

Integrating CLAHE and H-MHI with CNN for prompt recognition in children with a diagnosis of autism

Indah Werdiningsih^{1,2}, Ira Puspitasari^{2,3}, Rimuljo Hendradi²

¹Doctoral Program of Mathematics and Natural Sciences, Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

²Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

³Research Center of Quantum Engineering Design, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

Article Info

Article history:

Received Oct 22, 2025

Revised May 4, 2026

Accepted Sep 4, 2026

Keywords:

Autism
Canny edge detection
Contrast limited adaptive
histogram equalization
Disabilities
Otsu thresholding

ABSTRACT

Individualized therapeutic prompts such as physical, gestural, or verbal cues are necessary to enable children diagnosed with autism to become independent in daily activities. However, the automated detection of the prompts in naturalistic conditions is difficult. The solutions currently proposed also have difficulty with unstructured video data obtained from real-world therapeutic settings with different camera angles, uneven lighting, and random movement of children. Therefore, this study aims to introduce a new model that combines contrast limited adaptive histogram equalization (CLAHE) along with hybrid motion history image (H-MHI) to extract motion-based features with convolutional neural networks (CNNs) to classify features and provide timely responses in the context of autism therapy. The H-MHI method was based on the regular MHI with the addition of Otsu thresholding and Canny edge detector. These features were subsequently categorized with the CNN architectures in the form of VGG19 and MobileNetV2. The suggested algorithm was tested on 1,083 video samples and the results showed the ability of H-MHI to enhance accuracy by 2-3% at less computing time than the conventional MHI. The trend reflected the suggested method's effectiveness in enhancing recognition performance without compromising the computational efficiency.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Ira Puspitasari
Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga
Surabaya, Indonesia
Email: ira-p@fst.unair.ac.id

1. INTRODUCTION

Autistic children are often faced with the problems of independently conducting daily activities which require acting with proper assistance [1]. However, current interventions are aimed at developing the most basic skills of prompting such as physical, gestures, or words in line with the routines related to eating and drinking. Some other interventions are also focused on the enhancement of application process across different settings [2]. Eating and drinking are important daily activities [3] but children with autism tend to have an atypical routine such as refusal to eat, selective consumption, slow chewing process, or temper tantrums [4]. Caregivers and therapists are often required to monitor the behaviors but the traditional methods of observation are time-consuming, subjective, and demand continuous presence. The trend shows the non-feasibility of applying these methods in the real world.

The most recent innovations in artificial intelligence (AI) have been very important in the autism spectrum disorder (ASD) diagnosis and treatment as well as the trend emphasizes the importance in the current

healthcare system [5]. Explicable AI methods have also been presented to enhance transparency and credibility in clinical implementations [6]. These include deep learning methods like convolutional neural networks (CNNs) which have proved to be excellent in classifying images, recognizing activities, and detecting patterns [7]. However, the application of CNNs to raw video data of children with autism often does not produce the best results due to several challenges such as poor lighting, fine details or fleeting motion patterns, and cluttered backgrounds that are common in uncontrolled natural environments. Furthermore, most existing studies focus on activity recognition rather than prompt recognition [8], [9], which plays a crucial role in autism interventions. Furthermore, limited research integrates motion-based preprocessing techniques with CNN architectures in this domain.

Research by Singh *et al.* [8], general footage of people with autism engaging in different activities with their parents was used to diagnose autism. However, this study did not incorporate video preprocessing or enhancement techniques. The enhancement methods are carefully selected and considered to have a significant influence on CNN performance [10]. A previous study [9] classified the activities of children with autism by using a combination of enhancement methods with CNN and the highest accuracy was achieved with the inclusion of contrast limited adaptive histogram equalization (CLAHE). The eating and drinking behaviors' sequences were also examined but the scope excluded the recognition of specific prompts considered most important to the development of intervention strategies [11]. Moreover, the method developed by [9] did not include the feature extraction which led to a significantly longer computation time. This study aims to address the limitations by presenting a motion-enhanced CNN model for prompt recognition which is important in autism treatment. A dedicated feature extraction step is also included in the model to enhance efficiency.

The focus is on identifying cues during eating and drinking behaviors by combining CNN with CLAHE and hybrid motion history image (H-MHI). The purpose of developing H-MHI from traditional MHI with Otsu thresholding and Canny edge detector is to improve the motion's representation. Therefore, the combination of CLAHE and H-MHI is expected to ensure more accurate recognition, shorten the time of computation, and generate more discriminatory motion features. The aim is to ensure timely recognition of the cues in children with autism. This is based on the ability of CLAHE to increase local image contrast to emphasize the visual features of interest [12]. Motion-related preprocessing is also considered important in enhancing CNN. The phenomenon leads to the introduction of MHI as a motion capture system to summarize the sequences into a single image [13], [14] but the method is vulnerable to noises [15]. Otsu thresholding and Canny edge detector are subsequently adopted as segmentation methods to isolate foreground and background and refine motion edges in order to overcome the challenges in MHI [16]–[19]. This allows H-MHI to offer a more powerful and informative motion representation for prompt recognition through the combination of the methods.

The background information shows the two contributions of this study with the first being the data. The eating and drinking video footage were captured in the uncontrolled natural environments by the parents or therapists. The data obtained is complex and unstructured because the camera angles are varied and there are several interruptions due to the spontaneous behavior of the children. The second contribution is related to the methodology which includes the combination of CNN with CLAHE and H-MHI to ensure timely recognition of eating and drinking activity with the aim of improving the output. The H-MHI, an extension of the standard MHI, combines Otsu thresholding and Canny edge detection to generate sharper and more discriminative motion features.

2. METHOD

There are seven stages in this study: collecting the data, labeling the videos, converting them into grayscale, the use of CLAHE, extracting the feature, recognizing prompts, and evaluating the results. Feature extraction uses the proposed H-MHI technique, while recognition is carried out using the visual geometry group (VGG) 19 and MobileNetV2 architectures. Metrics such as accuracy, recall, precision, F1-score, and AUC/ROC are utilized to evaluate recognition. Figure 1 depicts the steps of the investigation.

2.1. Video acquisition

The datasets used in this investigation were acquired according to the same technique used in our earlier work [9]. The videos were recorded using a 64-MP main camera with an f/1.8 aperture at a resolution of 1920×1080 pixels and a frame rate ranging from 25 to 30 frames per second (fps).

In brief, the primary data are footage of the activities of children with autism (i.e., eating and drinking). The data were collected from a special education school and a developmental disabilities center in East Java, Indonesia. Eight pupils from the Public Special Needs School (SLB) in Indonesia and ten from the Regional Technical Implementation Unit (UPTD) for youngsters with Special Needs (ABK) in Indonesia were among the 18 youngsters, ages 4 to 12, who took part in the study. The final datasets comprise 1,083 video recordings, with 606 videos depicting eating activities and 477 videos depicting drinking activities.

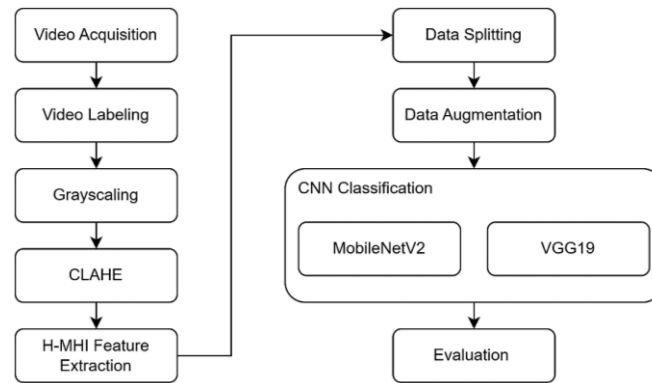


Figure 1. The study stages

2.2. Video labeling

The dataset consisted of eating and drinking activities' full-length and pre-segmented videos. The eating was subdivided into eight steps and drinking into six [9]. The segmentation and labelling of full recordings were conducted while the pre-segmented clips were tagged through direct guidance from educators and therapists. If uncertainty arose during the video labeling process, discussions were held with the therapists and educators to determine the most appropriate prompt category. Each of the 18 participants contributed one or more types of prompts, including physical, gestural, and verbal. The physical prompt was defined as a literal hand-over-hand assistance provided in making the right move. The gestural prompt was explained as a nonverbal signal such as pointing or other indicative gestures reflecting the activity to be conducted. Meanwhile, the verbal prompt was the instructions or prompts provided through words such as questions or keys to elicit the desired behavior [20]. The 606 eating activity videos obtained were subsequently divided into 159 as physical, 97 as gestural, and 350 as verbal prompts. The same method was applied to the 477 drinking activity videos with 98, 75, and 304 respectively.

2.3. Grayscale

Bit depth shows the amount of pixel intensity which determines the brightness levels of a grayscale image. An 8-bit picture, for instance, contains 256 grey levels, ranging from 0 (black) to 255 (white). The video is converted into frames at its frame rate of 25–30 fps. Each extracted frame is then transforming the results to grayscale [21].

2.4. Contrast limited adaptive histogram equalization

CLAHE and H-MHI were combined in this study and the process required the transformation of the initial frame to grayscale where the brightness values were presented without the color. The product was refined with CLAHE to increase local contrast towards making the visual features of the images sharper and clearer. CLAHE was reported to have been successful in enhancing image contrast through the use of local contextual information with the darker and brighter areas that previously had obscured details made clearer and more visually understandable [22]. CLAHE was applied using a clip limit of 0.2 and a tile grid size of 8×8 .

2.5. Hybrid motion history image feature extraction

H-MHI is a variation of the standard MHI produced by combining Otsu thresholding and the Canny edge detector. The difference between H-MHI and the conventional MHI is presented in Figure 2, the stage of H-MHI is presented in Figure 2(a), whereas the conventional MHI is presented in Figure 2(b). The difference between H-MHI and the conventional MHI only Otsu threshold-based segmentation is applied to develop the Ψ image which is further summed to produce the motion energy image (MEI) and MHI. Noise sensitivity is identified as an issue with the traditional pipeline particularly at the edges of objects which lowers the quality of motion representation. Therefore, this study proposes a modified flow to overcome the challenge. The first process was the application of frame differencing on the original product from the CLAHE image (img_{CL}) to obtain more detailed variations across the frames. The second was the production of c_fill using the Canny edge detector with thresholds calculated through Canny thresholds: lower threshold = $200 T_0$ and upper threshold = $1000 T_0$, where T_0 is the Otsu threshold. The result was subsequently used to shape Ψ with the aim of improving the crispness of motion boundaries. Third, v_fill is produced through Gaussian filtering followed by Otsu thresholding, generating a smoother binary motion map. Finally, the proposed MHI is constructed by

combining Ψ (derived from frame differencing and c_fill and v_fill , in contrast to the standard MHI pipeline, which depends only on Otsu-based Ψ .

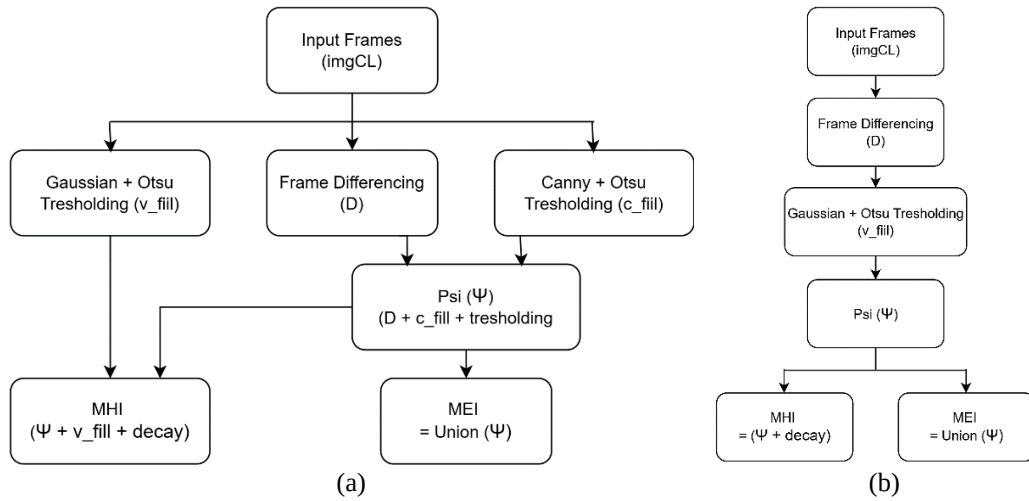


Figure 2. The comparison between H-MHI and standard MHI; (a) H-MHI and (b) standard MHI

The combination of two binary image representations is what makes H-MHI unique— c_fill from Canny edge detection and v_fill from Otsu thresholding—into the construction of the MHI, which has not been implemented in standard MHI approaches. The pipeline of our proposed approach is presented below:

- a. Input. A sequence of video frames from CLAHE image ($imgCL$) is provided as input.
- b. Preprocessing—generates two different images:
 - v_fill : each frame is converted to float, smoothed using a Gaussian filter (G_σ) as in (1). The result is stored in v_fill , as in (2):

$$G_t(x, y) = G_\sigma(imgCL_t(x, y)) \quad (1)$$

$$v_fill(x, y, t)G_t(x, y) = \begin{cases} 1, & G_t(x, y) > T_0 \\ 0, & G_t(x, y) \leq T_0 \end{cases} \quad (2)$$

- c_fill : canny edge detection is applied to ($imgCL$) with thresholds adapted from Otsu's result. The output is stored in c_fill , as in (3):

$$c_{fill,t} = Canny(imgCL_t; 200T_0, 1000T_0) \quad (3)$$

- c. Frame differencing (D). Differences between consecutive frames ($imgCL[t] - imgCL[t-1]$) are calculated, as in (4) and stored in D_list .

$$D_t(x, y) = imgCL_t(x, y) - imgCL_{t-1}(x, y) \quad (4)$$

- d. Ψ (Psi). Ψ_t is generated by thresholding the absolute frame difference D_t using a predefined threshold ξ and combining the result with the Canny edge image $c_{fill,t}$. The resulting Ψ_t is stored in Psi_{list} , as in (5):

$$\Psi_t(x, y) = \begin{cases} 1, & (|D_t(x, y)| > \xi) \vee c_{fill,t} = 1 \\ 0, & \text{Otherwise} \end{cases} \quad (5)$$

- e. Motion energy image (MEI, E). Ψ values are accumulated within a temporal window (τ), producing the MEI, which is stored in E_list . MEI is calculated using (6):

$$E_\tau(x, y, t) = \begin{cases} 1, & \sum_{i=t-\tau+1}^t \Psi_i(x, y) > 0 \\ 0, & \text{Otherwise} \end{cases} \quad (6)$$

- f. Hybrid motion history image (H-MHI, H):

- H-MHI Initialized with $v_fill[0]$, as in (7):

$$H_0(x, y) = v_fill(x, y, 0) \quad (7)$$

– For each subsequent frame, the H-MHI is updated according to: the updated results are stored in H_list .

$$H_t(x, y) = \Psi_t(x, y) \tau + |1 - \Psi_t(x, y)| \max |0, H_{t-1}(x, y) - \delta| \quad (8)$$

where δ is the decay parameter. The updated H-MHI images are stored in H_list , as in (8).

g. Output. Otsu-threshold images (v_fill), Canny edge images (c_fill), Ψ images (Psi_list), MEI, and H-MHI (H_list) are outputs.

2.6. Data splitting

The dataset was evaluated using 5-fold cross-validation, where the data were randomly divided into five subsets. In each iteration, four subsets (approximately 80%) were used for training and one subset (approximately 20%) was used for testing. The partitioning was performed at the video level by randomly assigning video samples to one of the five folds.

2.7. Data augmentation

To boost sample variety and enhance the model's capacity extend to new cases, data augmentation was only used on the training set in each fold. Three augmentation methods were used: shearing (20% intensity) to imitate perspective changes, zooming (20% range) to add scale invariance, and horizontal flipping (randomly turning the pictures left or right) to add orientation variety.

2.8. Convolutional neural network classification

The VGG19 model, consisting of 19 layers with 3×3 convolutional filters, was fine-tuned using pre-trained ImageNet weights [23] with input images resized to 224×224 pixels. The pre-trained feature extraction layers were frozen to preserve learned representations. The head of the classification was a flattened layer followed by two fully connected layers containing 4,096 neurons each and the generalization was enhanced through batch normalization and dropout. A single output layer with three neurons was subsequently used to produce immediate classification. Moreover, MobileNetV2 was utilized because of its efficiency and the high level of accuracy [24] with ImageNet domain pre-trained weights and 224×224 input scale. The pre-trained layers were frozen and global average pooling (GAP) was used before the classification layer applied to incorporate the batch normalization and dropout. This was followed by the adoption of the three-neuron output layers to predict eating and drinking prompts.

The CNN models were developed using Python 3.9.11 and TensorFlow 2.13.0 in visual studio code and trained on a Windows 11 (64-bit) equipped with an AMD Ryzen 7 5800H processor, 16 GB RAM, and an NVIDIA GeForce RTX 3050 Ti Laptop GPU with 4 GB GDDR6 VRAM.

The set of hyperparameters specified in advance during model training was determined through experimentation. The same values were used throughout the proposed models and all folds to provide a fair comparison. The configurations employed in this study are illustrated in Table 1.

Table 1. Hyperparameter values applied

Hyper parameter	Value
Learning rate	0.0001
Early stopping	Epochs=100
Batch size	32
Optimizer	Adam
Loss	Categorical cross-entropy

The two datasets including eating and drinking were used to test the proposed models. The images were enhanced through CLAHE and motion-based feature extraction was conducted using H-MHI. The framework was subsequently validated using two CNN architectures which were VGG19 and MobileNetV2. Furthermore, the four experimental settings conducted were assessed through the assistance of a 5-fold cross-validation. To assess the suggested approach's effectiveness, the best outcomes were contrasted with the standard MHI procedure. The experimental scenarios utilized in this study are demonstrated in the following Table 2.

Table 2. Experimental scenarios

Scenario	Dataset	Image enhancement	Feature extraction	Model
1	Eating	CLAHE	H-MHI	VGG19
2		CLAHE	H-MHI	MobileNetV2
3	Drinking	CLAHE	H-MHI	VGG19
4		CLAHE	H-MHI	MobileNetV2

2.9. Evaluation

Four assessment measures based on the confusion matrix were used to assess the model's performance: accuracy, precision, recall, and F1-score [7]. Area under the curve (AUC) and receiver operating characteristic (ROC) curve were also used to assess the capability.

3. RESULTS AND DISCUSSION

The purpose of this section is to compare the model with the state-of-the-art methods and outline the primary contributions. The H-MHI's and the experimental recognition system's results in the form of a CNN with CLAHE and H-MHI were also recapped.

3.1. Hybrid motion history image feature extraction

Figure 3 shows the feature extraction pipeline for one of the sample videos from the eating dataset containing 124 frames. The method applied to frames 20-40 was the same as the trend in v_fill which used Gaussian filtering and Otsu thresholding as presented in Figure 3(a), c_fill through the Canny edge detector with Otsu thresholding in Figure 3(b), and frame differencing with c_fill to produce Ψ in Figure 3(c). The accumulated Ψ sequence was subsequently added to the MEI in Figure 3(d) and multiplied by v_fill to obtain the final image of H-MHI in Figure 3(e) and the standard of MHI image in Figure 3(f).

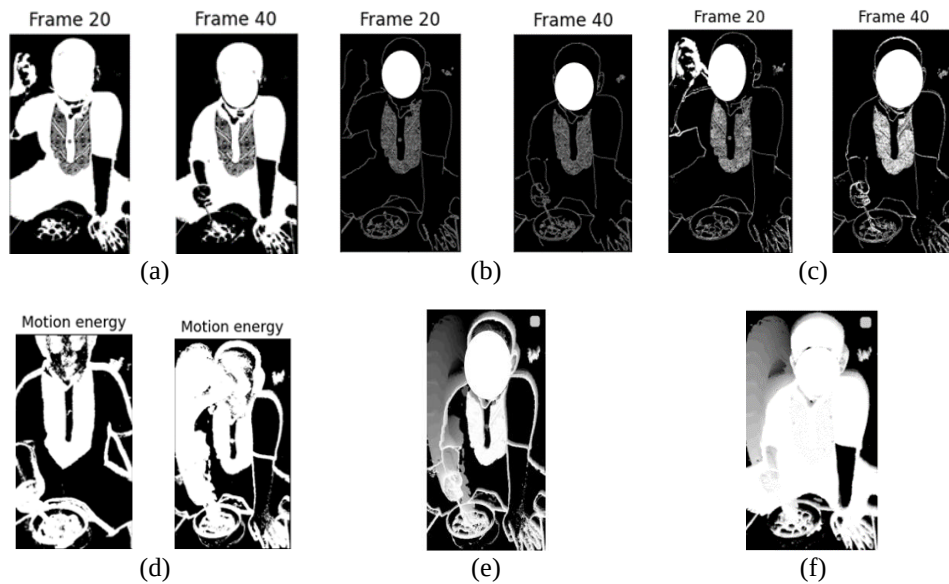


Figure 3. H-MHI pipeline; (a) Otsu-thresholding images (v_fill), (b) Canny edge images (c_fill), (c) Ψ images, (d) MEI image, (e) H-MHI image, and (f) standard MHI image

H-MHI is a representation of motion with an accent on edge features and temporal dynamics in terms of color gradients which allows the differentiation of salient movement patterns more clearly. This representation assisted in accentuating behaviorally salient behavior such as hand-to-mouth patterns considered significant for quick recognition. The combination of v_fill and c_fill in Figure 3(e) which led to the adoption of the Canny edge detector and Otsu thresholding to guide the edges made the motion boundaries become sharper and more discriminative. This was due to the combination of the region-based segmentation with accurate edge information. Meanwhile, Figure 3(f) shows that the v_fill is determined by Gaussian filtering and Otsu thresholding provides a smoother binary representation of motion regions. This method led to better and robust changes in illumination but tended to miss fine motion information, particularly along object boundaries.

3.2. Results of convolutional neural network classification

During 5-fold cross-validation, the model was trained on each training subset and tested on the remaining data. Table 3 summarizes of training scenarios using 5-fold cross-validation. For each scenario, the reported results represent the average performance across all folds.

Table 3. Summary of training scenarios

Scenario	Dataset	Training		Testing	
		Accuracy (%)	Loss	Accuracy (%)	Loss
Scenario 1 (CLAHE-H-MHI+VGG19)	Eating	89.8±1.72	0.2397±0.03	59.26±1.79	1.2052±0.03
Scenario 2 (CLAHE-H-MHI+MobileNetV2)		96.58±1.36	0.0516±0.02	62.71±1.50	1.3247±0.02
Scenario 3 (CLAHE-H-MHI+VGG19)	Drinking	95.44±1.62	0.0690±0.03	67.61±1.72	1.0448±0.03
Scenario 4 (CLAHE-H-MHI+MobileNetV2)		98.18±1.17	0.0179±0.01	79.97±1.36	1.0527±0.02
Scenario 5 (CLAHE-MHI+VGG19)	Eating	87.12±1.83	0.2821±0.03	53.94±1.85	0.9205±0.03
Scenario 6 (CLAHE-MHI+VGG19)	Drinking	87.55±1.67	0.3161±0.02	68.73±1.94	1.3009±0.03

The final performance of all scenarios was evaluated using the average across all folds in 5-fold cross-validation, as presented in Table 4. Figure 4 shows the testing accuracy comparison between MHI and H-MHI in this study. Scenario 1 achieved the highest accuracy of 71%±1.79% (543 seconds), while scenario 2 yielded the lowest accuracy at 65±1.41 (533 seconds). Scenario 3 attained the highest accuracy of 74%±1.67% (526 seconds), whereas scenario 4 resulted in the lowest accuracy of 65.2±1.17 (514 seconds). Overall, the VGG19-based models (scenarios 1 and 3) outperformed their MobileNetV2 counterparts (scenarios 2 and 4). Therefore, 2 scenarios were added to further compare the effect of H-MHI and MHI on our approaches. Specifically, VGG19 integrated with CLAHE and the standard MHI were applied to the Eating and Drinking datasets as scenarios 5 and 6, respectively. Scenarios 1 and 3 achieved higher accuracy with reduced computation time compared to Scenarios 5 and 6, reaching 68.6%±1.85 (1,108 seconds) and 71.4%±1.74 (820 seconds), respectively. The Wilcoxon signed-rank test showed no statistically significant difference between the H-MHI and MHI configurations for either the eating dataset ($p=0.3125$) or the drinking dataset ($p=0.0625$).

Table 4. Experimental scenario results

No	Scenario	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC/ROC	Time (seconds)	Wilcoxon p-value
1	CLAHE-H-MHI+VGG19	Eating	71±1.79	70.4±1.84	70.2±1.72	70.6±1.74	81.0±0.02	543	0.3125 ^a
2	CLAHE-H-MHI+MobileNetV2		65±1.41	65.2±1.72	65.4±1.36	65.4±1.50	72.2±0.04	533	-
3	CLAHE-H-MHI+VGG19	Drinking	74±1.67	73.6±1.85	73.8±1.72	73.4±1.62	82.6±0.02	526	0.0625 ^b
4	CLAHE-H-MHI+MobileNetV2		65.2±1.17	66±1.67	65.2±1.33	65.4±1.36	68.9±0.04	514	-
5	CLAHE-MHI+VGG19	Eating	68.6±1.85	68±1.90	68±1.83	67.6±1.85	74.8±0.03	1,108	0.3125 ^a
6	CLAHE-MHI+VGG19	Drinking	71.4±1.74	70.6±1.96	70.8±1.94	71±1.67	71.1±0.03	820	0.0625 ^b

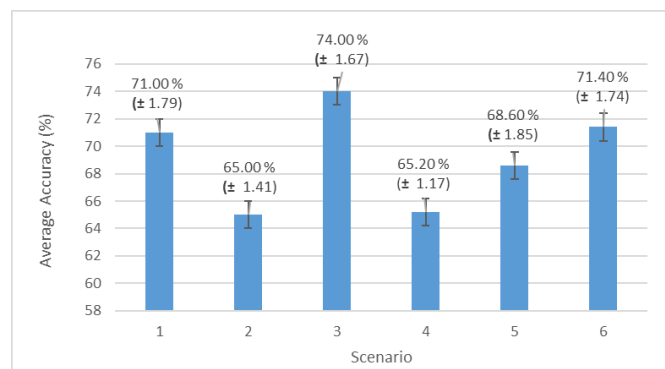


Figure 4. Comparison of testing accuracy between H-MHI and MHI

VGG19 achieves the best performance across all metrics, while MobileNetV2 shows lower variability in classification-based metrics, indicating more stable predictions. However, VGG19 demonstrates more consistent AUC values, suggests stronger and more stable class separability. Figure 5 present the AUC/ROC curve from a representative fold for the eating and drinking dataset. Figure 5(a) presents the ROC curve in scenario 1 with the highest AUC value of 0.91 for physical prompt and Figure 5(b) presents the ROC curve in scenario 3 with the highest AUC value of 0.92 for physical prompt.

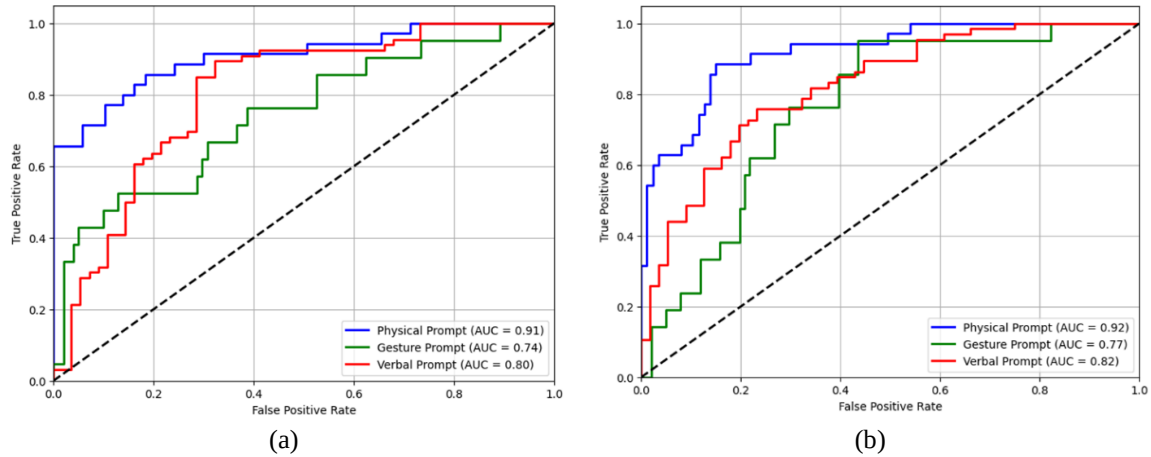


Figure 5. AUC/ROC curve; (a) Scenario 1 and (b) Scenario 3

To evaluate the effectiveness of the proposed models, classification metrics were computed. Based on the results in Table 4, the confusion matrix values for the best-performing scenario are presented in Table 5. The results showed that the model successfully recognized verbal. However, misclassification occurred between the gestural and verbal, with gestures sometimes being classified as verbal and vice versa. This indicates overlapping movement characteristics between the gestural and verbal.

Table 5. The confusion matrix for Scenarios 1 and 3

Actual\predicted	Scenario 1			Scenario 3		
	Physical	Gestural	Verbal	Physical	Gestural	Verbal
Physical	25	3	7	16	1	8
Gestural	2	10	9	0	4	6
Verbal	4	8	54	5	3	53

Compared to physical and verbal methods, gesture cues have a higher degree of ambiguity. Physical methods are able to provide evident, direct directions, while verbal methods offer explicit linguistic meanings that facilitate the interpretation process. On the other hand, the process of labeling gesture data faces significant challenges because it relies on visual interpretation of movements, which increases the risk of annotation errors and potentially decreases model accuracy. This challenge is exacerbated by the characteristics of gestures in children with autism, which are often faint, irregular, and highly variable between individuals, making them difficult to interpret consistently [25]. As a result, there is overlapping movement patterns in the H-MHI representation—especially between gestures and verbal cues—which ultimately increases the risk of misclassification.

3.3. Comparison of the proposed approach

A comparison of the suggested approach and previous research on ASD activity identification is shown in Table 6. Transformer-based methods perform well in a research in [26] that uses video-based analysis for behavior identification in ASD because they can represent intricate spatiotemporal patterns. Nevertheless, these models usually need more processing power, which makes them less appropriate for real-time use. In contrast, the proposed CNN-based approach offers a more computationally efficient solution while maintaining competitive performance. Similarly, the study in [27] demonstrates the effectiveness of CNN-based models for recognizing activities in ASD intervention settings. However, it does not address fine-grained motion patterns or computational efficiency. In contrast, the proposed method focuses on specific daily behaviors and

emphasizes real-time performance. The vision-assisted approach in [28] achieves an accuracy of approximately 79% using motion-based descriptors. However, the study did not describe computational efficiency which led to the difficulty in determining the suitability for real-time deployment. Another similar effort in [9] used CNN with image enhancement methods to recognize daily activities in children with ASD. The method enhanced visual features based on space but was unable to capture motion dynamics over time. However, the method combined the spatial (CLAHE) and temporal (H-MHI) information which allowed the extraction of more useful discriminative spatio-temporal features. The accuracy was observed to be higher but the computational efficiency was not assessed which was part of the main gaps bridged in this study.

The method proposed generally had a good trade-off between accuracy along with computational efficiency. This was because the combination of CLAHE and H-MHI improved motion representation and allowed real-time performance. The ability to quickly and accurately recognize everyday activities is especially important for assistive applications that help children with autism become more autonomous.

Table 6. Comparison of the proposed approach

Reference	Dataset	Baseline architecture	Performance	Time (seconds)
[26]	Self-stimulatory behavior dataset (SSBD)	Transformer	AUROC=93–95%	Not reported
[27]	Multimodal-autism intervention dataset (MMASD)	LSTM	Accuracy=85–95.72%	Not reported
[28]	Expanded stereotype behavior dataset (ESBD)+SSBD	Histogram of optical flow (HOF)+MLP	Accuracy=79%	Not reported
[9]	Eating and drinking activity	CLAHE+CNN	Accuracy=85%	39,448
Proposed approach (scenario 3)	Eating and drinking activity	CLAHE and H-MHI with VGG19	Accuracy=76%	526

3.4. Discussion

The results showed the presence of overfitting because training accuracies were more than 90% and validation accuracies were less than 70%. The model had a significant discrepancy between training and validation performances despite the use of dropout, data augmentation, and transfer learning. The trend was explained by the relatively small sample, class imbalanced, and natural complexity of quick recognition within non-structured settings. Furthermore, this study employs 5-fold cross-validation at the video level by randomly partitioning video samples into distinct folds; this approach means that videos from the same participant may appear in both the training and testing sets, potentially leading to overly optimistic estimates of the model's generalization performance. Future research will investigate participant-based cross-validation methods to provide a more rigorous evaluation of the proposed framework.

The combination of CLAHE and H-MHI with VGG19 was more accurate with MobileNetV2. This was due to the improved preprocessing features of H-MHI over the normal MHI. The phenomenon was associated with the opportunity provided by H-MHI to allow further differentiation between objects and the background through the assistance of the adaptive threshold selection obtained by including Otsu thresholding. A Canny edge detector was also integrated to maintain important image contours which led to more informative motion patterns [17]. These methods combined enhance motion representation which was considered important for improving the capability of the model in identifying prompts. Moreover, the better performance was also explained by the VGG19 architecture that possessed a more profound and highly complicated layer structure compared to MobileNetV2. The process enabled VGG19 to capture more detailed spatio temporal features [23] which led to more accuracy but MobileNetV2 provided a more computationally efficient alternative and was considered more suitable in resource-constrained settings. By using depthwise separable convolutions, whereby split ordinary convolution into depthwise as well as pointwise operations and drastically reduce the amount of parameters and floating-point operations, MobileNetV2 lowers computational costs [29]. Although H-MHI achieved higher mean accuracy than MHI, the differences were not statistically significant, possibly due to the limited dataset size and number of cross-validation folds.

The implementation of H-MHI can significantly decrease the computation time of the VGG19 model. For example, the operation of the VGG19 model with CLAHE required 39,448 seconds as reported in [9] but the inclusion of H-MHI reduced the time to 526 seconds. This large saving in computation time was possible because H-MHI compressed the video sequence of movements into a single compound image. The process eliminated the need to work with each of several frames and retained the motion dynamic between movements in a single image [14].

The accuracy obtained in this study was less than the value in [9] but the results had a critical benefit of significantly decreasing the time of computation from 39,448 seconds [9] to the 526 seconds. This decrease in processing time does not reflect a performance cost because the method has the highest accuracy relative to

the other experiments performed. The enhancement in computation efficiency increases the possibility of performing real-time detection which is considered important in real life. Moreover, the reduction of the processing time has a direct positive influence on scalability [30] and enables the system to deliver a quick response to the daily routines of children with autism. Computational burden is also significantly reduced through the decrease in the number of frames required to be processed into a single image per video by MHI or H-MHI. This method can cause a certain loss of time information and a relative reduction in classification accuracy but is identified as a fair trade-off in real-world settings where resources are limited and time interaction is very important.

The discussion shows that H-MHI requires less time to compute than methods without MHI and is more accurate than conventional MHI. The combination of Otsu thresholding and Canny edge detector led to the performance gain which produced a succinct and informative motion representation to be used in feature generation. Moreover, both CLAHE and H-MHI complemented each other and worked together to provide better performance. CLAHE enhanced the local contrast which allowed clearer presentation of subtle details of motion while H-MHI focused on the temporal dynamics by underlining the changes across the frames. The combination of the methods enabled the CNN to acquire more discriminative spatio-temporal features to identify the subtle behavioral patterns of the autistic children. The experimental findings on the benchmark datasets demonstrated the efficacy of the suggested strategy. In order to support the creation of reliable assistive technologies that guarantee the independence of autistic children, the technique can be practically applied to identify daily routines in real-time. Future studies can expand the framework to a wider range of activities and prompt categories. It is also possible to adopt explainable AI in increasing the interpretability in addition to the assessment of sophisticated methods such as multimodal and temporal modeling in enhancing prompt recognition.

4. CONCLUSION

In conclusion, this study showed that the suggested framework was effective for early recognition in autism treatment, as follows: i) the main results reported that the combination of CLAHE and H-MHI enhanced the motion representation and recognition accuracy in addition to the achievement of less computation time. VGG19 with CLAHE and H-MHI were identified as the most successful models with $71\% \pm 1.79\%$ (543 seconds) and $74\% \pm 1.67\%$ (526 seconds) accuracy for eating and drinking activity respectively. The trend showed the suitability for application in real-time assistive systems to facilitate more objective and scalable interventions for autism, ii) the limitations identified were possible overfitting of the model, a relatively small dataset size, class imbalanced, bias towards the majority classes, and insufficient generalizability to more diverse real-world situations, and iii) future studies should investigate participant-based cross-validation methods, expand the framework to other types of activities and prompt categories, use explainable AI to promote interpretability, as well as investigate real-time implementation and longitudinal analyses to understand the practical effectiveness.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Indah Werdiningsih	✓	✓	✓	✓	✓	✓		✓	✓	✓				
Ira Puspitasari	✓	✓			✓	✓			✓	✓		✓		
Rimuljo Hendadi	✓	✓			✓	✓			✓			✓		

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nterpretation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**diting

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

There is no conflict of interest.

INFORMED CONSENT

Every participant in this study has given their informed permission. On July 12 and 13, 2023, the informed consent procedure was carried out.

ETHICAL APPROVAL

The investigation was carried out with human participants in accordance with the guiding principles of the Helsinki Declaration. All applicable institutional and national rules were adhered to, and Universitas Airlangga's Faculty of Public Health granted approval (Approval No: 76/EA/KEPK/2023).

DATA AVAILABILITY

Due to ethical and privacy concerns concerning research participants, the data supporting the study's conclusions are not publicly accessible. The dataset contains sensitive information that cannot be shared, in order to protect participant confidentiality. However, detailed descriptions of the data collection process, preprocessing steps, and experimental setup are provided to support reproducibility.

REFERENCES

- [1] P. Ghanouni, S. Quirke, J. Blok, and A. Casey, "Independent Living in Adults with Autism Spectrum Disorder: Stakeholders' Perspectives and Experiences," *Research in developmental disabilities*, vol. 119, pp. 633–6387, 2021, doi: 10.1016/j.ridd.2021.104085.
- [2] A. I. Roncati, A. C. Souza, and C. F. Miguel, "Exposure to A Specific Prompt Topography Predicts Its Relative Efficiency when Teaching Intraverbal Behavior to Children With Autism Spectrum Disorder," *Journal of Applied Behavior Analysis*, vol. 52, no. 3, pp. 739–745, 2019, doi: 10.1002/jaba.568.
- [3] K. Porter *et al.*, "Supporting People Who Have Eating and Drinking Difficulties," *Clinical Medicine*, vol. 21, no. 4, pp. e344–e350, 2021, doi: 10.7861/CLINMED.2021-0161.
- [4] A. Brzóska *et al.*, "Eating Behaviors of Children with Autism—Pilot Study, Part II," *Nutrients*, vol. 13, no. 11, pp. 1–18, 2021, doi: 10.3390/nu13113850.
- [5] N. Wankhede *et al.*, "Leveraging AI for the diagnosis and treatment of autism spectrum disorder: Current trends and future prospects," *Asian Journal of Psychiatry*, vol. 101, pp. 1–8, 2024, doi: 10.1016/j.ajp.2024.104241.
- [6] R. Agrawal and R. Agrawal, "Explainable AI in early autism detection: a literature review of interpretable machine learning approaches," *Discover Mental Health*, vol. 5, no. 1, pp. 1–21, 2025, doi: 10.1007/s44192-025-00232-3.
- [7] D. L. Long, "Efficient DenseNet Model with Fusion of Channel and Spatial Attention for Facial Expression Recognition," *Cybernetics and Information Technologies*, vol. 24, no. 1, pp. 171–189, 2024, doi: 10.2478/cait-2024-0010.
- [8] V. A. Singh, A. Rawat, M. Laroia, and K. R. Seeja, "Autism Spectrum Disorder Screening on Home Videos Using Deep Learning," *International Journal of Image, Graphics and Signal Processing (IJIGSP)*, vol. 16, no. 4, pp. 106–115, 2024, doi: 10.5815/ijigsp.2024.04.08.
- [9] I. Werdiningsih, I. Puspitasari, and R. Hendradi, "Recognizing Daily Activities of Children with Autism Spectrum Disorder using Convolutional Neural Network Based on Image Enhancement," *Cybernetics and Information Technologies*, vol. 25, no. 1, pp. 78–96, 2025, doi: 10.2478/cait-2025-0005.
- [10] N. Mitschke, Y. Ji, and M. Heizmann, "Task Specific Image Enhancement for Improving the Accuracy of CNNs," in *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods ICPRAM*, pp. 174–181, 2021, doi: 10.5220/0010186301740181.
- [11] J. Wang, D. M. Maggin, and K. Zarate, "Peer Prompting and Reinforcement for Supporting Positive Behavioral Development in Students With EBD," *Beyond Behav*, vol. 31, no. 2, pp. 125–135, 2022, doi: 10.1177/10742956211022045.
- [12] I. M. Mohammed and N. A. M. Isa, "Contrast Limited Adaptive Local Histogram Equalization Method for Poor Contrast Image Enhancement," *IEEE Access*, vol. 13, pp. 62600–62632, 2025, doi: 10.1109/ACCESS.2025.3558506.
- [13] P. Pareek and A. Thakkar, "A Survey on Video-Based Human Action Recognition: Recent Updates, Datasets, Challenges, and Applications," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 2259–2322, 2021, doi: 10.1007/s10462-020-09904-8.
- [14] H. Chen, M. C. Leu, and Z. Yin, "Real-Time Multi-Modal Human–Robot Collaboration Using Gestures and Speech," *Journal of Manufacturing Science and Engineering*, vol. 144, no. 10, 2022, doi: 10.1115/1.4054297.
- [15] A. Dhillon and G. K. Verma, "Convolutional Neural Network: A Review of Models, Methodologies and Applications to Object Detection," *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 85–112, 2020, doi: 10.1007/s13748-019-00203-0.
- [16] A. S. Ahmed, "Comparative Study Among Sobel, Prewitt and Canny Edge Detection Operators Used in Image Processing," *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 19, pp. 6517–6525, 2018.
- [17] M. A. Ansari, D. Kurchaniya, and M. A. Dixit, "Comprehensive Analysis of Image Edge Detection Techniques," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 12, no. 11, No 11, pp. 1–12, 2017, doi: 10.14257/ijmue.2017.12.11.01.
- [18] M. T. Nyo, F. Mebarek-Oudina, S. S. Hlaing, and N. A. Khan, "Otsu's Thresholding Technique for MRI Image Brain Tumor Segmentation," *Multimedia Tools and Applications*, vol. 81, no. 30, pp. 43837–43849, 2022, doi: 10.1007/s11042-022-13215-1.
- [19] T. Y. Goh, S. N. Basah, H. Yazid, M. J. A. Safar, and F. S. A. Saad, "Performance Analysis of Image Thresholding: Otsu Technique," *Measurement*, vol. 114, pp. 298–307, 2018, doi: 10.1016/j.measurement.2017.09.052.
- [20] J. L. Seaver and J. C. Bourret, "An Evaluation of Response Prompts for Teaching Behavior Chains," *Journal of Applied Behavior Analysis*, vol. 47, no. 4, pp. 777–792, 2014, doi: 10.1002/jaba.159.

- [21] C. -H. Yeh *et al.*, "Lightweight Deep Neural Network for Joint Learning of Underwater Object Detection and Color Conversion," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6129–6143, Nov. 2022, doi: 10.1109/TNNLS.2021.3072414.
- [22] G. F. C. Campos, S. M. Mastelini, G. J. Aguiar, R. G. Mantovani, L. F. D. Melo, and S. Barbon Jr, "Machine Learning Hyperparameter Selection for Contrast Limited Adaptive Histogram Equalization," *EURASIP Journal on Image and Video Processing*, no 1, 2019, doi: 10.1186/s13640-019-0445-4.
- [23] C. B. R. Killi, N. Balakrishnan, and C. S. Rao, "Deep Fake Image Classification Using VGG-19 Model," *International Information and Engineering Technology Association*, vol. 28, no. 2, pp. 509–515, 2023, doi: 10.18280/isi.280228.
- [24] H. Lokhande and S. R. Ganorkar, "Object detection in video surveillance using MobileNetV2 on resource-constrained low-power edge devices," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 1, pp. 357–365, 2025, doi: 10.11591/eei.v14i1.8131.
- [25] M. G. Logrieco *et al.*, "Nonverbal Skills Evolution in Children with Autism Spectrum Disorder One-Year Post-Diagnosis," *Children*, vol. 11, no. 12, pp. 1–13, 2024, doi: 10.3390/children11121520.
- [26] A. J. R. and R. Senthilkumar, "DSVTN-ASD: Detection of Stereotypical Behaviors in Individuals with Autism Spectrum Disorder using a Dual Self-Supervised Video Transformer Network," *Neurocomputing*, vol. 624, 2025, doi: 10.1016/j.neucom.2025.129397.
- [27] A. Kanwal, A. Hanzala, K. Javed, and T. Ahmed, "ARAIIS-Activity Recognition of Autism Children During Intervention Sessions," in *2024 7th International Conference on Machine Learning and Natural Language Processing (MLNLP)*, Chengdu, China, 2024, pp. 1–6, doi: 10.1109/MLNLP63328.2024.10800500.
- [28] 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, doi: 10.1016/j.neucom.2021.03.004.
- [29] M. Akay *et al.*, "Deep Learning Classification of Systemic Sclerosis Skin Using the MobileNetV2 Model," *Engineering in Medicine and Biology*, vol. 2, pp. 104–110, 2021, doi:10.1109/OJEMB.2021.3066097.
- [30] V. A. Petricia, M. Jacksy, and R. Bhavadharani, "Autism Spectrum Disorder Detection Using Convolutional Neural Network," in *2025 International Conference on Visual Analytics and Data Visualization (ICVADV)*, Tirunelveli, India: IEEE, Mar. 2025, pp. 743–748, doi: 10.1109/ICVADV63329.2025.10961573.

BIOGRAPHIES OF AUTHORS



Indah Werdiningsih    received her B.Sc. degree in Mathematics in 2003 and her M.Eng. degree in Informatics Engineering in 2011, both from Institut Teknologi Sepuluh Nopember (ITS), Indonesia. She is currently pursuing her Ph.D. degree at the Faculty of Science and Technology, Universitas Airlangga, Indonesia. She is also a lecturer with the Information Systems Study Program at Universitas Airlangga, Indonesia. Her research interests include computer vision, machine, deep learning, and expert systems. She can be contacted at email: indah-w@fst.unair.ac.id.



Ira Puspitasari    is an Associate Professor at the Information Systems Study Program, Universitas Airlangga, where she has been a faculty member since 2012. She graduated with honors B.Eng. degree in Informatics Engineering and M.Eng. degree in Informatics, both from Institut Teknologi Bandung, Indonesia. She then completed her Ph.D. in Information and Physical Sciences from the University of Osaka, Japan, in 2015. Her research interests are primarily in the area of data science, consumer health informatics, and IT strategy and strategic planning where she is the author/co-author of over 40 research publications. She can be contacted at email: ira-p@fst.unair.ac.id.



Rimuljo Hendradi    received his B.Sc. and M.Sc. degrees in Mathematics Department, Universtiy of Gadjah Mada, Indonesia, in 1995 and 2002, respectively. He received Ph.D. degree at Electrical Engineering, Institute of Technology Sepuluh Nopember in 2016. He is presently an Assistant Professor at Department of Mathematics, Information Systems Study Program, Faculty of Science and Technology, Universitas Airlangga, Indonesia. His present research interests include the decision support systems, business intelligence, biomedical signal processing and particular in time-frequency phonocardiogram analysis. He can be contacted at email: rimuljohendradi@fst.unair.ac.id.