming raw video data into a structured format, thereby enabling effective frame-level analysis for the identification and classification of stereotypic behaviors in autistic children.

Table 2. Summary of Clip Counts by Stereotypic Behavior (SSBD+New)

	Armflapping	Headbanging	Spinning	Normal	Total
Train	466	319	536	709	2030
Test	116	79	133	177	505
Total	582	398	669	886	2535

4.3 Experiment

In this study, experiments were conducted using three models—ConvLSTM, LRCN, and CNN+GRU—on both the original SSBD dataset and an augmented dataset that combined the SSBD data with additional video segments. These datasets comprised video frames extracted from recordings of autistic children's stereotypic behaviors, categorized into Arm Flapping, Head Banging, Spinning, and Normal. The 'Normal' category consisted of video segments not exhibiting stereotypic behaviors, derived from the residual portions of the videos after behavioral clips were extracted.

For these experiments, the ConvLSTM model, which integrates convolutional layers with Long Short-Term Memory (LSTM) units, was employed to capture both spatial and temporal features from the video frames. The model was compiled using the categorical cross-entropy loss function and optimized with the Adam optimizer. The training process was regulated by an early stopping callback, which monitored the validation loss with a patience of 10 epochs and restored the best model weights upon termination. The ConvLSTM model was trained for a maximum of 200 epochs with a batch size of 4, employing 80% of the data for training and 20% for validation. This model achieved an accuracy of 66.85% on the overall dataset and 66.47% on the original SSBD dataset.

The Long-term Recurrent Convolutional Network (LRCN) model, which combines convolutional neural networks (CNNs) for spatial feature extraction with LSTM units for sequential data processing, was also utilized. This model was compiled with the categorical cross-entropy loss function and Adam optimizer. An early stopping callback with a patience of 15 epochs was implemented to avoid overfitting. The LRCN model was trained for up to 200 epochs with a batch size of 4, using an 80-20 split for training and validation. The LRCN model demonstrated superior performance, achieving an accuracy of 79.61% on the overall dataset and 79.37% on the original SSBD dataset.

The CNN+GRU model was developed to leverage convolutional layers for feature extraction and Gated Recurrent Units (GRUs) for handling temporal dependencies. This model was trained over 100 epochs. The training data were partitioned with 80% allocated for training and 20% for testing. To ensure balance among the labels, a sampling process was applied, and an under-sampling technique was employed for the 'Normal' label due to its larger volume compared to other labels. Additionally, 20% of the training data were randomly extracted to form a validation dataset, which was used to periodically evaluate the model's performance during training. This validation dataset was crucial in preventing overfitting and maintaining the model's generalization capability. The optimization of the model was conducted based on its performance on the validation data, with the model demonstrating the best performance selected as the final model. The CNN+GRU model achieved an accuracy of 62% on the overall dataset and 53% on the original SSBD dataset.

The experimental results underscore that the LRCN model attained the highest accuracy, followed by the ConvLSTM model, with the CNN+GRU model exhibiting the lowest accuracy across both datasets. These findings highlight the importance of selecting appropriate architectures for the analysis of complex video data, particularly in the context of identifying and classifying stereotypic behaviors in autistic children.

5. DISCUSSON

This study aimed to augment the early detection and intervention response for autism by empowering parents and caregivers to identify early symptoms in children. To achieve this objective, the research expanded the existing Self-Stimulatory Behavior Dataset (SSBD) by incorporating additional datasets, resulting in a comprehensive, extended dataset. This enriched dataset was utilized to train and evaluate three distinct deep learning architectures: ConvLSTM, LRCN, and CNN+GRU.

The ConvLSTM model, which integrates convolutional layers with Long Short-Term Memory (LSTM) units, achieved an accuracy of 66.85% on the overall dataset and 66.47% on the original SSBD dataset. The Long-term Recurrent Convolutional Network (LRCN) model, combining Convolutional Neural Networks (CNNs) for spatial feature extraction with LSTM units for sequential data processing, demonstrated superior performance, with an accuracy of 79.61% on the overall dataset and 79.37% on the original SSBD dataset. The CNN+GRU model, designed to leverage convolutional layers for feature extraction and Gated Recurrent Units (GRUs) for temporal dependencies, achieved an accuracy of 62% on the overall dataset and 53% on the original SSBD dataset.

These results indicate that while there remains scope for performance enhancement, the LRCN model, in particular, exhibits significant potential for accurately identifying stereotypical behaviors in children with autism from video recordings. This capability could substantially improve the speed and reliability of early autism interventions, especially in settings where professional medical expertise is not readily accessible.

The study underscores the critical importance of utilizing an expanded dataset to enhance the precision of automated identification models for early autism detection. By leveraging a more comprehensive dataset, the models demonstrated improved capacity to identify stereotypic behaviors, thereby enhancing accuracy and robustness. Moreover, the findings highlight the potential for rapid preliminary identification of autism symptoms by non-specialists such as parents, caregivers, and educators. This capability could facilitate earlier interventions and support for children, thereby mitigating the challenges associated with delayed diagnosis.

6. CONCLUSION

This study provides compelling evidence that expanding the dataset used to train deep learning models markedly improves the accuracy of automated systems designed for the early detection of autism spectrum disorder (ASD). Among the evaluated models, the Long-term Recurrent Convolutional Network (LRCN) demonstrated superior accuracy, underscoring its considerable potential for practical application in early autism detection, especially within resource-constrained settings.

The findings of this research highlight the critical role of equipping non-specialists, such as parents and caregivers, with advanced tools powered by these models. Such empowerment can significantly enhance the timeliness and effectiveness of early identification and intervention efforts, which are crucial for optimizing developmental outcomes in children with autism.

Future research directions should emphasize the continuous expansion of datasets through automated collection techniques and focus on iterative model refinement to further enhance detection capabilities. These advancements are expected to not only make autism diagnosis more accessible but also broaden the involvement of non-specialist individuals in the early detection process, thereby improving long-term prognoses for children with ASD.

REFERENCES

[1] F. Negin, B. Ozyer, S. Agahian, S. Kacdioglu, and G. T. Ozyer, "Vision-assisted recognition of stereotype behaviors for early diagnosis of autism spectrum disorders," *Neurocomputing*, vol. 446, pp. 145-155, 2021.

[2] H. Alkahtani, Z. A. Ahmed, T. H. Aldhyani, M. E. Jadhav, and A. A. Alqarni, "Deep learning algorithms for behavioral analysis in diagnosing neurodevelopmental disorders," *Mathematics*, vol. 11, no. 19, p. 4208, 2023.

[3] E. Fernell, M. A. Eriksson, and C. Gillberg, "Early diagnosis of autism and impact on prognosis: A narrative review," *Clinical Epidemiology*, pp. 33-43, 2013.

[4] J. I. Kim, and H. J. Yoo, "Diagnosis and assessment of autism spectrum disorder in South Korea," *Journal of the Korean Academy of Child and Adolescent Psychiatry*, vol. 35, no. 1, p. 15, 2024.

[5] S. Lee, and J. Moon, "A review of early screening for autism spectrum disorder," *Journal of Psychological Movement*, vol. 8, no. 1, pp. 55-79, 2022.

[6] C. Yoon, "Standardization of the Korean-Autism Diagnostic Scale and development of a web-based evaluation system," *Journal of Emotional and Behavioral Disorders*, vol. 20, no. 3, pp. 27-43, 2004.

[7] J. Choi, J. Park, J. Kim, and M. Kim, "Validation study of early diagnosis tools for autism spectrum disorder based on smart platform," *Journal of Autism and Developmental Disorders*, vol. 22, no. 1, pp. 137-156, 2022.

[8] H. Sunwoo, D. Noh, K. Kim, J. Kim, and H. Yoo, "Perceptions of professionals related to early screening for autism spectrum disorder," *Journal of Child and Adolescent Psychiatry*, vol. 28, no. 2, pp. 96-105, 2017.

[9] I. Lovaas, C. Newsom, and C. Hickman, "Self-stimulatory behavior and perceptual reinforcement," *Journal of Applied Behavior Analysis*, vol. 20, no. 1, pp. 45-68, 1987.

[10] P. Wei, D. Ahmedt-Aristizabal, H. Gammulle, S. Denman, and M. A. Armin, "Vision-based activity recognition in children with autism-related behaviors," *Heliyon*, vol. 9, no. 6, p. e12345, 2023.

[11] D. L. Robins, D. Fein, M. L. Barton, and J. A. Green, "The Modified Checklist for Autism in Toddlers: An initial study investigating the early detection of autism and pervasive developmental disorders," *Journal of Autism and Developmental Disorders*, vol. 31, pp. 131-144, 2001.

[12] A. Anzulewicz, K. Sobota, and J. T. Delafield-Butt, "Toward the autism motor signature: Gesture patterns during smart tablet gameplay identify children with autism," *Scientific Reports*, vol. 6, no. 1, p. 31107, 2016.

[13] A. Lakkapragada, A. Kline, O. C. Mutlu, K. Paskov, B. Chrisman, N. Stockham, et al., "The classification of abnormal hand movement to aid in autism detection: Machine learning study," *JMIR Biomedical Engineering*, vol. 7, no. 1, p. e33771, 2022.

[14] S. Rajagopalan, A. Dhall, and R. Goecke, "Self-stimulatory behaviours in the wild for autism diagnosis," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 755-761, 2013.

[15] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[16] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625-2634, 2015.