Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MMVX-Net: Multispectral Mamba Representation Learning for Dysphagia Screening

Derek Ka Hei Lai

derekkh.lai@connect.polyu.hk

Hong Kong Polytechnic University

Jing Li

Hong Kong Polytechnic University

Mi-Ya Po

Hong Kong Sheng Kung Hui

Armstrong Tat-San Chiu

The Hong Kong Society for the Blind

James Chung-Wai Cheung

Hong Kong Polytechnic University

Research Article

Keywords: Dysphagia, Multispectral Video, Deep Learning

Posted Date: October 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11262558/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

MMVX-Net: Multispectral Mamba Representation Learning for Dysphagia Screening

Derek Ka-Hei Lai
Department of Biomedical Engineering
The Hong Kong Polytechnic University
Hong Kong, China
derekkh.lai@connect.polyu.hk

Jing Li
Department of Biomedical Engineering
The Hong Kong Polytechnic University
Hong Kong, China
jing-aggie.li@connect.polyu.hk

Mi-Ya Po
Hong Kong Sheng Kung Hui Li Ka Shing Care and Attention Home For The Elderly
Hong Kong Sheng Kung Hui
Hong Kong, China
mypo@skhwc.org.hk

Armstrong Tat-San Chiu
Kowloon Home for the Aged Blind
The Hong Kong Society for the Blind
Hong Kong, China
armstrong.chiu@hksb.org.hk

James Chung-Wai Cheung
Department of Biomedical Engineering
The Hong Kong Polytechnic University
Hong Kong, China
james.chungwai.cheung@polyu.edu.hk

Abstract— **Dysphagia assessment remains difficult to scale beyond specialist settings. This study introduces MMVX-Net, a Mamba-based framework for binary classification of manually selected, pre-segmented swallowing clips from synchronized RGB-D video. The cohort comprised 136 participants (68 healthy controls and 68 patients with clinically diagnosed dysphagia), and the analyzed dataset contained 282 clips acquired across IDDSI Levels 0, 2, and 4. Modality-specific embeddings and bidirectional cross-modal fusion preserve RGB and depth representations while exchanging contextual information. The architecture has linear sequence-complexity in its state-space blocks and requires 6.47 GFLOPs and 0.69 million parameters for the evaluated 16-frame input. Five-fold cross-validation yielded accuracy of 0.929, precision of 0.927, recall of 0.928, F1-score of 0.927, ROC-AUC of 0.963, and PR-AUC of 0.955. These results support the feasibility of efficient RGB-D classification of curated swallowing clips in a controlled acquisition setting. Participant-disjoint validation, complete dataset accounting, external validation, and automated swallow localization remain necessary before claims of end-to-end clinical screening or deployment can be made.**

Keywords— Dysphagia, Multispectral Video, Deep Learning

I. INTRODUCTION

Dysphagia, or swallowing difficulty, is common among older adults, with prevalence estimates varying substantially across populations [1], [2]. It is associated with aspiration pneumonia, malnutrition, poorer health, and reduced quality of life, making timely screening clinically important [3], [6]. Instrumental assessments such as VFSS and FEES provide detailed evaluation but require specialized equipment and personnel, motivating complementary, accessible, non-invasive approaches [3], [6].

Wearable systems using accelerometry and acoustics have been developed to address these limitations, but their contact-based nature can cause discomfort and reduce compliance, particularly among older adults and those with dementia [4, 5]. For example, dual-axial and tri-axial accelerometers have been widely used to detect hyoid and laryngeal motion, while contact microphones are used to recognize swallowing or aspiration sounds [4, 6].

Software-wise, most systems rely on traditional machine learning algorithms like support vector machines (SVM), decision trees, and linear discriminant analysis (LDA), with features extracted from time and frequency domains [4, 6]. Recent studies have investigated deep learning models including convolutional neural networks (CNN) and transformers, yet challenges remain. Performance varies widely (e.g., sensitivity ranging from 21.7% to 97%) and generalizability is often untested due to small sample sizes (many with <50 subjects) and heterogeneous protocol [4, 6].

Given the limitations of invasive and contact-based tools, camera-based systems, particularly multispectral video, offer a contact-free way to observe visible swallowing-related motion. Unlike wearable sensors, video preserves spatial and temporal surface information without attaching a sensor to the participant. However, surface RGB-D signals should be interpreted as correlates of swallowing motion rather than direct measurements of internal hyoid or laryngeal displacement.

Multispectral video provides complementary information for swallowing analysis. RGB video captures visible cues such as lip, mandibular, and surface motion, whereas depth video provides three-dimensional surface geometry [7] – [9]. In the present study, these modalities are used to model external oral and neck motion during swallowing; the system does not directly measure internal biomechanics.

Deep learning plays a central role in video understanding and has begun to be applied to dysphagia screening. A previous study evaluated the performance of Transformer models for classifying swallowing and non-swallowing tasks using depth video data acquired with an infrared depth camera [10]. This research demonstrated the feasibility of combining depth video with deep learning for swallow classification as a more accessible and cost-effective screening approach compared with traditional VFSS.

Self-attention has quadratic time and space complexity with respect to token sequence length, whereas state-space models provide an alternative sequence-mixing mechanism with linear scaling [11]–[14]. Although such scaling is attractive for longer videos, the present study evaluates only 16 sampled frames and therefore does not establish a long-sequence advantage.

Mamba uses selective state-space updates for content-dependent sequence processing with linear complexity in sequence length [12]–[15]. We use this mechanism as a parameter-efficient sequence mixer within a dual-stream

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

RGB-D architecture; we do not claim that the current 16-frame experiment demonstrates length-scaling benefits.

The architectural motivation is therefore twofold: modality-specific processing can preserve distinct RGB and depth information, and bidirectional cross-modal exchange can integrate complementary surface-motion cues [14], [16]. Recent work also emphasizes the need for clinically interpretable and externally validated dysphagia models [23].

The contributions of this study are: (1) a synchronized RGB-D dataset of swallowing tasks from 68 healthy controls and 68 participants with clinically diagnosed dysphagia; (2) MMVX-Net, a dual-stream bidirectional Mamba architecture with repeated cross-modal exchange; and (3) an internal five-fold evaluation of clip-level dysphagia classification together with model-size and FLOP comparisons. The intended scope is classification of manually selected swallowing clips in the controlled acquisition setting described below.

II. Methods

A. Subject

We recruited 68 healthy participants and 68 participants with clinically diagnosed dysphagia from day centers and hostels (overall mean age: 85 ± 7.4 years). The cohort comprised 32 male and 104 female participants. Participants in the dysphagia group had a clinical diagnosis of swallowing dysfunction and were prescribed dysphagia diets (IDDSI level ≤ 4). The dysphagia status was determined as presence or absence of dysphagia. Participants were excluded if they had a history of neck surgery, tracheostomy, feeding-tube use, or cognitive limitations that prevented understanding of or compliance with the study protocol. This study was approved by the Ethics Committee of the Hong Kong Polytechnic University Institutional Review Board (Reference Number: HSEARS20230302009).

B. Experiment

The experimental swallowing protocol included bolus consistencies corresponding to IDDSI Levels 0, 2, and 4. Healthy participants were eligible to perform all three conditions. For participants with dysphagia, the bolus consistencies administered were determined according to their clinical swallowing status and the judgment of the responsible speech therapist or occupational therapist. If a particular IDDSI level was considered inappropriate or potentially unsafe for an individual participant, that swallowing condition was not performed. Therefore, not every participant with dysphagia completed all three IDDSI conditions.

During the experiment, we recorded the motions of the lower face and neck (including the lips, mandible, and throat) during swallowing tasks. The procedure involved capturing a subject's movements during a series of swallowing tasks using an RGB-D camera system (RealSense D435i, Intel Corp., California, United States) [17]. As shown in Fig. 1, the subjects sat on a chair in a neutral position. The RealSense D435i camera, operating at 30 frames per second with a resolution of 480 × 848 pixels, was mounted on a table and aimed at the subject's neck and mandible regions. It was positioned at a 45-degree angle relative to the horizontal plane, at an approximate distance of 35 cm from the target area. An eye-level marker was placed on the table to serve as a fixed reference point, helping the subject maintain a neutral position. A connected computer unit recorded the acquired data.

Each subject was asked to sequentially swallow three distinct boluses representing specific consistencies defined by the IDDSI framework [18]. The first bolus consisted of plain water, representing IDDSI Level 0. The second and third boluses were juice-based and modified with THICKENUP® (Nestlé, Vevey, Switzerland) instant food and drink thickener—the second bolus was prepared to achieve IDDSI Level 2 (Mildly Thick), while the third bolus corresponded to IDDSI Level 4 (Extremely Thick). The entire process for each food type was recorded for subsequent analysis of swallowing kinematics and characteristics. These procedures were repeated five times to capture data variations and enhance model generalization.

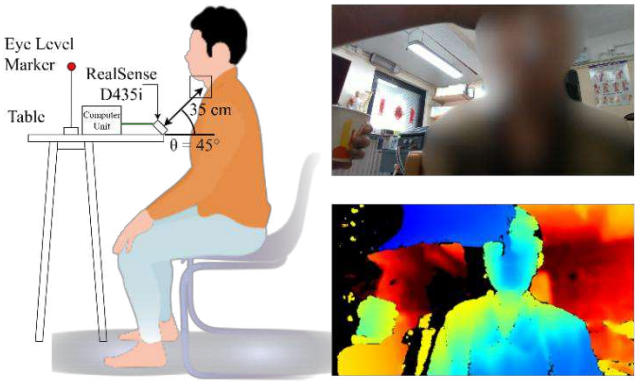


Fig. 1. Fig. 1. The schematic of experiment setup and actual field trial

C. Pre-processing

All recordings underwent manual quality review before dataset construction. Recordings were excluded when the participant was unable or unwilling to continue the experiment, when uncontrolled movement caused occlusion or displacement of the lower-face/neck region from the camera view, or when the swallowing event could not be identified reliably from the video. These issues were observed particularly in cases of limited subject compliance or substantial involuntary motion. Because the original quality-control process did not retain a structured per-record exclusion code, exact exclusion counts by reason could not be reconstructed retrospectively.

After quality review, recordings were organized into aligned three-condition samples. Each sample consisted of one valid RGB-D swallowing clip from IDDSI Level 0, one from IDDSI Level 2, and one from IDDSI Level 4. To maintain a consistent model input structure, only samples with all three IDDSI recordings available were retained for analysis.

The final analyzed dataset comprised 282 aligned three-condition samples from 98 participants, including 57 healthy participants and 41 participants with dysphagia. Of the retained samples, 164 were from healthy controls and 118 were from participants with dysphagia. Since each retained sample contained one clip for each IDDSI condition, the dataset included 282 clips at IDDSI Level 0, 282 clips at IDDSI Level 2, and 282 clips at IDDSI Level 4, for a total of 846 condition-specific RGB-D clips.

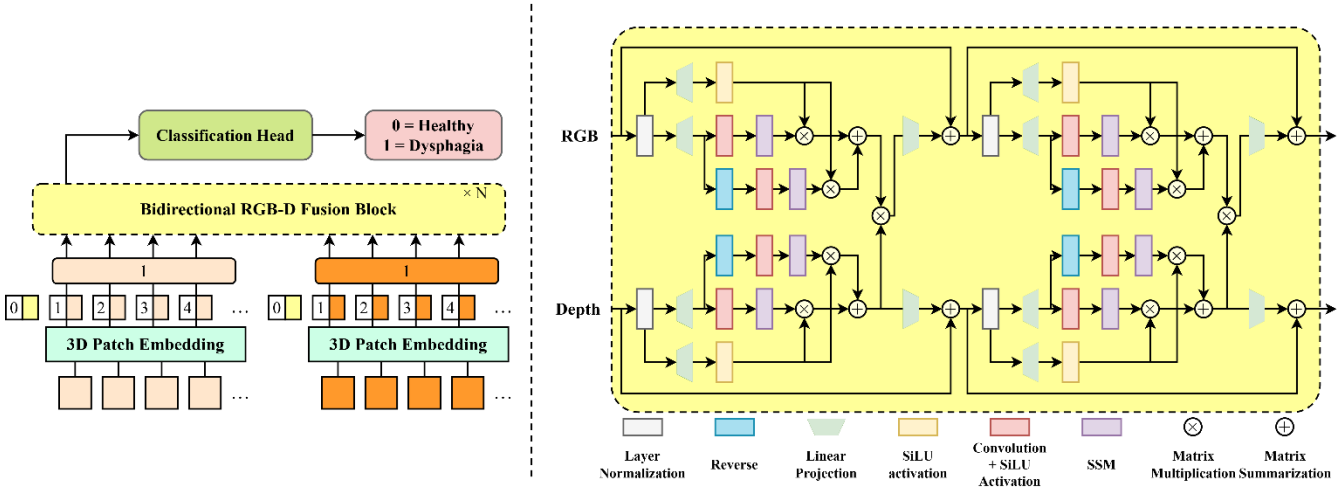
Preprocessing differed by modality. RGB frames were retained without depth filtering. Depth frames were processed with the Intel RealSense SDK by conversion to disparity, edge-preserving spatial filtering, temporal filtering, and conversion back to depth. The acquisition targeted the lower

face and neck, including the lips, mandible, and throat. During model input preparation, stored frames were used without additional spatial cropping and resized to 224 × 224 pixels using bilinear interpolation with anti-aliasing for RGB and nearest-neighbor interpolation without anti-aliasing for depth. RGB and depth were separately min–max normalized to [0,1] for each frame. No random data augmentation was applied. Resizing used fixed settings, and normalization used only the individual frame; no dataset-level normalization statistics were fitted or shared across folds.

$$S^{\triangle} = Softplus(Linear_{\triangle}(x(t) + bias_{\triangle}) \quad (9)$$

E. Model

Fig. 2 presents MMVX-Net, a dual-stream architecture for concurrent processing and fusion of RGB and depth clips. Each modality is processed by a bidirectional Mamba block with forward and reverse convolution/SSM pathways. Cross-modal exchange is applied repeatedly while modality-specific streams are retained.



a) MMVX-Net b) Bidirectional RGB-D Fusion Block

Fig. 1. Fig. 1. (a)The model architecture of MMVX-Net and the structure of (b) bidirectional multispectral fusion block

For each selected clip, 16 frames were sampled uniformly across the clip. The sampled frames from IDDSI Levels 0, 2, and 4 were combined to form one 48-frame model input. These inputs combine three separate clips rather than one continuous recording. Because the model input length was fixed at 48 frames, the present experiment does not evaluate long-sequence scaling.

D. State Space Models and Mamba

Mamba is a sequence-modeling architecture based on state-space models (SSMs) and a selective scan mechanism. Its input-dependent parameters enable content-aware state updates with linear time complexity O(n) in sequence length n [12]–[14]. The continuous-time linear SSM can be written as follows:

$$h'(t) = Ah(t) + Bx(t) \quad (1)$$
$$y(t) = Ch(t) \quad (2)$$

where A is the state-transition matrix, B maps the input to the latent state, and C maps the latent state to the output. For discrete sequences, zero-order hold discretization gives:

$$A = exp(\triangle A) \quad (3)$$
$$B = (\triangle A)^{-1}(exp(\triangle A) - I) \cdot \triangle B \quad (4)$$
$$h_k = \overline{A}h_{k-1} + \overline{B}x_k \quad (5)$$
$$y_k = Ch_k \quad (6)$$

where $\bar{A}$ and $\bar{B}$ are the discrete counterparts of A and B and $\triangle$ is the step size. Mamba replaces the fixed B, C, and $\triangle$ terms with input-dependent projections before discretization. Specifically,

$$S^B = Linear_B(x(t)) \quad (7)$$
$$S^C = Linear_C(x(t)) \quad (8)$$

The RGB-D input is divided into separate RGB and depth streams, each embedded with a 3D convolution. The resulting spatiotemporal patches are flattened and projected to token embeddings, and positional embeddings encode their locations. We use the spatial-first bidirectional scan shown in Fig. 3 so that spatially adjacent tokens are traversed coherently in both directions.

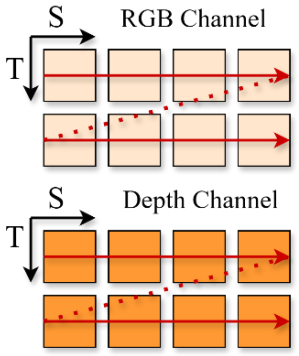


Fig. 2. Fig. 3. Spatial-first scanning method

In the core module, parallel bidirectional Mamba blocks process the RGB and depth streams. For each modality, the input Hin is projected along two paths:

$$H_a = Linear_a(H_{in}) \quad (10)$$
$$H_b = Linear_b(H_{in}) \quad (11)$$

The second projection is passed through a sigmoid linear unit (SiLU):

$$H_{SiLU} = SiLU(H_b) \quad (12)$$

where SiLU(x) = x • σ (x). The other path contains forward and reverse convolution/SSM branches. The reverse branch

TABLE I. MODEL COMPARISON WITH 5-FOLD CROSS-VALIDATION

Model	FLOPs (G)	Parameters (M)		Acc	Pc	Rc	F1	ROC-AUC	PR-AUC
X3D [20]	15.31	2.98	Avg.	0.894	0.901	0.886	0.888	0.936	0.938
			Std.	0.011	0.021	0.014	0.010	0.033	0.027
ViViT [21]	18.73	3.75	Avg.	0.915	0.916	0.905	0.909	0.942	0.920
			Std.	0.035	0.035	0.041	0.038	0.027	0.030
Swin-T [22]	106.56	25.71	Avg.	0.851	0.873	0.844	0.841	0.893	0.861
			Std.	0.081	0.068	0.087	0.091	0.039	0.064
VideoMamba [14]	3.70	0.39	Avg.	0.915	0.913	0.918	0.912	0.952	0.948
			Std.	0.024	0.024	0.020	0.024	0.019	0.028
MMVX-Net (ours)	6.47	0.69	Avg.	0.929	0.927	0.928	0.927	0.963	0.955
			Std.	0.029	0.028	0.032	0.030	0.010	0.010

Pc: Precision; Rc: Recall; F1: F1-Score; ROC-AUC: Area under receiver-operating curve; PR-AUC: area under precision-recall curve.

operates on the reversed token order. Eq. (16) therefore takes the backward convolution output Cbwd as its input.

$$C_{fwd} = Conv_{fwd}(H_a) \quad (13)$$

$$S_{fwd} = SSM_{fwd}(C_{fwd}) \quad (14)$$

$$C_{bwd} = Conv_{bwd}(Reverse(H_a)) \quad (15)$$

$$S_{bwd} = SSM_{bwd}(C_{bwd}) \quad (16)$$

The forward and backward state-space outputs are gated by the SiLU path and summed:

$$G_{fwd} = S_{fwd} \otimes H_{SiLU} \quad (17)$$

$$G_{bwd} = S_{bwd} \otimes H_{SiLU} \quad (18)$$

$$G_{combine} = G_{fwd} + G_{bwd} \quad (19)$$

Cross-modal fusion multiplies the combined output of one modality into the other stream. In the first bidirectional pair, depth features modulate the RGB stream; the direction is reversed in the second pair. After the fusion blocks, a linear projection produces the binary healthy/dysphagia clip-level prediction.

F. Evaluation

Evaluation uses five participant-disjoint outer fold with all repetitions and IDDSI conditions from each participant assigned to the same outer fold. Dysphagia is designated class 1. Precision, recall, and F1-score are macro-averaged across both classes. Performance will be reported as the mean ± standard deviation of the five outer-fold metrics. A pooled confusion matrix will summarize out-of-fold predictions at the combined-sample level, with each sample counted once.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (20)$$

$$Precision = \frac{TP}{TP + FP} \quad (21)$$

$$Recall = \frac{TP}{TP + FN} \quad (22)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (23)$$

ROC-AUC and PR-AUC were calculated with roc_auc_score and average_precision_score from scikit-learn [19]. Fold-level metrics are reported as mean ± standard deviation. We compared MMVX-Net with X3D [20], ViViT [21], Swin-T [22], and VideoMamba [14]. All baseline runs will use the same saved participant-disjoint folds and shared evaluation approaches. Each baseline will receive four-channel RGB-D inputs comprising 48 frames resized to 224 × 224 pixels, with identical modality-specific normalization and no random data augmentation. Models will be trained from random initialization without pretrained weights, using Adam with a learning rate of 1×10^{-4}, zero weight decay, no learning-rate scheduler, batch size 1, and 50 epochs per fold. Checkpoints will be selected by minimum cross-entropy on a separate participant-disjoint inner validation set. One fixed configuration per model will be evaluated without hyperparameter search.

III. RESULTS

MMVX-Net was evaluated with five-fold cross-validation. Table I reports the mean and standard deviation across validation folds. The consistent metric set is accuracy 0.929 ± 0.029, precision 0.927 ± 0.028, recall 0.928 ± 0.032, and F1-score 0.927 ± 0.030. These values agree with the pooled clip-level confusion matrix in Fig. 4, in which 262 of 282 clips were classified correctly (accuracy 0.929).

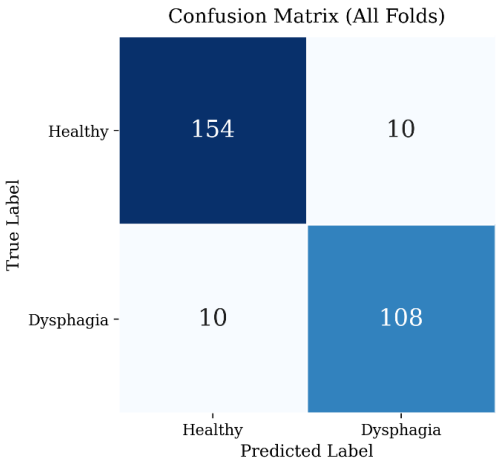


Fig. 3. Fig. 4. Pooled sample-level confusion matrix across the five validation folds

The pooled confusion matrix in Fig. 4 contains 154 correctly classified healthy clips and 108 correctly classified dysphagia clips, with 10 errors in each direction. This is a clip-level summary of predictions across the five held-out folds; it should not be interpreted as participant-level sensitivity or specificity unless participant-level aggregation is added.

The mean ROC-AUC was 0.963 ± 0.010 and the mean PR-AUC was 0.955 ± 0.010. Per-fold ROC and precision-recall curves are shown in Fig. 5. These results indicate strong discrimination within the evaluated cross-validation protocol, subject to confirmation of participant-disjoint folds.

MMVX-Net required 6.47 GFLOPs and 0.69 million parameters. This was lower than X3D (15.31 GFLOPs, 2.98 million parameters), ViViT (18.73 GFLOPs, 3.75 million parameters), and Swin-T (106.56 GFLOPs, 25.71 million parameters), but higher than VideoMamba (3.70 GFLOPs, 0.39 million parameters). These measurements describe model complexity for the evaluated input; hardware latency and memory use were not measured.

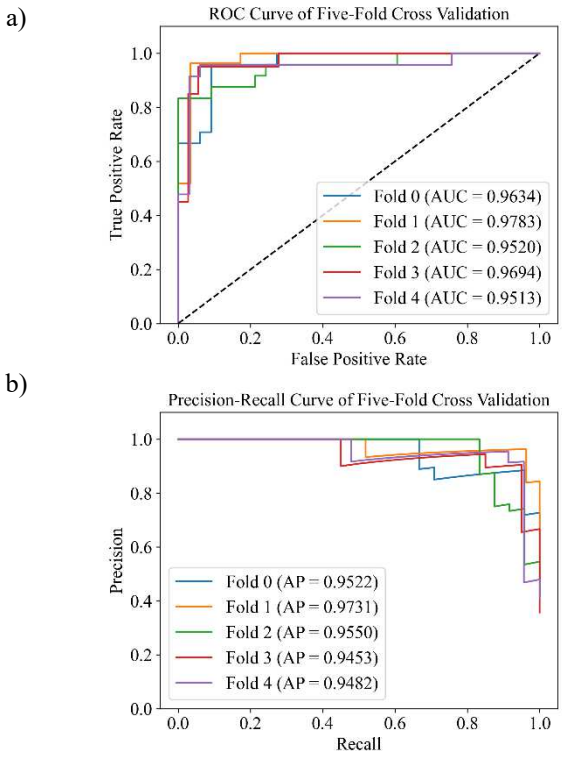


Fig. 4. Fig. 5. (a) ROC curves and (b) precision-recall curves for the five validation folds of MMVX-Net.

IV. Discussion

This study evaluates a dual-stream Mamba architecture for classifying pre-segmented RGB-D swallowing clips. The dataset contains synchronized video from 68 healthy controls and 68 participants with clinically diagnosed dysphagia performing controlled swallowing tasks across three bolus consistencies. The contribution is the combination of modality-specific state-space processing and repeated bidirectional RGB-depth exchange, evaluated with an internal five-fold protocol. The study does not establish end-to-end screening in continuous video, direct measurement of internal swallowing biomechanics, or clinical deployment performance.

MMVX-Net preserves separate RGB and depth streams while exchanging information between them. This design is intended to retain modality-specific surface-motion cues and model cross-modal relationships. However, the current comparison does not include RGB-only, depth-only, early-fusion, late-fusion, unidirectional, or no-cross-exchange ablations. The contribution of each component therefore remains to be isolated experimentally, and statements attributing the observed gain specifically to the fusion mechanism should be interpreted as architectural motivation rather than causal evidence.

Within the reported five-fold evaluation, MMVX-Net achieved accuracy of 0.929 ± 0.029, precision of 0.927 ± 0.028, recall of 0.928 ± 0.032, F1-score of 0.927 ± 0.030, ROC-AUC of 0.963 ± 0.010, and PR-AUC of 0.955 ± 0.010. Its average accuracy was 0.014 higher than VideoMamba, while the fold standard deviations overlapped. Without paired testing on identical participant folds and repeated seeds, this difference should be treated as descriptive rather than evidence of statistical superiority. MMVX-Net is compact relative to X3D, ViViT, and Swin-T, although VideoMamba is smaller and requires fewer FLOPs.

Several limitations constrain interpretation. First, the evaluation is conditional on obtaining usable recordings during the instructed swallowing protocol; excluding recordings affected by limited cooperation or poor visibility may limit applicability to participants who experience these difficulties. Second, the final dataset flow and group-specific demographic and clinical characteristics are not yet fully reported, limiting assessment of selection effects and confounding. Third, labels reflect participant-level clinical status rather than swallow-level VFSS/FEES-confirmed physiology; models may therefore exploit age, morphology, posture, assistance, background, or session cues. Fourth, clips were manually selected to contain visible swallowing actions in a fixed setup, so the study evaluates curated clip classification rather than automated detection in continuous, unconstrained video. Fifth, only 48 sampled frames were tested, so no empirical conclusion can be drawn about long-sequence scaling. Finally, MMVX-Net lacks a tailored explanation analysis, whereas recent dysphagia AI research has emphasized interpretability and external validation [23].

Future work should prioritize participant-disjoint and external validation, transparent dataset accounting, group-specific demographic analysis, and paired evaluation on identical folds. Modality and fusion ablations, temporally shuffled and pre-swallow controls, background/neck masking, IDDSI-stratified performance, calibration, sensitivity and specificity, and sequence-length/latency measurements would clarify what information drives predictions and where the method generalizes. An end-to-end system would additionally require automatic swallow-event localization and robustness to occlusion, head motion, camera placement, and lighting variation. Model explanations should be evaluated for clinical plausibility rather than presented only as qualitative visualizations.

V. Conclusion

We proposed MMVX-Net, a dual-stream bidirectional Mamba architecture for binary classification of pre-segmented RGB-D swallowing clips. In the reported five-fold evaluation, the model achieved accuracy of 0.929, F1-score of 0.927, ROC-AUC of 0.963, and PR-AUC of 0.955 with 6.47 GFLOPs and 0.69 million parameters. The results support the feasibility of efficient multispectral clip classification in the controlled acquisition setting. Confirmation of participant-disjoint folds, complete reporting of dataset composition and baseline settings, ablation studies, and external validation are required before broader claims about end-to-end screening or clinical deployment.

REFERENCES

[1] M. Ribeiro et al., "The Prevalence of Oropharyngeal Dysphagia in Adults: A Systematic Review and Meta-analysis," Dysphagia, pp. 1-14, 2023.

[2] F. Rajati, N. Ahmadi, Z. A. Naghibzadeh, and M. Kazeminia, "The global prevalence of oropharyngeal dysphagia in different populations: a systematic review and meta-analysis," Journal of translational medicine, vol. 20, no. 1, p. 175, 2022, doi: 10.1186/s12967-022-03380-0.

[3] E. V. Sreedevi et al., "Artificial intelligence tools for the assessment and management of dysphagia: a scoping review," BMJ Open, vol. 16, no. 7, p. e119582, 2026, doi: 10.1136/bmjopen-2026-119582.

[4] D. K.-H. Lai et al., "Computer-aided screening of aspiration risks in dysphagia with wearable technology: a Systematic Review and meta-analysis on test accuracy," Frontiers in Bioengineering and Biotechnology, vol. 11, p. 1205009, 2023, doi: 10.3389/fbioe.2023.1205009.

[5] B. P.-H. So et al., "Swallow Detection with Acoustics and Accelerometric-Based Wearable Technology: A Scoping Review," International Journal of Environmental Research and Public Health, vol. 20, no. 1, p. 170, 2022, doi: 10.3390/ijerph20010170.

[6] D. W.-C. Wong et al., "Current Technological Advances in Dysphagia Screening: Systematic Scoping Review," Journal of Medical Internet Research, vol. 27, p. e65551, 2025.

[7] T. Liu, J. Kong, and M. Jiang, "RGB-D Action Recognition Using Multimodal Correlative Representation Learning Model," IEEE Sensors Journal, vol. 19, no. 5, pp. 1862-1872, 2019, doi: 10.1109/JSEN.2018.2884443.

[8] M. M. S. Ahmed *et al.*, "Three-dimensional evaluation of upper pharyngeal airway, hyoid bone, and craniocervical changes following stabilization splint therapy in adult patients with temporomandibular joint disorders and mandibular deviation: A retrospective study," *Journal of Stomatology, Oral and Maxillofacial Surgery,* vol. 124, no. 6, p. 101646, 2023, doi: https://doi.org/10.1016/j.jormas.2023.101646.

[9] W.-J. Lee et al., "Real-Time Tracking of Laryngeal Motion via the Surface Depth-Sensing Technique for Radiotherapy in Laryngeal Cancer Patients," Bioengineering, vol. 10, no. 8, p. 908, 2023, doi: 10.3390/bioengineering10080908.

[10] D. K.-H. Lai et al., "Transformer models and convolutional networks with different activation functions for swallow classification using depth video data," Mathematics, vol. 11, no. 14, p. 3081, 2023, doi: 10.3390/math11143081.

[11] S. Nerella et al., "Transformers in Healthcare: A Survey," 2023, doi: 10.48550/arXiv.2307.00067.

[12] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," 2023, doi: 10.48550/arXiv.2312.00752.

[13] M. M. Rahman et al., "Mamba in Vision: A Comprehensive Survey of Techniques and Applications," 2024, doi: 10.48550/arXiv.2410.03105.

[14] K. Li et al., "VideoMamba: State Space Model for Efficient Video Understanding," 2024, doi: 10.48550/arXiv.2403.06977.

[15] S. Li, H. Singh, and A. Grover, "Mamba-nd: Selective state space modeling for multi-dimensional data," in European Conference on Computer Vision, Milano, Italy, 2024, pp. 75-92.

[16] Y. Zhan, Z. Zeng, H. Liu, X. Tan, and Y. Tian, "MambaSOD: Dual Mamba-driven cross-modal fusion network for RGB-D Salient Object Detection," Neurocomputing, vol. 631, p. 129718, 2025/05/28/ 2025, doi: 10.1016/j.neucom.2025.129718.

[17] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel RealSense Stereoscopic Depth Cameras," 2017, doi: 10.48550/arXiv.1705.05548.

[18] H.-J. Lim et al., "A comprehensive assessment protocol for swallowing (CAPS): Paving the way towards computer-aided dysphagia screening," International Journal of Environmental Research and Public Health, vol. 20, no. 4, p. 2998, 2023.

[19] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, Nov. 2011.

[20] C. Feichtenhofer, "X3D: Expanding Architectures for Efficient Video Recognition," 2020, doi: 10.48550/arXiv.2004.04730.

[21] A. Arnab et al., "ViViT: A Video Vision Transformer," 2021, doi: 10.48550/arXiv.2103.15691.

[22] Z. Liu et al., "Video Swin Transformer," 2021, doi: 10.48550/arXiv.2106.13230.

[23] Y. Dai et al., "Interpretable machine learning for accessible dysphagia screening and staging in older adults," iScience, vol. 29, no. 1, p. 114451, 2026, doi: 10.1016/j.isci.2025.114451.