Computer Modeling in Engineering & Sciences

Tech Science Press

Doi:10.32604/cmes.2025.075064

ARTICLE

Real-Time Mouth State Detection Based on a BiGRU-CLPSO Hybrid Model with Facial Landmark Detection for Healthcare Monitoring Applications

Mong-Fong Horng[1,#], Thanh-Lam Nguyen[1,#], Thanh-Tuan Nguyen[2,*], Chin-Shiuh Shieh[1,*], Lan-Yuen Guo[3], Chen-Fu Hung[4] and Chun-Chih Lo[1]

[1]Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, 807618, Taiwan
[2]Faculty of Electrical and Electronics Engineering, Nha Trang University, Nha Trang, 650000, Vietnam
[3]Department of Sports Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung, 807378, Taiwan
[4]Program in Biomedical Engineering, College of Medicine, Kaohsiung Medical University, Kaohsiung, 807378, Taiwan

*Corresponding Authors: Thanh-Tuan Nguyen. Email: tuannt@ntu.edu.vn; Chin-Shiuh Shieh. Email: csshieh@nkust.edu.tw

[#]These authors contributed equally as co-first authors

Received: 24 October 2025; Accepted: 19 December 2025; Published: 29 January 2026

ABSTRACT: The global population is rapidly expanding, driving an increasing demand for intelligent healthcare systems. Artificial intelligence (AI) applications in remote patient monitoring and diagnosis have achieved remarkable progress and are emerging as a major development trend. Among these applications, mouth motion tracking and mouth-state detection represent an important direction, providing valuable support for diagnosing neuromuscular disorders such as dysphagia, Bell's palsy, and Parkinson's disease. In this study, we focus on developing a real-time system capable of monitoring and detecting mouth state that can be efficiently deployed on edge devices. The proposed system integrates the Facial Landmark Detection technique with an optimized model combining a Bidirectional Gated Recurrent Unit (BiGRU) and Comprehensive Learning Particle Swarm Optimization (CLPSO). We conducted a comprehensive comparison and evaluation of the proposed model against several traditional models using multiple performance metrics, including accuracy, precision, recall, F1-score, cosine similarity, ROC–AUC, and the precision–recall curve. The proposed method achieved an impressive accuracy of 96.57% with an excellent precision of 98.25% on our self-collected dataset, outperforming traditional models and related works in the same field. These findings highlight the potential of the proposed approach for implementation in real-time patient monitoring systems, contributing to improved diagnostic accuracy and supporting healthcare professionals in patient treatment and care.

KEYWORDS: Remote patient monitoring; mouth state detection; dysphagia; facial landmark detection; bidirectional gated recurrent unit; comprehensive learning particle swarm optimization

1 Introduction

Swallowing disorders and neuromuscular dysfunctions in the orofacial region, such as facial paralysis, Parkinson's disease, or cranial nerve injury, significantly impair patients' ability to eat and communicate, thereby reducing their quality of life and increasing the risk of respiratory complications. Several studies have reported that the prevalence of dysphagia in the elderly ranges from 13% to 38% [1] and can reach up to 78% in stroke patients [2]. This highlights the urgent need for effective monitoring of oral motor functions in rehabilitation. Continuous observation of mouth movements not only enables clinicians to objectively assess neuromuscular conditions but also facilitates timely adjustments in treatment plans. In particular, real-time and remote monitoring systems have opened new opportunities in telemedicine, allowing patients to be

supervised and guided directly at home. Such systems can reduce follow-up visits, alleviate the workload on healthcare facilities, and extend medical services to regions with limited professional resources [3,4]. However, there remains a lack of lightweight and reliable solutions that can operate stably on common edge devices, such as IP cameras or wearable sensors, which are typically constrained in energy consumption and computational resources [5].

The real-time mouth-state detection problem faces multiple technical challenges. On edge platforms, CPU power, memory, and bandwidth are limited, whereas the system demands low latency and high stability for immediate response. Deep learning models with heavy architectures, such as multi-layer CNNs, though capable of achieving high accuracy, often fail to maintain consistent performance and reasonable energy efficiency under edge-deployment conditions. Conversely, handcrafted feature-based methods, such as those using the Mouth Aspect Ratio (MAR) with fixed thresholds, are lightweight and easy to implement but lack adaptability. They tend to degrade in sensitivity or produce high false-positive rates under variations in individual anatomy and environmental conditions, such as lighting, head pose, or occlusion [6]. Here, the MAR threshold can be understood as a relative ratio between the vertical and horizontal dimensions of the mouth. When the mouth opens, the vertical distance increases while the horizontal distance tends to decrease. However, fixed-threshold MAR systems are difficult to generalize in practice, as they cannot dynamically adapt across groups of patients differing in gender, age, or ethnicity.

To overcome these limitations, this study proposes an optimized pipeline composed of several coordinated components designed to balance accuracy, sensitivity, and processing speed. First, mouth landmarks are extracted from a lightweight facial-landmark detector, yielding a total of 68 spatial coordinates per video frame. The 2D spatial coordinates of these landmarks are then used as input features for deep learning models to classify mouth states as open or closed. The learning models automatically extract key intermediate features, such as distances and geometric ratios between landmarks, without relying on manually defined parameters. In this framework, a Bidirectional Gated Recurrent Unit (Bi-GRU) serves as the primary model for processing sequential data and identifying mouth states [7]. Finally, the Comprehensive Learning Particle Swarm Optimization (CLPSO) algorithm is employed to perform hyperparameter tuning for the Bi-GRU network. This optimization strategy aims to achieve an optimal configuration within resource constraints, reducing inference latency while improving sensitivity and accuracy compared with baseline approaches [8].

Compared with previous studies, this approach differs by introducing a landmark-coordinate-based mouth-state detection method that replaces MAR thresholds while maintaining a lightweight model architecture suitable for edge deployment. The main contributions of this study are summarized as follows:

(i) Proposing the use of 2D spatial coordinates of facial landmarks as model input for mouth-state recognition, replacing MAR thresholds that are often sensitive to inter-subject variations.
(ii) Introducing a lightweight hybrid BiGRU-CLPSO architecture for mouth-state detection from sequential landmark data.
(iii) Successfully deploying the proposed system on common edge devices with low latency, suitable for home-based monitoring scenarios using only CPU and standard cameras.
(iv) Conducting cross-validation on our self-collected dataset to evaluate model robustness and compare with multiple machine learning models such as MLP, CNN, LSTM, and XGBoost. Experimental results show that landmark-based geometric models achieve high accuracy and sensitivity with stable real-time inference. Furthermore, the BiGRU-CLPSO model demonstrates superior stability and overall performance compared with other methods under real-world operating conditions.

The remainder of this paper is organized as follows: Section 2 reviews the related works; Section 3 describes the proposed BiGRU-CLPSO method and feature design; Section 4 presents the experiments, multi-source evaluation, and discussion; and Section 5 concludes the study and outlines future research directions.

2 Related Work

2.1 Clinical Applications of Mouth-State Monitoring

Monitoring the open–close states of the mouth holds significant clinical value across multiple medical domains. In the management of dysphagia, detecting abnormal oral movements or prolonged mouth opening can serve as an early indicator of swallowing difficulty or aspiration risk, allowing timely intervention to prevent complications such as aspiration pneumonia [9]. Continuous monitoring also enables the detection of subtle deviations from normal swallowing patterns, which are often overlooked during brief or periodic clinical assessments [4]. In fatigue detection, particularly in driver monitoring and neurological evaluation, the frequency of yawning or prolonged mouth opening can provide a reliable indicator of drowsiness or reduced alertness, thereby supporting timely safety warnings [10,11]. Another important application is rehabilitation compliance monitoring: patients recovering from facial nerve paralysis, stroke, or temporomandibular joint disorders can be remotely supervised to ensure correct execution and frequency of prescribed oral exercises [12,13]. Moreover, assistive communication systems for individuals with severe motor impairments can employ mouth-state recognition as a control signal, enhancing accessibility to assistive technologies [14].

Despite these promising applications, current "gold-standard" clinical tools, such as video fluoroscopic swallowing study (VFSS) and fiberoptic endoscopic evaluation of swallowing (FEES), remain limited by their discontinuous nature, high cost, and dependence on specialized equipment and trained personnel [15,16]. These constraints highlight the need for affordable, continuous, and remote patient monitoring (RPM) solutions that are compatible with commonly available consumer cameras, thereby extending care beyond hospital settings, particularly for patients in remote areas or with mobility limitations.

However, deploying mouth-state monitoring systems outside laboratory conditions faces several constraints. Target platforms are typically resource-constrained edge devices, with restricted memory, computational power, and battery capacity, making the deployment of heavy neural models challenging [17,18]. Environmental variability, including changes in lighting, head posture, or partial occlusion by masks or hands, further challenges system robustness [6]. In addition, privacy requirements demand minimizing the storage and processing of raw facial images, while model interpretability remains critical in medical contexts where clinicians must understand and trust algorithmic decisions [19,20].

Regarding evaluation metrics, medical applications typically prioritize f1-score and recall to ensure high sensitivity in detecting clinically relevant events, while also considering median inference latency, model size, and cross-subject or cross-session stability as indicators of overall system reliability [21]. To meet these demands, recent studies have increasingly focused on developing lightweight, stable, and interpretable processing pipelines capable of being deployed on edge devices and operating reliably in real-world, non-laboratory environments [20].

2.2 Deep Learning Approaches for Mouth-State Detection

In recent years, there has been a rapid growth in the use of deep learning techniques for mouth-state recognition, with various approaches leveraging the power of spatiotemporal feature extraction models [22,23]. Two-dimensional convolutional neural networks (2D-CNNs) focus on learning static

shape features from still images or individual video frames, often adopting lightweight architectures such as MobileNet or EfficientNet to enhance deployability on mobile or embedded devices [24,25]. Three-dimensional convolutional neural networks (3D-CNNs) extend this capability into the temporal domain, extracting both spatial and motion features from frame sequences. Architectures such as C3D, I3D, and S3D have achieved high performance in classifying mouth-related behaviors, such as swallowing, chewing, or speaking, but they demand substantial computational resources [26,27].

Recurrent neural networks (RNNs) and their variants, LSTM and GRU, have been widely employed to model sequential temporal features, enabling the detection of dynamic mouth patterns such as open–close cycles or repetitive swallowing phases [28]. More recently, Transformer-based architectures, including lightweight variants such as Video Swin Transformer and TimeSformer, have been applied to video data. By leveraging the self-attention mechanism, they can learn long-range temporal dependencies between frames and improve robustness under noisy or varying environmental conditions [27]. Notably, in the domain of facial analysis, novel architectures such as the Reference Heatmap Transformer have set new benchmarks for precise landmark localization [29]. However, most of these models still require high-end hardware and large memory capacity to make them unsuitable for edge deployment on resource-limited hardware platforms. To address this limitation, several lightweight approaches have emerged, combining facial landmark features with compact recurrent networks to reduce input dimensionality and computational cost. These studies have shown that when geometric features are well designed, lightweight sequential models can achieve performance close to that of heavier end-to-end architectures [22]. Nevertheless, feature selection and hyperparameter tuning in these models are often based on manual or empirical heuristics, leading to suboptimal results and limited adaptability to multi-source or cross-subject datasets [30,31].

In this study, we adopt the Bidirectional Gated Recurrent Unit (Bi-GRU) as the core component. Bi-GRU retains the lightweight structure of GRU with a bidirectional mechanism, enabling the model to capture both past and future contextual information at each step. Specifically, the sequence is encoded by two parallel GRU branches named as forward and backward, and their hidden states are concatenated or combined to form context-rich representations. This design is particularly effective for distinguishing short transitions, such as quick swallows, and lip compressions. With landmark coordinates as input, Bi-GRU can exploit inter-frame dynamics more effectively than static geometric descriptors [30].

To overcome the limitations of manual hyperparameter tuning, we integrate Comprehensive Learning Particle Swarm Optimization (CLPSO), an enhanced swarm intelligence algorithm, to optimize the Bi-GRU hyperparameters and simultaneously perform feature selection for geometric descriptors. CLPSO is particularly well-suited for hybrid search spaces, combining binary dimensions for feature selection with continuous or integer dimensions for network parameters. It balances exploration and exploitation, avoids premature convergence, and supports multi-objective optimization goals: maximizing f1-score and recall, while minimizing median latency and model size [32].

The proposed integration of Bi-GRU and CLPSO with spatial landmark coordinates has been initially explored in prior researches, especially within the context of simultaneous optimization of lightweight sequential models and geometric-feature pipelines for edge devices. This study, therefore, addresses this research gap by developing a solution that is computationally efficient for real-time edge deployment to maintain the accuracy and the sensitivity required for clinical monitoring applications.

3 Real-Time Mouth State Detection Based on a BiGRU-CLPSO Hybrid Model with Facial Landmark

In this study, we develop a patient monitoring system designed to track changes in mouth movements, as illustrated in Fig. 1.

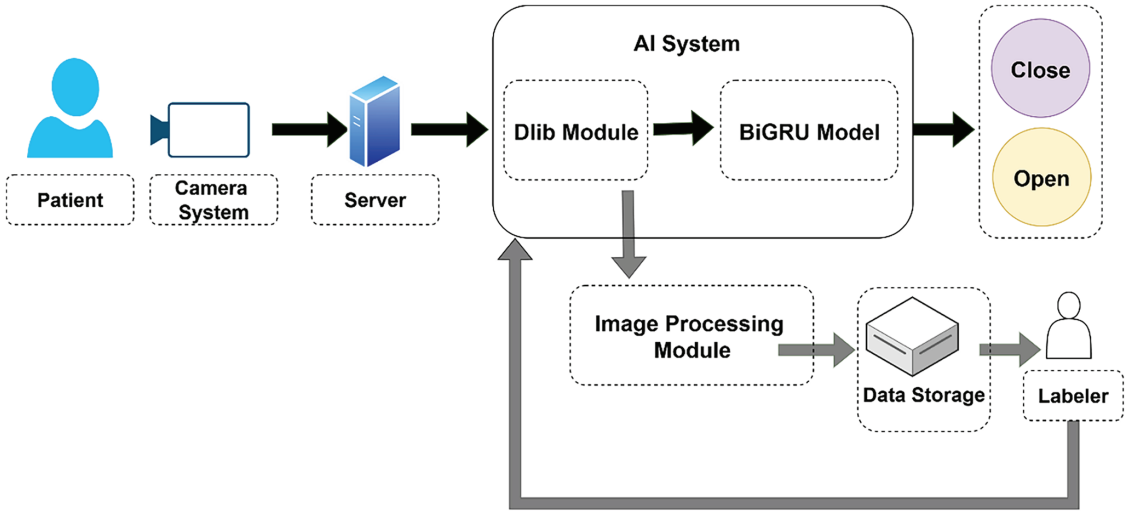

Figure 1: Overall architecture of the proposed system

In this system, a camera module is installed to capture and transmit patient images to a central server system. The AI module processes the incoming video by extracting individual frames and detecting facial landmarks with the Dlib library. The spatial coordinates of the detected landmarks are then fed into a BiGRU model to determine the mouth state, open or closed. In addition, the system includes a mechanism to store video frames along with the corresponding landmark information. These images are forwarded to human annotators for manual labeling of mouth states. This process helps expand and refine the mouth-state dataset of patients, enabling further training to improve model accuracy over time.

As mentioned earlier, this study employs the BiGRU model in combination with the heuristic optimization algorithm CLPSO as the primary approach for mouth-state classification. Furthermore, we compare the proposed method with several other representative deep learning algorithms, including CNN, MLP, LSTM, and XGBoost. Multiple algorithms are evaluated to highlight the superior performance of deep learning models trained on facial landmark coordinates, as opposed to traditional approaches that rely on fixed MAR thresholds. Thehe deep learning models utilized in this study are introduced later.

3.1 Facial Landmark Detection Using Dlib

In the overall pipeline, the Dlib module performs two key functions: face detection using the Histogram of Oriented Gradients (HOG)+ SVM method and landmark localization using Ensemble of Regression Trees (ERT). Dlib is an open-source C++ library that integrates multiple optimized modules for computer vision and machine learning applications. In this study, we utilize built-in face detection and facial landmark localization capabilities of Dlib library [23]. The detection mechanism combines HOG and Support Vector Machine (SVM) classifiers [5]. Once a face bounding box is detected, the Ensemble of Regression Trees (ERT) module predicts the facial landmark points [24]. The landmark predictor used in this study was trained on the 300 Faces In-the-Wild Challenge (300-W) dataset [25–27]. Fig. 2 illustrates the standard set of 68 facial landmark points identified by the algorithm.

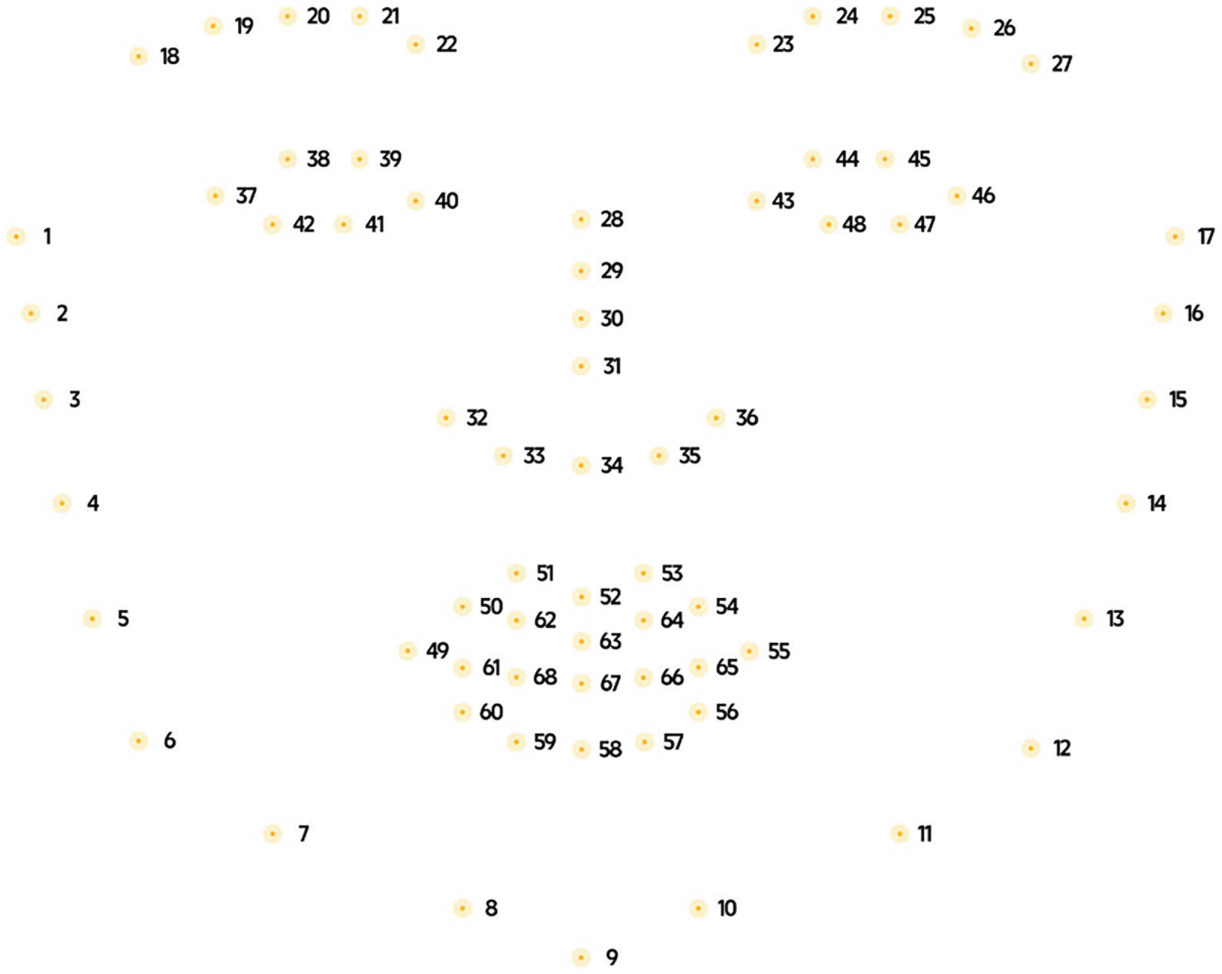

Figure 2: Position the 68 facial landmarks detected by Dlib on the face

In the grayscale image, the gradient components are approximated as:

$$g_x = I(x+1, y) - I(x-1, y) \tag{1}$$

$$g_y = I(x, y+1) - I(x, y-1) \tag{2}$$

where $I(x, y)$ denotes the intensity at the coordinate (x, y), and g_x, g_y represent the horizontal and vertical gradient components, respectively. The gradient magnitude and orientation can then be derived as:

$$m(x, y) = \sqrt{g_x^2 + g_y^2}, \quad \theta(x, y) = \operatorname{atan2}(g_y, g_x) \tag{3}$$

where m measures the magnitude of intensity variation, and θ indicates the dominant direction of change. Next, the image is divided into cells of size 8×8. For each cell $\mathcal{C}$, an orientation histogram with B bin is computed as the weighted sum of magnitudes:

$$h_b = \sum_{(x,y)\in\mathcal{C}} m(x, y), \mathbf{1}\{\theta(x, y) \in \mathcal{B}_b\}, \quad b = 1, \ldots, B \tag{4}$$

where $\mathcal{B}_b$ is the angle range of the bin b, and $\mathbf{1}$ is an indicator function. Several neighboring cells are grouped into a block of size 2×2 for L2-Hys normalization, which enhances robustness to illumination variations:

$$\tilde{\mathbf{v}} = \frac{\mathbf{v}}{\sqrt{|\mathbf{v}|_2^2 + \varepsilon^2}}, \quad \hat{\mathbf{v}} = \frac{\min(\tilde{\mathbf{v}}, \tau)}{\sqrt{|\min(\tilde{\mathbf{v}}, \tau)|_2^2 + \varepsilon^2}} \tag{5}$$

where $\mathbf{v}$ denotes the block feature vector, $\varepsilon > 0$ is a stabilization constant, and α is a clipping threshold. By aggregating all blocks within a multi-scale sliding window, a composite HOG descriptor $\mathbf{h} \in \mathbb{R}^d$. A linear SVM classifier then performs binary classification as follows:

$$f(\mathbf{h}) = \mathbf{w}^\top \mathbf{h} + b, \quad \hat{y} = \text{sign!}(f(\mathbf{h})) \tag{6}$$

where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ are the trained parameters. When a soft decision boundary is required for noise reduction, Platt scaling is applied to calibrate the margin score:

$$p = \sigma(a, f(\mathbf{h}) + c) = \frac{1}{1 + \exp(-(a, f(\mathbf{h}) + c))} \tag{7}$$

where $(a, c) \in \mathbb{R}^2$ are scaling coefficients, σ is a sigmoid function. This HOG + SVM architecture is particularly effective for objects with complex shapes, such as human faces, while maintaining real-time performance owing to compact features and a linear decision model [5]. Once the face region of interest (ROI) is detected, Dlib employs an Ensemble of Regression Trees (ERT) to infer the facial shape comprising LLL landmark points:

$$\mathbf{s} = [x_1, y_1, \ldots, x_L, y_L]^\top \in \mathbb{R}^{2L} \tag{8}$$

where (x_ℓ, y_ℓ) denotes the coordinate of the ℓ landmark, and in Dlib's default configuration $L = 68$. ERT refines the predicted landmark positions iteratively through cascaded regression stages:

$$\mathbf{s}^{(t+1)} = \mathbf{s}^{(t)} + \mathcal{R}_t\left(\Phi\left(I, \mathbf{s}^{(t)}\right)\right), \quad t = 0, \ldots, T-1 \tag{9}$$

where $\mathbf{s}^{(0)}$ is the mean shape aligned to the face bounding box, $\mathcal{R}_t$ is an ensemble of shallow regression trees at the stage t, and $\Phi\left(I, \mathbf{s}^{(t)}\right)$ represents the local feature vector extracted around the current predicted positions. A typical pixel-difference feature computes the intensity difference between two anchor points mapped by a similarity transform:

$$\phi_j = I\left(\mathbf{p}_j^{(1)}\right) - I\left(\mathbf{p}_j^{(2)}\right), \qquad \mathbf{p}_j^{(k)} = s, \mathbf{R}, \mathbf{q}_j^{(k)} + \mathbf{t} \tag{10}$$

where $\mathbf{q}_j^{(k)}$ is fixed anchor coordinates in the mean-shape space, $s > 0$ is the scaling factor, $\mathbf{R} \in SO(2)$ is the rotation matrix, and $\mathbf{t} \in \mathbb{R}^2$ is the translation vector. All parameters of the regression stages $\mathcal{R}_t$ are learned by minimizing the squared error over N labeled samples:

$$\min_{\mathcal{R}t} \sum i = 1^N \left|\mathbf{s}_i^\star - \left(\mathbf{s}i^{(0)} + \sum t = 0^{T-1} \mathcal{R}_t\left(\Phi\left(I_i, \mathbf{s}_i^{(t)}\right)\right)\right)\right|_2^2 \tag{11}$$

where $\mathbf{s}_i^\star$ denotes the ground-truth landmarks of image I_i, and $\mathbf{s}_i^{(0)}$ is the mean shape aligned to the corresponding ROI. Owing to its simple local features and shallow regression trees, ERT achieves low latency and remains stable under variations in lighting, pose, and facial expressions [24]. The predictor used in Dlib was trained on the 300-W dataset [25–27].

3.2 BiGRU Model for Mouth State Classification

The overall architecture of the proposed BiGRU attention, is presented in Fig. 3, which was applied in the proposed model.

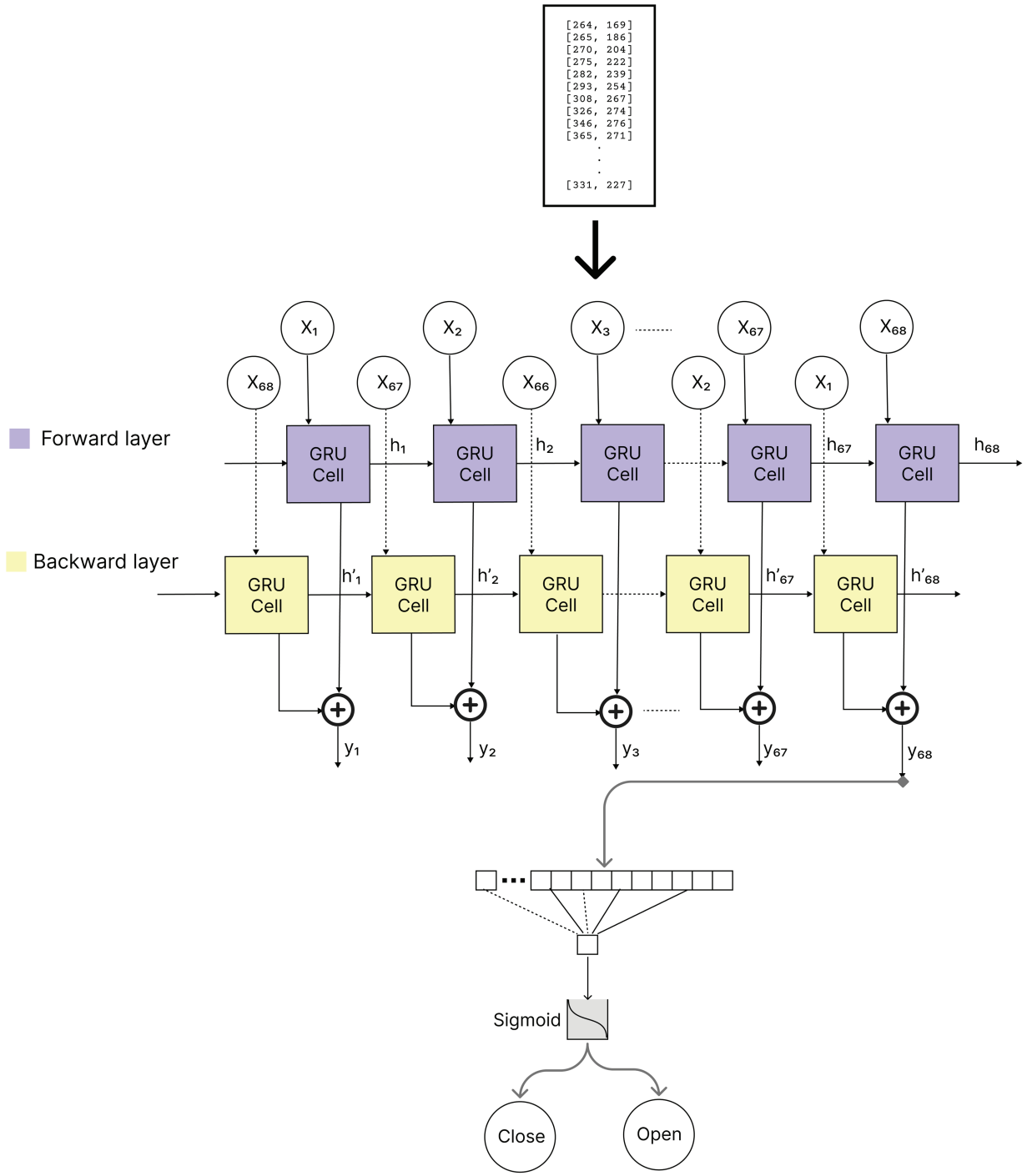

Figure 3: The architecture of our BiGRU model

The time-normalized landmark sequence $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_T]^\top$ is processed by a bidirectional GRU to capture the dynamics of mouth opening across short temporal windows. On the forward branch, the reset and update gates regulate how past memory is blended with new evidence at a time t:

$$\begin{cases} \mathbf{r}_t = \sigma\left(\mathbf{W}_r\, \mathbf{z}_t + \mathbf{U}_r\, \mathbf{h}_{t-1} + \mathbf{b}_r\right) \\ \mathbf{u}_t = \sigma\left(\mathbf{W}_u\, \mathbf{z}_t + \mathbf{U}_u\, \mathbf{h}_{t-1} + \mathbf{b}_u\right) \end{cases} \tag{12}$$

here, $\mathbf{z}_t \in \mathbb{R}^{2L}$ denotes the landmark vector at frame t, $\mathbf{h}_{t-1} \in \mathbb{R}^H$ is the previous hidden state, $\mathbf{W}, \mathbf{U}, \mathbf{b}$ are trainable parameters, and $\sigma(\,)$ is the logistic sigmoid. Given these gates, the candidate state injects current evidence modulated by the reset gate, stabilizing the dynamics:

$$\tilde{\mathbf{h}}_t = \tanh\left(\mathbf{W}_h, \mathbf{z}_t + \mathbf{U}_h\left(\mathbf{r}_t \odot \mathbf{h}_{t-1}\right) + \mathbf{b}_h\right) \tag{13}$$

where $\odot$ is the Hadamard product and tanh is the hyperbolic tangent. The new hidden state then balances memory retention and innovation via the update gate:

$$\mathbf{h}_t = (\mathbf{1} - \mathbf{u}_t) \odot \mathbf{h}_{t-1} + \mathbf{u}_t \odot \tilde{\mathbf{h}}_t \tag{14}$$

From (14), $\mathbf{1}$ is the all ones vector in $\mathbb{R}^H$. A backward branch mirrors these updates on the time-reversed sequence to encode short-term future context; both directions are concatenated into a time-local representation:

$$\mathbf{h}_t^{\text{bi}} = \left[\overrightarrow{\mathbf{h}}_t\,;\, \overleftarrow{\mathbf{h}}_t\right] \in \mathbb{R}^{2H} \tag{15}$$

To form a compact sequence descriptor aligned with our downstream classifier and latency constraints, we apply additive attention. This mechanism emphasizes frames near state transitions rather than averaging uniformly:

$$e_t = \mathbf{v}^\top \tanh\left(\mathbf{W}_a, \mathbf{h}_t^{\text{bi}} + \mathbf{b}_a\right), \quad \alpha_t = \frac{\exp\left(e_t\right)}{\sum j = 1^T \exp\left(e_j\right)}, \quad \overline{\mathbf{h}} = \sum t = 1^T \alpha_t, \mathbf{h}_t^{\text{bi}} \tag{16}$$

where $\alpha_t \geq 0\ with \sum_t \alpha_t = 1$ are temporal weights, and $\overline{\mathbf{h}} \in \mathbb{R}^{2H}$ is the sequence embedding used for classification. The logistic head produces the probability that the mouth is open:

$$p = \sigma\left(\mathbf{w}_o^\top \overline{\mathbf{h}} + b_o\right), \tag{17}$$

where $p \in (0,1)$ is the predicted probability and $(\mathbf{w}_o, b_o)$ are trainable parameters. To optimize discriminability under class imbalance typical of real-world monitoring streams, we minimize a weighted binary cross entropy:

$$\mathcal{L}_{\text{BCE}} = -w_1 \cdot y \cdot \log p - w_0 \cdot (1 - y) \cdot \log(1 - p) \tag{18}$$

with label $y \in \{0,1\}$ and positive weights w_1, w_0 compensating class skew. For more biased regimes, we adopt focal loss to emphasize hard cases near decision boundaries:

$$\mathcal{L}_{\text{focal}} = -\alpha\, y\, (1 - p)^\gamma \log p - (1 - \alpha)\,(1 - y)\, p^\gamma \log(1 - p) \tag{19}$$

where $\alpha \in (0,1)$ and $\gamma \geq 0$ control the focus on difficult samples. Finally, to maintain a compact recurrent core consistent with real time deployment, we regularize all trainable weights:

$$\mathcal{L} = \mathcal{L}_{\text{BCE/focal}} + \lambda_w \|\Theta\|_2^2, \tag{20}$$

where Θ denotes the full parameter set and $\lambda_w \geq 0$ tunes the strength of L_2 regularization. This BiGRU with attention design thus concentrates capacity on the most informative temporal segments while preserving the low latency required by our healthcare monitoring pipeline.

3.3 Multi-Objective CLPSO for Feature Selection and Hyperparameter Tuning

To satisfy the real-time constraint without sacrificing accuracy, we cast model configuration as a joint search over a sparse feature set derived from the landmark sequence and the BiGRU hyperparameters. We employ Comprehensive Learning Particle Swarm Optimization (CLPSO), which mitigates premature convergence by letting each coordinate learn from different exemplars. We encode each candidate as a concatenation of a binary feature mask and continuous hyperparameters is $\mathbf{x} = [\mathbf{m}; \boldsymbol{\theta}]$. Given $\mathbf{x}$, we train and evaluate with short epoch K-fold CV and measure single-batch inference latency on the target hardware. The unified objective balances both goals and discourages bloated inputs/models:

$$J(\mathbf{x}) = \lambda [1 - \text{F1}_{\text{cv}}(\mathbf{x})] + (1-\lambda) \frac{\text{lat}(\mathbf{x})}{\text{lat}_{\text{ref}}} + \rho \max(0, \text{lat}(\mathbf{x}) - \text{lat}_{\text{bud}}) + \mu_1 \frac{\|\mathbf{m}\|_0}{d} + \mu_2 \frac{\text{Params}(\boldsymbol{\theta})}{\text{Params}_{\text{ref}}}. \tag{21}$$

From (21), $\lambda \in [0,1]$ trades accuracy vs. latency; lat is end to end latency; lat_{ref} normalizes scale; lat_{bud} is the latency budget with penalty $\rho \gg 0$; $\|\mathbf{m}\|_0$ counts selected features; and Params $(\boldsymbol{\theta})$ is model size relative to a reference, controlled by $\mu_1, \mu_2 \geq 0$ to enforce sparsity and compactness.

To reduce variance in accuracy estimation during the search, we average Recall across folds with early stopping inside each fold:

$$\text{Recall}_{\text{cv}}(\mathbf{x}) = \frac{1}{K} \sum k = 1^K \text{Recall}^{(k)}(\mathbf{x}) \tag{22}$$

where $\text{Recall}^{(k)}(\mathbf{x})$ is the fold-k score obtained by training the BiGRU configured by $(\mathbf{m}, \boldsymbol{\theta})$ for a small number of epochs with early stopping. CLPSO updates each coordinate using a comprehensive learning rule every dimension may learn from a different exemplar personal best then integrates velocity to position:

$$v_{i,j}^{(t+1)} = \omega v_{i,j}^{(t)} + c\, r_{i,j}^{(t)} \left(p_{f(i,j),j}^{(t)} - x_{i,j}^{(t)} \right), \qquad x_{i,j}^{(t+1)} = x_{i,j}^{(t)} + v_{i,j}^{(t+1)}. \tag{23}$$

From (23), $x_{i,j}$ and $v_{i,j}$ are the j-th coordinate and velocity of particle i; ω is inertia; c is acceleration; $r_{i,j}^{(t)} \sim \mathcal{U}(0,1)$ and $p_{f(i,j),j}^{(t)}$ is the exemplar personal best at dimension j chosen by CLPSO's learning strategy. Since $\mathbf{m}$ is binary while the swarm evolves in $\mathbb{R}^d$, we recover sparsity by a sigmoid gate followed by thresholding:

$$m_{i,j} = \mathbf{1}\{ \sigma(x_{i,j}) \geq \tau_b \}, \tag{24}$$

where $\tau_b \in (0,1)$ sets sparsity, σ is the logistic sigmoid, and $\mathbf{1}\{\cdot\}$ is the indicator. With these updates and the objective in (21), CLPSO returns $(\mathbf{m}^*, \boldsymbol{\theta}^*)$ that meets the latency budget while maximizing accuracy, yielding a compact BiGRU configuration aligned with our healthcare monitoring constraints.

4 Experiments and Results

4.1 Dataset

The dataset was collected from 5 healthy volunteers (3 males and 2 females) aged between 22 and 28 years. All participants were recruited from the university research laboratory. None of the participants had a history of neuromuscular disorders or facial paralysis. They were instructed to simulate slow mouth movements under controlled laboratory lighting conditions. The demographic details of the participants are summarized in Table 1. This study serves as a pre-clinical pilot to validate the feasibility of the proposed algorithm before deploying it on patients with dysphagia.

Table 1: Participant demographics

Participant ID	Gender	Age	Health status	Role
P01	Male	24	Healthy	Researcher
P02	Female	23	Healthy	Student
P03	Male	28	Healthy	Researcher
P04	Female	22	Healthy	Student
P05	Male	25	Healthy	Student
Total	3M/2F	24.4 (Mean)	–	–

To ensure the reliability of the dataset, a rigorous manual labeling process was implemented. Two independent annotators were trained to classify mouth states based on standard visual criteria: the mouth was labeled as "Open" if there was a visible separation between the lips or teeth exceeding 2 mm; otherwise, it was labeled as "Closed" [15]. Any discrepancies between the two annotators were reviewed and resolved by a third senior researcher to reach a final consensus. This cross-verification process minimizes subjective bias and ensures high inter-rater agreement for the binary classification task. We established an experimental environment to record mouth movement images. Participants were asked to sit in front of the camera module in the Raspberry Pi 4 and perform the actions as follows:

- The participant sat approximately one meter away from the recording camera.
- The participant looked straight ahead and opened and closed their mouth at least three times while maintaining a forward gaze.
- The participant turned their face to the left and opened and closed their mouth at least three times.
- The participant turned their face to the right and opened and closed their mouth at least three times.
- The participant tilted their head upward and opened and closed their mouth at least three times.
- The participant tilted their head downward and opened and closed their mouth at least three times.

After the recording, video frames were extracted for analysis. Each frame was manually labeled to determine whether the participant's mouth was open or closed based on the visual appearance in the image. Table 2 summarizes the dataset used in this study.

Table 2: Dataset statistics

# of Persons	Mouth state	# of samples	Ratio	Total
5	Closed	1570	50.55%	3106
	Open	1536	49.45%	

4.2 Experimental Environment

In this experiment, the environment has been designed with two architectures the first is composed of a workstation with Windows 10 Pro and the second is a light weight Raspberry Pi 4. The workstation was equipped with an Intel Core i5-10500 processor operating at 3.1 GHz, featuring six cores and twelve threads, along with 24.0 GB of DDR4 memory. In addition, an NVIDIA RTX 3070 GPU was employed to accelerate parallel computations through CUDA (Compute Unified Device Architecture). The lightweight solution includes Raspberry Pi 4 board and the Camera Module 3. The Raspberry Pi 4 Model B represents the latest generation of the Raspberry Pi microcomputer series, powered by the Broadcom BCM2711 SoC, which integrates a quad-core ARM Cortex-A72 (ARM v8, 64-bit) CPU running at 1.5 GHz. The board supports 4 GB of LPDDR4 RAM and offers modern connectivity options such as USB 3.0, Gigabit Ethernet, Bluetooth 5.0, and dual-band Wi-Fi. Notably, it includes two micro-HDMI ports capable of 4K video output and a VideoCore VI GPU that provides strong graphics processing performance. Fig. 4 illustrates the Raspberry Pi 4 Model B board.

Figure 4: Raspberry Pi 4 model B board

The hardware specifications of the Camera Module 3 is presented in Table 3.

Table 3: Hardware specification of camera module 3

Specification	Description
Net price	$25
Size	Around 25 × 24 × 11.5 mm
Weight	4 g
Still resolution	11.9 Megapixels
Video modes	2304 × 1296p56, 2304 × 1296p30 HDR, 1536 × 864p120
Sensor	Sony IMX708
Sensor resolution	4608 × 2592 pixels

4.3 Model Preprocessing and Training

In this study, we set up the Camera Module 3 of the Raspberry Pi 4 board as an edge device. The Dlib library was then employed to detect 68 facial landmark points in each image. The two-dimensional coordinates of these landmarks were directly used to train the machine learning models. The dataset was divided into 85% for training and 15% for testing. During training, we applied the k-fold cross-validation technique with k equal to 5. Table 4 summarizes the training configuration parameters. We also employed the heuristic optimization algorithm CLPSO to search for the optimal parameters of the GRU model. Table 5 summarizes the search space of the CLPSO algorithm.

Table 4: Training parameters

Parameters	Value
# of Epochs	100
Learning rate	5.00×10^{-4}
Optimizer	Adam
Batch size	64
Training split ratio	0.85 training, 0.15 testing
K-fold cross validation	On the training set with k = 5

Table 5: CLPSO search space for GRU hyperparameter optimization

Variable	Search range	Type	Description
sequence length	68	Integer	Input time frames for BiGRU.
population Size	24	Integer	Number of particles in CLPSO.
max Iterations	30	Integer	Maximum optimization cycles.
hidden_size	32 → 128	Integer	Number of hidden units in each GRU layer.
num_layers	1 → 2	Integer	Number of stacked GRU layers.
bidirectional	0 → 1	Binary	Indicates whether the GRU is bidirectional (0 = unidirectional, 1 = bidirectional).

To standardize the temporal input, the sequence length was fixed at T = 68. At each time step, the BiGRU receives a 2-D feature vector $l_t = (x_t, y_t)$ corresponding to the t-th landmark. This sequential formulation allows the network to model contextual dependencies among neighboring landmarks while keeping the number of trainable parameters manageable. Since the input data had a size of (68, 2), the search space for the GRU network's hidden size could not be too large. The range from 32 to 128 is appropriate for optimizing the performance of the GRU model. The number of layers of the GRU model was considered within the range [1,2]. A larger number of layers can improve the representation capacity of the model, though it may also lead to overfitting for a sequence length of 68. In addition, we tested the Bidirectional configuration as both False and True to compare the performance between the traditional GRU and the Bidirectional GRU models. For the optimization execution, the CLPSO algorithm was configured with a population size of 24 and a maximum of 30 iterations.

4.4 Evaluation Metrics

In machine learning and pattern recognition, evaluation metrics serve as quantitative indicators for measuring the performance and effectiveness of a model. These metrics are used to assess and compare how well a model performs on a given task or problem. In the context of this study, we employed the following metrics:

- **Accuracy:** Measures the overall correctness of the model's predictions.
- **Precision:** Evaluates the proportion of correctly predicted positive samples among all samples predicted as positive.
- **Recall:** Measures the proportion of correctly predicted positive samples among all actual positive samples.
- **F1 score:** A comprehensive measure that combines both precision and recall.
- **Cosine Similarity:** Assesses the similarity between the predicted labels vector and the true labels vector by computing the cosine of the angle between them, reflecting their closeness in vector space.
- **ROC AUC (Area Under the Receiver Operating Characteristic Curve):** Measures the model's ability to distinguish between classes by considering all possible classification thresholds.
- **Precision–Recall (Average Precision—AP):** Summarizes the relationship between precision and recall across multiple thresholds, particularly useful in cases of class imbalance.

These metrics collectively reflect the model's performance in distinguishing between different classes. Table 6 presents the definition of the confusion matrix used in machine learning.

Table 6: Confusion matrix

		Predicted values	
		Negative (0)	**Positive (1)**
Actual values	Negative (0)	TN (True Negative)	FP (False Negative)
	Positive (1)	FN (False Positive)	TP (True Positive)

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{25}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{26}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{27}$$

$$\text{F1 score} = \frac{2 * Precision * Recall}{Precision + Recall} \tag{28}$$

4.5 Experiment Result

After executing the CLPSO algorithm to search for the optimal configuration, we obtained the best setup for the GRU model with parameters: hidden_size = 128, num_layers = 2, bidir = True. Fig. 5 shows the training plot of the BiGRU-CLPSO model during k-fold cross-validation. To further analyze the contribution of CLPSO-based optimization, we conducted comparative experiments using a baseline GRU model configured with hidden_size = 32, num_layers = 1, and bidirectional = True. In addition, the GRU was further tuned using the Random Search algorithm. The optimal GRU configuration obtained by Random Search was hidden_size = 101, num_layers = 1, and bidirectional = True. Table 7 summarizes the performance comparison among the baseline GRU, the GRU optimized by Random Search, and the proposed BiGRU-CLPSO model on the testing dataset.

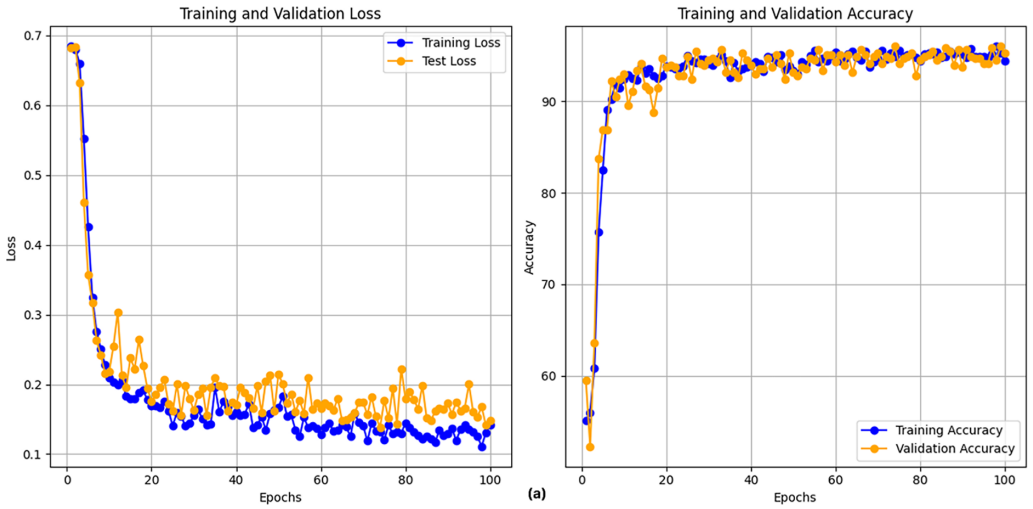

Figure 5: (Continued)

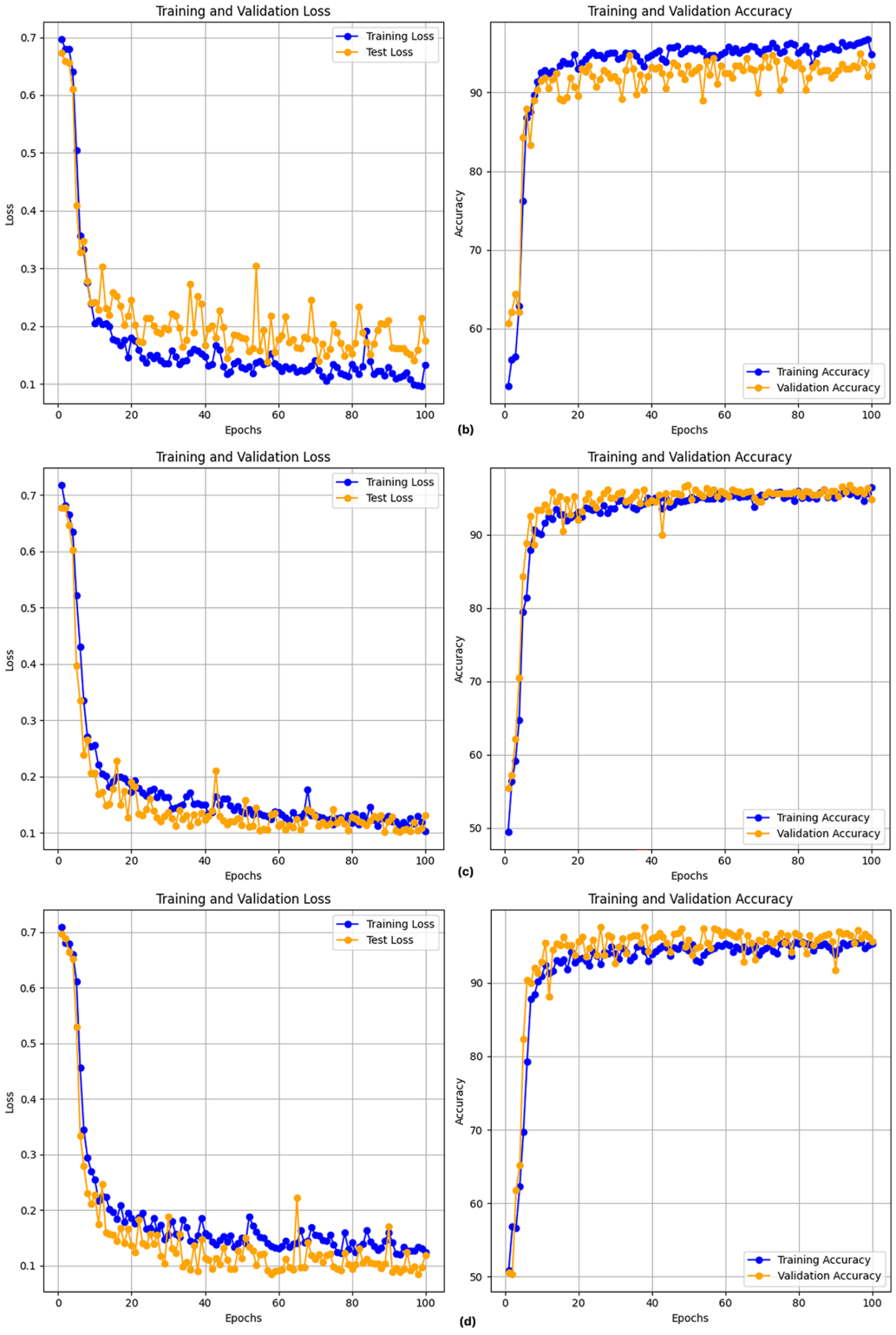

Figure 5: (Continued)

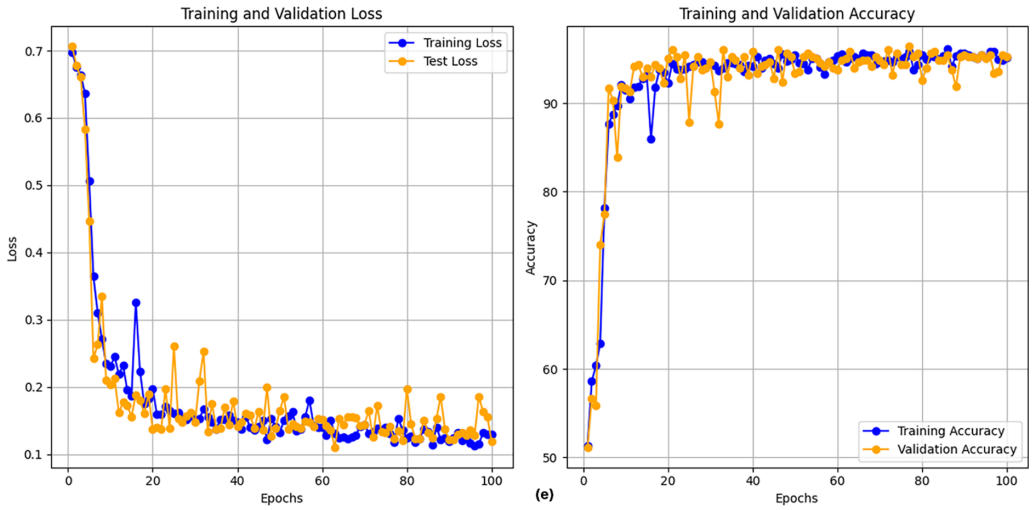

Figure 5: Training plot with cross-validation of of BiGRU-CLPSO model

Table 7: Ablation study—impact of CLPSO optimization on model performance

Model	Loss	Accuracy	Precision	Recall	F1 score	CS
GRU (Baseline)	0.5737	0.6760	0.6840	0.6695	0.6767	0.6767
GRU (Random Search)	0.1771	0.9442	0.9816	0.9068	0.9427	0.9435
BiGRU-CLPSO	0.1084	0.9657	0.9825	0.9492	0.9655	0.9657

Table 7 serves as an ablation study to quantify the specific contribution of the CLPSO component. When the GRU model is used without optimization (Baseline), the accuracy is limited to 67.60%. Applying a standard Random Search improves performance significantly to 94.42%. However, the proposed BiGRU-CLPSO integration yields the highest accuracy of 96.57% and an F1-score of 0.9657. This incremental improvement demonstrates that the CLPSO algorithm effectively navigates the hyperparameter search space to locate a global optimum that manual or random tuning might miss, thereby validating the necessity of the optimization module in the proposed architecture.

The cross-validation results on the training dataset, as shown in Fig. 6, reveal clear performance differences among the models in the binary classification task based on sequences of two-dimensional coordinate features.

Starting with the loss function values, when trained for 100 epochs, XGBoost achieved the lowest training loss but showed clear signs of overfitting, ranking only third out of five algorithms in validation loss. In contrast, both LSTM and GRU maintained stable loss values across training and validation phases. Notably, GRU achieved the second-lowest loss during training and the lowest loss during validation. Conversely, CNN exhibited relatively high loss values, while MLP had the highest loss in both training and validation phases.

Overall, GRU achieved the highest and most stable performance across all evaluation metrics, including Accuracy, Precision, Recall, F1-score, and Cosine Similarity. The model maintained a high median value with a narrow range across folds, reflecting strong generalization capability. Importantly, the evaluation emphasized not only the training phase but also the validation phase, where overfitting was evident in

models such as XGBoost. LSTM followed GRU in terms of stability and performance, ranking second on most metrics, including Accuracy, Recall, and F1-score. However, GRU outperformed LSTM in stability, as reflected by the smallest gap between training and validation results.

CNN demonstrated moderate performance, with relatively high precision but significantly lower recall, leading to a reduced F1-score. This suggests that CNN tended to minimize false positives at the expense of missing many true positives. XGBoost exhibited typical overfitting behavior, performing almost perfectly on the training set with near-zero loss and an F1-score close to 1.0. However, its validation performance did not remain consistent and was lower than that of LSTM, and notably inferior to GRU. Finally, MLP showed the weakest performance, with uniformly low results across most metrics, indicating its limitation in capturing the sequential structure of temporal data.

Regarding the precision–recall relationship, LSTM and GRU maintained a good balance, resulting in high and stable F1-scores. CNN favored precision, while XGBoost achieved high precision but showed large fluctuations in recall across folds, and MLP performed poorly in both metrics. For Cosine Similarity, LSTM and GRU again outperformed other models, indicating that the logit distributions of these two models were more closely aligned with the true label distribution.

From these findings, it can be concluded that GRU and LSTM are the two best-performing models on the training dataset. In particular, GRU stands out as the optimal choice due to its superior stability and slightly better generalization compared to LSTM. CNN and XGBoost may be considered for specific use cases but are not suitable as primary models, while MLP demonstrated the weakest performance overall.

Fig. 7 illustrates the average values of all evaluation metrics obtained during cross-validation on the training dataset, presented in the form of a radar chart. This visualization provides an overview of the comparative performance among the models. The results clearly show that GRU and LSTM lead with high and well-balanced average scores across most criteria, including Loss, Cosine Similarity, Accuracy, Precision, Recall, and F1-score. The radar boundaries of these two models are broader and more stable than those of the others, with GRU consistently outperforming LSTM. This further reinforces the conclusion drawn from Fig. 6 that GRU is the optimal choice on the training dataset.

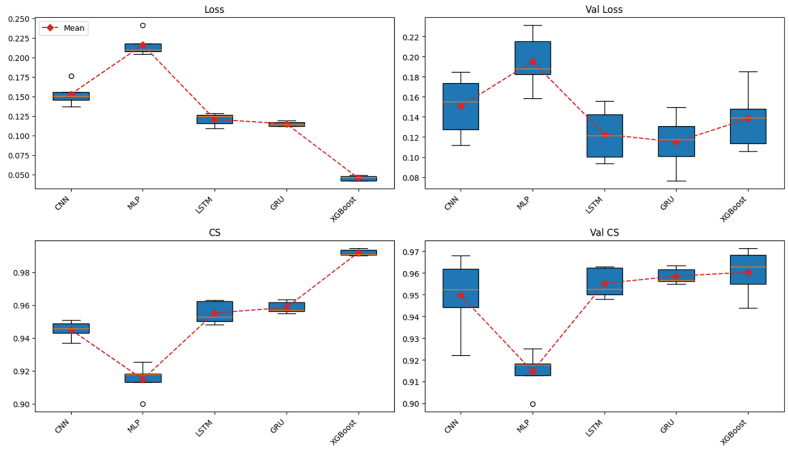

Figure 6: (Continued)

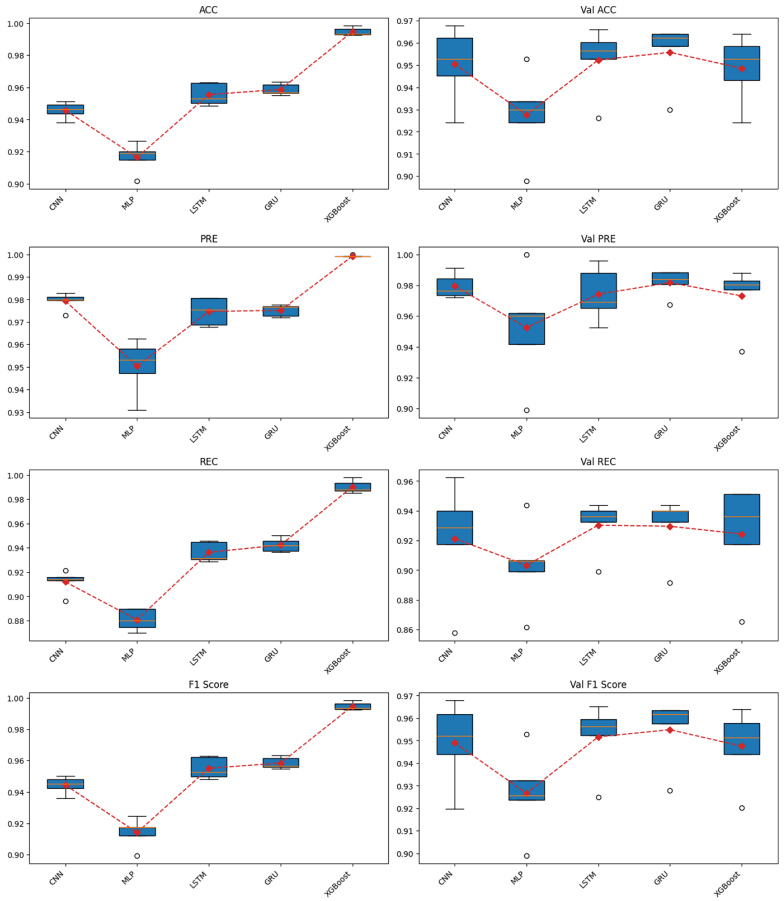

Figure 6: Cross-validation results across multiple models on the training dataset

CNN achieved moderate overall performance, showing a relatively strong result in Precision; however, its lower Recall reduced the overall balance of the model. XGBoost, despite achieving very high values on the training set, demonstrated considerable discrepancies when compared with validation results, clearly reflecting the overfitting tendency previously discussed. Lastly, MLP was the weakest model, with all average metrics remaining low and its radar boundary concentrated near the center of the chart, indicating limited capability in capturing sequential data features.

Fig. 7 not only confirms the findings from Fig. 6 but also provides an intuitive view of the overall comprehensiveness of each model. GRU and LSTM demonstrate superiority in both accuracy and balance across evaluation metrics.

After training all models, we proceeded to evaluate their performance on the testing dataset.

Fig. 8 presents the performance of the models on the independent test dataset, which directly reflects their generalization ability. The results show that the GRU model continues to lead, achieving the lowest Loss value of 0.11, while maintaining Accuracy, Precision, F1-score, and Cosine Similarity values above 0.97. Notably, GRU demonstrates a well-balanced relationship between Precision and Recall, resulting in the highest F1-score of 0.97. This indicates that the model remains stable and reliable when applied to previously unseen data.

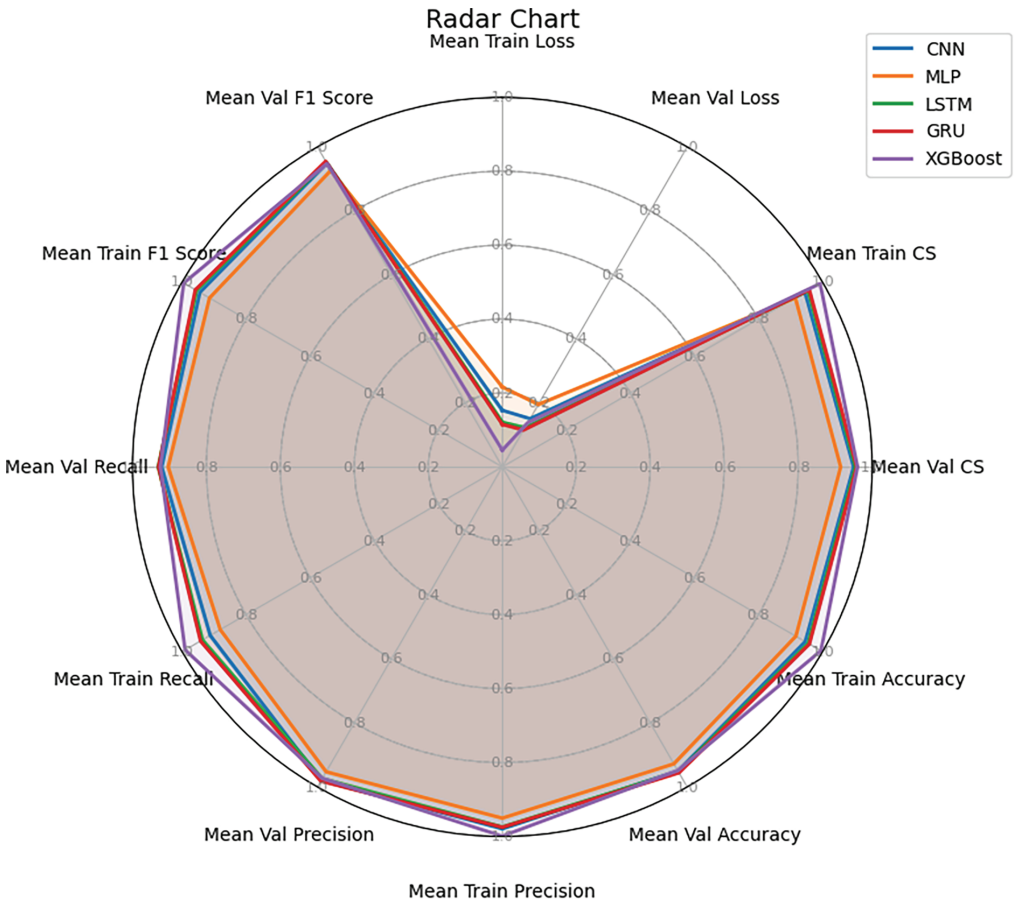

Figure 7: Radar chart of mean performance across multiple models

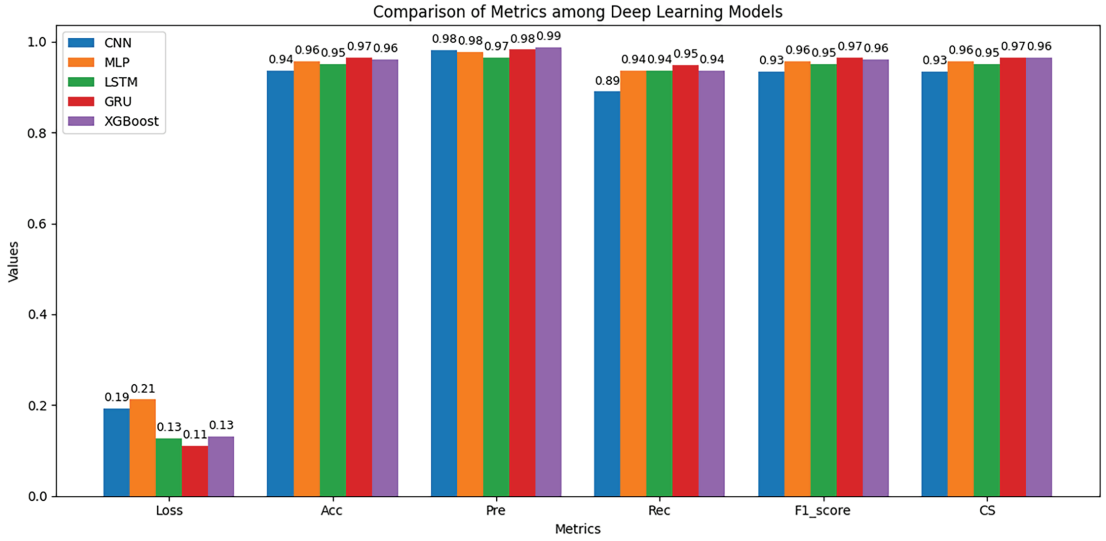

Figure 8: Performance comparison on the testing set

XGBoost ranks second, also exhibiting competitive performance, particularly in Precision, which reached 0.99. This suggests that XGBoost is effective at minimizing false positive predictions. However, its lower Recall compared to GRU indicates that it missed some true positive cases, leading to less balanced overall performance. In remote patient monitoring scenarios, the primary priority is to detect abnormal behaviors such as difficulty swallowing, swallowing errors, or prolonged mouth opening as early as possible. Therefore, any unusual inability of a patient to close the mouth must trigger an immediate alert. For such systems, Recall is the most critical metric. The high Recall and F1-score values achieved by GRU clearly demonstrate its superior performance and robustness for remote dysphagia monitoring applications.

Unlike the results observed during cross-validation on the training dataset, the LSTM model performed near the bottom on most evaluation metrics in this test phase. CNN yielded the lowest overall performance, with a Loss exceeding 0.19 and the lowest Recall value of 0.89, resulting in a reduced F1-score of 0.93. This outcome highlights the limitations of CNN in capturing the sequential dependencies inherent in temporal data. MLP ranked third, performing slightly better than CNN and LSTM but still falling short of GRU and XGBoost.

Overall, the results on the test dataset reinforce and extend the earlier analyses. GRU stands out as the most effective and suitable model for this task, excelling not only during cross-validation but also when evaluated on entirely new data. XGBoost may be considered as a supplementary model in applications that prioritize Precision, whereas LSTM, CNN, and MLP exhibit limited performance and are not yet suitable as primary models.

Fig. 9 presents the ROC curves and AUC values of the models on the test dataset. The results show that all models achieved very high AUC values above 0.976, indicating strong capability in distinguishing between the two classes. Among them, the GRU model obtained the highest AUC value of 0.990, confirming its superior classification performance compared with the other models. XGBoost with a value of 0.988 and LSTM with a value of 0.986 followed closely, both exhibiting ROC curves that lie near the upper-left corner, demonstrating their ability to maintain a high true positive rate even at low false alarm rates. CNN achieved an AUC of 0.982, which is relatively high; however, its overall performance was weaker due to the lower Recall observed in Fig. 8. Finally, MLP recorded the lowest AUC value of 0.976, consistent with its weaker performance on both the training and test datasets. Thus, the ROC–AUC analysis reinforces the previous findings, confirming that GRU is the most comprehensive and stable model on the test dataset.

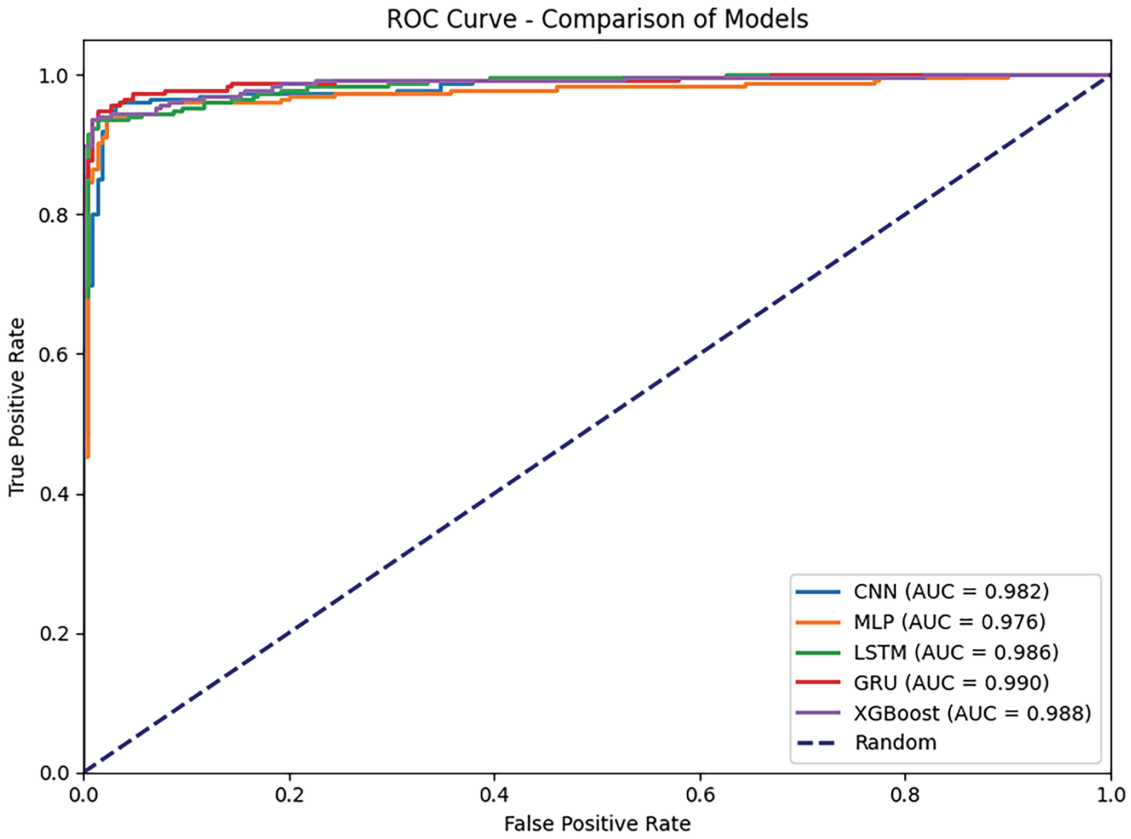

Figure 9: ROC-AUC comparison across multiple models

Fig. 10 presents the Precision–Recall (PR) curves and the Average Precision (AP) values of the models on the test dataset. This metric is particularly important for assessing the trade-off between precision and recall across different decision thresholds. The results show that all models achieved very high AP values, with values equal to or exceeding 0.982, confirming their stable classification capability. GRU achieved the highest AP of 0.993, with a PR curve closely following the upper boundary and maintaining stability in the high-recall region. This demonstrates the model's ability to sustain high precision while minimizing missed positive cases. XGBoost also exhibited highly competitive performance, with an AP of 0.991, approaching GRU's result and maintaining consistently high precision across the recall range. LSTM achieved an AP of 0.988, showing a good balance between precision and recall, though its curve was slightly lower than those of GRU and XGBoost in the high-recall region. CNN with AP value of 0.985 and MLP with AP value of 0.982 also achieved reasonably high results. However, their PR curves declined as recall approached 1.0, reflecting limitations in detecting all positive cases. This observation aligns with the analysis in Figs. 7 and 9, where both models exhibited lower recall compared to the top-performing group.

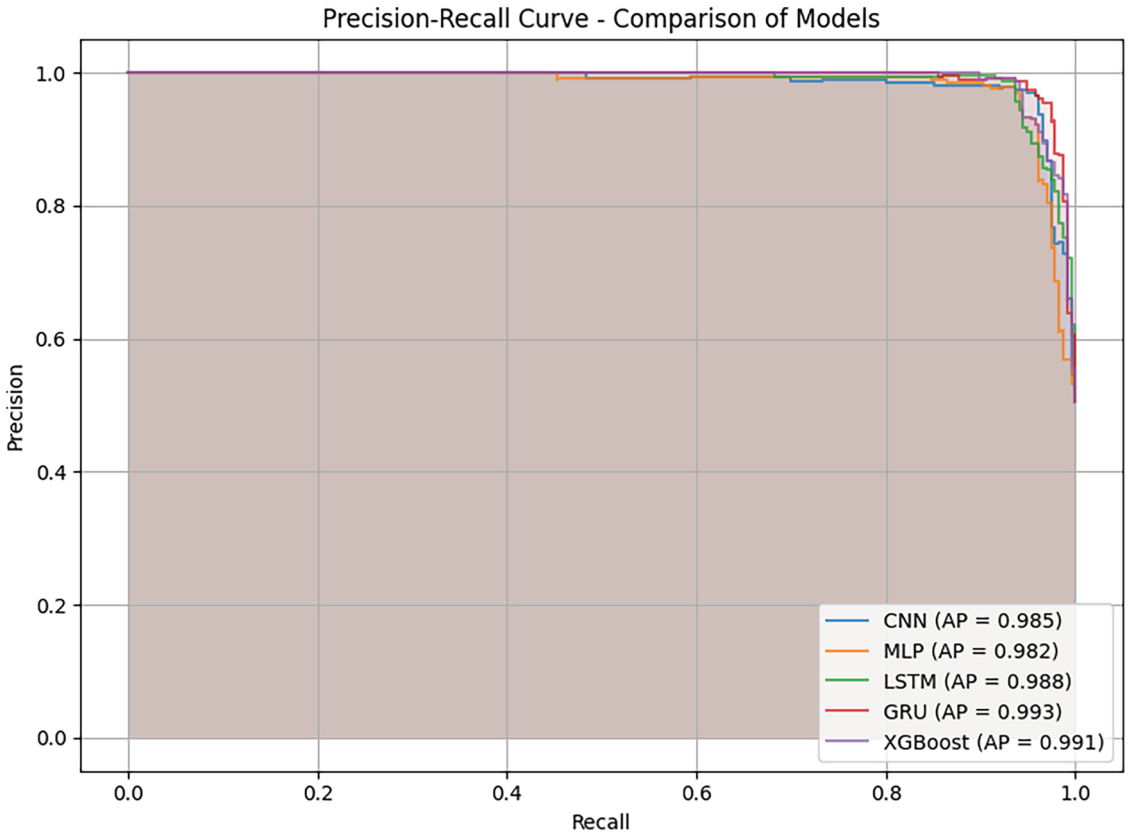

Figure 10: Precision–recall (PR) curves and average precision (AP) values comparison across multiple models

To validate the reliability of the performance improvements observed in the figures, we conducted paired t-tests (at a significance level of $\alpha = 0.05$) comparing the F1-scores of the proposed BiGRU-CLPSO model against the other classifiers. Table 8 summarizes these statistical results.

Table 8: Statistical significance test results

Comparison pair	t-value	p-value	Significance ($\alpha < 0.05$)
BiGRU-CLPSO vs. CNN	1.8186	0.1431	
BiGRU-CLPSO vs. MLP	5.6045	0.0050	Y
BiGRU-CLPSO vs. LSTM	2.4477	0.0706	
BiGRU-CLPSO vs. XGBoost	3.1277	0.0353	Y

First, the analysis confirms that BiGRU-CLPSO statistically outperforms the traditional machine learning baselines, specifically MLP ($p = 0.005 < 0.01$) and XGBoost ($p = 0.0353 < 0.05$). This statistically significant difference provides robust evidence that the proposed deep learning architecture generalizes far better than shallow models. While XGBoost demonstrated high training accuracy, its significantly lower validation stability indicates a tendency toward overfitting on limited datasets. In contrast, the BiGRU model effectively captures the temporal dependencies in landmark sequences that MLP and XGBoost fail to model adequately. Second, regarding the deep learning comparisons, the performance differences against CNN ($p = 0.1431$) and LSTM ($p = 0.0706$) were not statistically significant ($p > 0.05$). The p-value for LSTM, while

marginally close to the threshold, suggests that both RNN variants (LSTM and BiGRU) possess comparable predictive power for this specific task. However, in the context of edge computing, statistical accuracy is not the sole criterion for model selection; computational complexity is equally critical. The BiGRU architecture is structurally simpler, utilizing only two gating mechanisms (reset and update) compared to the three-gate structure (input, output, and forget) of the LSTM. This architectural difference translates to fewer trainable parameters and reduced matrix operations during inference. Consequently, since BiGRU-CLPSO achieves a statistically equivalent level of accuracy to LSTM but with greater computational efficiency, it is verified as the optimal choice for deployment on resource constrained platforms like Raspberry Pi.

The experimental results on both the training dataset via cross-validation and the independent test dataset demonstrate consistent performance patterns across models. The training analyses from Figs. 5 and 6 identified GRU and LSTM as the leading models, characterized by low loss values, high accuracy, and an optimal balance between precision and recall. CNN and MLP showed less stability, with CNN favoring precision but lacking recall, and MLP producing the lowest overall results. XGBoost achieved very high training accuracy but showed signs of overfitting, though it still maintained competitive test performance. The test results presented in Figs. 7–9 further supported these findings. GRU stood out as the most comprehensive and stable model, achieving the lowest loss, the highest F1-score, and excellent AUC and AP scores of 0.990 and 0.993, respectively. LSTM and XGBoost also demonstrated strong performance, with LSTM maintaining a good balance and XGBoost excelling in precision. Although CNN and MLP achieved high AUC and AP scores exceeding 0.97, they lacked the balance and stability exhibited by the leading models.

In conclusion, these experimental results demonstrate that BiGRU-CLPSO is the optimal model for classifying mouth states based on sequential coordinate data. The model not only outperforms the baseline GRU and the GRU optimized using the Random Search algorithm but also exhibits greater stability and superior performance compared with traditional models such as CNN, MLP, LSTM, and XGBoost. These improvements are consistent across multiple evaluation metrics, including accuracy, precision, recall, F1-score, cosine similarity, ROC–AUC, and the precision–recall curve. These findings highlight the robustness of BiGRU in handling sequential data and the effectiveness of CLPSO in searching for the optimal hyperparameter configuration for the BiGRU model.

4.6 Real-Time Performance on Edge Device

Beyond algorithmic accuracy, the practical viability of a healthcare monitoring system depends on its operational efficiency on resource-constrained hardware. To validate this, we monitored the system performance directly on the Raspberry Pi 4 device under continuous inference load. The hardware metrics derived from system resource analysis are summarized in Table 9.

Table 9: Hardware performance analysis on raspberry Pi 4

Metric	Value
Throughput	11.3 FPS
Inference latency	88.5 ± 6.7 ms
CPU usage	31.40%
Memory usage	598 MB
Power consumption	4.2 W

As detailed in Table 9, the system achieves a stable throughput of approximately 11.3 FPS, corresponding to an average inference latency of 88.5 ms. While this frame rate is lower than standard video streaming

(30 FPS), it is highly effective for patient monitoring applications. Physiological mouth movements, such as opening for feeding or yawning, typically occur over durations of 2 to 5 s. With a sampling rate of ~11 times per second, the system captures sufficient temporal granularity to reconstruct the motion dynamics without missing critical state transitions. The low latency variance (±6.7 ms) further confirms the temporal stability of the pipeline, which is crucial for real time alerting. Regarding computational resources, the system demonstrates exceptional efficiency. The memory usage is 598 MB, only up to 15% of the devicecapacity. Notably, the system-wide CPU user load is approximately 31.4%. Since the Raspberry Pi 4 utilizes a Quad-core Cortex-A72 processor, this moderate load profile indicates that the lightweight BiGRU architecture is computationally efficient, leaving substantial headroom (~68%) for concurrent background tasks. This is a critical feature for IoT deployment, as it allows the device to simultaneously handle data encryption, wireless transmission, and system health checks without experiencing thermal throttling or performance degradation.

4.7 Comparison

We compared the results of this study with several recent prominent works in the same field of mouth-state recognition. The selected studies include Ji et al. (2018) [33] and Ishmam et al. (2022) [34]. Ji et al. (2018) proposed a contour-based algorithm to detect whether the eyes and mouth are open or closed, aiming to identify driver fatigue states. Ishmam et al. (2022) introduced two lightweight CNN models for recognizing lip states, open or closed, to support non-verbal communication for individuals with hearing impairment. Instead of using raw images directly, the authors employed landmark detection tools (DLIB and MediaPipe) to extract six key feature points around the lips and then computed pairwise distances between these points as input to the CNN model. In our comparison, we refer to the results of the model developed by Ishmam et al. based on the Dlib landmark detector.

At present, comparisons with recent studies are not fully standardized because all datasets were self-collected. The main reason is that there is currently no publicly available benchmark dataset specifically designed for monitoring mouth-opening and closing states in patients. Some existing facial expression datasets, such as emotion recognition datasets (e.g., CK+, MUGFE, or AFEW), can be partially repurposed to construct mouth-state datasets. However, the mapping between emotional expressions and mouth-opening states is not always consistent, and therefore, the conversion rate cannot reach 100% accuracy. For example, in the study by Ji et al. (2018) [33], only a small subset of the FERET database (418 images) was selected for analysis. Consequently, comparisons with prior work can only serve as qualitative references, as the evaluations were performed on different datasets. However, the results summarized in Table 10 indicate that the proposed method achieved high and competitive performance in identifying mouth-opening states. This demonstrates that the present study provides a promising tool for remotely monitoring patients with conditions involving impaired oral motor control.

Table 10: Our method in comparison with recent ML algorithms in mouth state detection

Method	Accuracy	Precision	Recall	F1 score
Ji et al. (2018)	0.9656	N/A	N/A	N/A
Ishman et al. (2022)	0.9525	N/A	N/A	N/A
BiGRU-CLPSO	0.9657	0.9825	0.9492	0.9655

4.8 Discussion and Limitations

The experimental results provide several important implications for practical model deployment. First, GRU and LSTM not only achieved superior performance on both the training and test datasets but also maintained high stability, making them the most suitable candidates for implementation in real-time monitoring systems or embedded applications. In particular, GRU, with its more compact architecture, offers advantages in inference speed and resource efficiency, making it ideal for hardware-constrained environments such as embedded cameras or mobile devices. Second, although XGBoost exhibited signs of overfitting during training, it achieved very high accuracy and precision on the testing set. This suggests that XGBoost may serve as an effective alternative in scenarios with limited data or when rapid deployment is needed, offering lower computational cost compared to deep learning models. Finally, while CNN and MLP were not the optimal choices in this study, their results indicate potential for improvement through techniques such as data augmentation, architectural optimization, or integration into hybrid models. These enhancements could help leverage the simplicity and fast training capability of CNN and MLP in practical applications where short development cycles are required. In establishing the baseline for this study, we focused on learning based approaches (CNN, LSTM, MLP) rather than static geometric thresholds like the MAR. As noted in previous studies [6], fixed-threshold metrics are susceptible to variations in head pose and perspective, which are inherent in our multi-view dataset (frontal, tilted, and rotated orientations). Therefore, we prioritized the BiGRU architecture to automatically learn robust, pose-invariant features from the temporal landmark sequences, ensuring stable performance across the simulated head movements described in Section 4.1.

It is also worth noting why recent heavy architectures, such as Transformer-based models, were not prioritized in this specific comparative analysis. While Transformers have demonstrated state-of-the-art performance in video understanding, they typically demand substantial computational resources due to the quadratic complexity of self-attention mechanisms. In the context of this research, our primary objective was to develop an ultra-lightweight solution capable of operating on CPU-based edge devices (Raspberry Pi) with minimal latency. Furthermore, our input data consists of low-dimensional geometric vectors rather than high-dimensional raw pixel data. Applying complex Transformer encoders to such compact feature sets often results in computational redundancy without proportional gains in accuracy, particularly on smaller datasets where Transformers are prone to overfitting. Therefore, the BiGRU-CLPSO architecture was selected as the optimal balance between data efficiency, inference speed, and predictive accuracy for this specific edge-health monitoring orientation.

The proposed system demonstrates strong potential for real-world healthcare applications. In remote patient monitoring systems, accurate detection of mouth dynamics can provide valuable support for diagnosing neuromuscular disorders such as dysphagia, Bell's palsy, and Parkinson's disease [35]. In addition, it can assist in evaluating fatigue, drowsiness, and respiratory abnormalities, particularly in elderly or bedridden individuals. The high recall achieved by the BiGRU–CLPSO model indicates its suitability for early-warning tasks, where minimizing false negatives is critical. The system can be integrated into edge devices such as Raspberry Pi modules for on-device inference, thereby reducing transmission latency and protecting data privacy. Furthermore, the proposed system is highly convenient for expanding training data sources, facilitating continuous improvement of model performance, and enabling remote updates to maintain real-time operation of the system.

In addition, the developed system is designed to facilitate future data expansion through hardware integration with edge devices, enabling scalable and real-time data acquisition. This feature distinguishes our work from several previous studies that rely on specialized equipment, trained medical staff, and high-cost procedures such as VFSS and FEES [15,16], highlighting the practical potential of our approach. However,

strict interpretation of these results requires acknowledging certain limitations. First, the dataset used in this study is relatively small, involving 5 healthy participants with approximately 3100 samples, and currently lacks diversity in terms of age, ethnicity, and pathological conditions. Consequently, this work represents a pre-clinical pilot investigation aimed at demonstrating the algorithmic effectiveness of the BiGRU–CLPSO model on geometric inputs, rather than a large-scale clinical assessment. It is worth noting that, despite the limited sample size, data collection was conducted under rigorous, well-controlled conditions, where each participant performed mouth-opening actions in multiple fundamental orientations (frontal, right-tilted, left-tilted, upward, and downward) to ensure initial algorithmic robustness. Future work will focus on validating the system in uncontrolled environments and expanding the dataset to include patients with actual neuromuscular disorders to ensure broader clinical generalization.

5 Conclusion

This study proposes a BiGRU–CLPSO model for detecting the mouth-open and mouth-close states of patients, which is suitable for deployment in real-time systems on edge devices. The proposed model was also compared with traditional algorithms, including CNN, LSTM, MLP, and XGBoost. Through experiments involving cross-validation on the training dataset and independent testing on a held-out dataset, the models were thoroughly analyzed using several evaluation metrics, including Accuracy, Precision, Recall, F1-score, Cosine Similarity, ROC–AUC, and the Precision–Recall Curve.

The experimental results show that the proposed BiGRU–CLPSO model consistently achieved the highest performance, demonstrating an optimal balance between accuracy and coverage while maintaining stability across both training and testing datasets. BiGRU–CLPSO was particularly outstanding due to its compact structure and strong generalization ability, making it the optimal choice for real-time applications and edge deployment. Although more complex, LSTM maintained competitive performance and served as a reliable alternative. XGBoost also performed well, especially in terms of precision, though it exhibited a clear tendency toward overfitting during training. CNN and MLP showed noticeably lower performance. CNN favored precision but lacked recall, while MLP produced the weakest and least stable results.

Finally, the proposed model can be effectively integrated into real-time patient monitoring systems using cameras or edge devices. Such systems have the potential to enhance the processes of monitoring, diagnosis, and treatment of disorders related to the oral muscles and the corresponding neural mechanisms.

For future work, we plan to enrich and diversify the data sources to expand the dataset. In addition, more comprehensive analyses will be conducted to further refine and optimize the current system, with the goal of improving its accuracy and responsiveness to the demands of real-time applications. Specifically, we plan to enhance the interpretability of the proposed model using Explainable Artificial Intelligence (XAI) techniques to better understand the feature contributions and decision-making process. This will help identify which facial regions contribute most to mouth-state classification. Furthermore, real-time latency assessment will be conducted on edge devices to validate the model's suitability for continuous monitoring scenarios. Finally, ethical considerations such as informed consent, data anonymization, and bias mitigation across demographic groups will be integrated into the future deployment framework.

Acknowledgement: We would like to thank the National Science and Technology Council, Taiwan, for funding this research.

Funding Statement: This research was partly supported by the National Science and Technology Council, Taiwan, with grant numbers NSTC 114-2622-8-992-007-TD1 and 112-2811-E-992-003-MY3.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Thanh-Lam Nguyen and Mong-Fong Horng; data collection: Thanh-Lam Nguyen; analysis and interpretation of results: Thanh-Tuan Nguyen and Chin-Shiuh Shieh; draft manuscript preparation: Thanh-Tuan Nguyen and Thanh-Lam Nguyen; supervision: Chin-Shiuh Shieh and Mong-Fong Horng; validation: Chen-Fu Hung and Lan-Yuen Guo; critical review: Chun-Chih Lo. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the Corresponding Author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Doan TN, Ho WC, Wang LH, Chang FC, Nhu NT, Chou LW. Prevalence and methods for assessment of oropharyngeal dysphagia in older adults: a systematic review and meta-analysis. J Clin Med. 2022;11(9):2605. doi:10.3390/jcm11092605.
2. González-Fernández M, Ottenstein L, Atanelov L, Christian AB. Dysphagia after stroke: an overview. Curr Phys Med Rehabil Rep. 2013;1(3):187–96. doi:10.1007/s40141-013-0017-y.
3. Moritoyo R, Nakagawa K, Yoshimi K, Yamaguchi K, Nagasawa Y, Yanagida R, et al. Effects of telemedicine on dysphagia rehabilitation in patients requiring home care: a retrospective study. Dysphagia. 2025;40(6):1468–78. doi:10.1007/s00455-025-10844-0.
4. Yang W, Du Y, Chen M, Li S, Zhang F, Yu P, et al. Effectiveness of home-based telerehabilitation interventions for dysphagia in patients with head and neck cancer: systematic review. J Med Internet Res. 2023;25(6):e47324. doi:10.2196/47324.
5. Meuser T, Lovén L, Bhuyan M, Patil SG, Dustdar S, Aral A, et al. Revisiting edge AI: opportunities and challenges. IEEE Internet Comput. 2024;28(4):49–59. doi:10.1109/mic.2024.3383758.
6. Albadawi Y, Takruri M, Awad M. A review of recent developments in driver drowsiness detection systems. Sensors. 2022;22(5):2069. doi:10.3390/s22052069.
7. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014 Oct 25–29; Doha, Qatar. Stroudsburg, PA, USA: ACL; p. 1724–34. doi:10.3115/v1/d14-1179.
8. Zhang YE, Song X. A multi-strategy adaptive comprehensive learning PSO algorithm and its application. Entropy. 2022;24(7):890. doi:10.3390/e24070890.
9. Karunaratne TB, Clavé P, Ortega O. Complications of oropharyngeal dysphagia in older individuals and patients with neurological disorders: insights from Mataró hospital, Catalonia, Spain. Front Neurol. 2024;15:1355199. doi:10.3389/fneur.2024.1355199.
10. Florez R, Palomino-Quispe F, Alvarez AB, Coaquira-Castillo RJ, Herrera-Levano JC. A real-time embedded system for driver drowsiness detection based on visual analysis of the eyes and mouth using convolutional neural network and mouth aspect ratio. Sensors. 2024;24(19):6261. doi:10.3390/s24196261.
11. Majeed F, Shafique U, Safran M, Alfarhood S, Ashraf I. Detection of drowsiness among drivers using novel deep convolutional neural network model. Sensors. 2023;23(21):8741. doi:10.3390/s23218741.
12. de Sire A, Marotta N, Agostini F, Drago Ferrante V, Demeco A, Ferrillo M, et al. A telerehabilitation approach to chronic facial paralysis in the COVID-19 pandemic scenario: what role for electromyography assessment? J Pers Med. 2022;12(3):497. doi:10.3390/jpm12030497.
13. Machetanz K, Lins M, Roder C, Naros G, Tatagiba M, Hurth H. Innovative mobile app solution for facial nerve rehabilitation: a usability analysis. Front Digit Health. 2024;6:1471426. doi:10.3389/fdgth.2024.1471426.

14. MacLellan LE, Stepp CE, Fager SK, Mentis M, Boucher AR, Abur D, et al. Evaluating *Camera Mouseas* a computer access system for augmentative and alternative communication in cerebral palsy: a case study. Assist Technol. 2024;36(3):217–23. doi:10.1080/10400435.2023.2242893.
15. Helliwell K, Hughes VJ, Bennion CM, Manning-Stanley A. The use of videofluoroscopy (VFS) and fibreoptic endoscopic evaluation of swallowing (FEES) in the investigation of oropharyngeal dysphagia in stroke patients: a narrative review. Radiography. 2023;29(2):284–90. doi:10.1016/j.radi.2022.12.007.
16. Ferrari de Castro MA, Dedivitis RA, Luongo de Matos L, Baraúna JC, Kowalski LP, de Carvalho Moura K, et al. Endoscopic and videofluoroscopic evaluations of swallowing for dysphagia: a systematic review. Braz J Otorhinolaryngol. 2025;91(5S):101598. doi:10.1016/j.bjorl.2025.101598.
17. Xu Y, Khan TM, Song Y, Meijering E. Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. Artif Intell Rev. 2025;58(3):93. doi:10.1007/s10462-024-11033-5.
18. Singh R, Gill SS. Edge AI: a survey. Internet Things Cyber Phys Syst. 2023;3(5):71–92. doi:10.1016/j.iotcps.2023.02.004.
19. Khalid N, Qayyum A, Bilal M, Al-Fuqaha A, Qadir J. Privacy-preserving artificial intelligence in healthcare: techniques and applications. Comput Biol Med. 2023;158(3):106848. doi:10.1016/j.compbiomed.2023.106848.
20. Sadeghi Z, Alizadehsani R, Cifci MA, Kausar S, Rehman R, Mahanta P, et al. A review of explainable artificial intelligence in healthcare. Comput Electr Eng. 2024;118(5):109370. doi:10.1016/j.compeleceng.2024.109370.
21. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep. 2022;12(1):5979. doi:10.1038/s41598-022-09954-8.
22. Friji R, Chaieb F, Drira H, Kurtek S. Geometric deep neural network using rigid and non-rigid transformations for landmark-based human behavior analysis. IEEE Trans Pattern Anal Mach Intell. 2023;45(11):13314–27. doi:10.1109/TPAMI.2023.3291663.
23. Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding?. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. 2021. p. 813–24.
24. Tan M, Le QV. EfficientNetV2: smaller models and faster training. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 10096–106.
25. Vasu PKA, Gabriel J, Zhu J, Tuzel O, Ranjan A. MobileOne: an improved one millisecond mobile backbone. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023 Jun 17–24; Vancouver, BC, Canada, p. 7907–17. doi:10.1109/CVPR52729.2023.00764.
26. Huang X, Cai Z. A review of video action recognition based on 3D convolution. Comput Electr Eng. 2023;108(1):108713. doi:10.1016/j.compeleceng.2023.108713.
27. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, et al. Video swin transformer. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 3202–11. doi:10.1109/CVPR52688.2022.00320.
28. Zhu T, Zhang C, Wu T, Ouyang Z, Li H, Na X, et al. Research on a real-time driver fatigue detection algorithm based on facial video sequences. Appl Sci. 2022;12(4):2224. doi:10.3390/app12042224.
29. Wan J, Liu J, Zhou J, Lai Z, Shen L, Sun H, et al. Precise facial landmark detection by reference heatmap transformer. IEEE Trans Image Process. 2023;32(3):1966–77. doi:10.1109/tip.2023.3261749.
30. Lu Y, Li K. Research on lip recognition algorithm based on MobileNet + attention-GRU. Math Biosci Eng. 2022;19(12):13526–40. doi:10.3934/mbe.2022631.
31. Easwaran S, Venugopal JP, Subramanian AAV, Sundaram G, Naseeba B. A comprehensive learning based swarm optimization approach for feature selection in gene expression data. Heliyon. 2024;10(17):e37165. doi:10.1016/j.heliyon.2024.e37165.
32. Liang JJ, Qin AK, Suganthan PN, Baskar S. Comprehensive learning particle swarm optimizer for global optimization of multimodal functions. IEEE Trans Evol Computat. 2006;10(3):281–95. doi:10.1109/tevc.2005.857610.
33. Ji Y, Wang S, Lu Y, Wei J, Zhao Y. Eye and mouth state detection algorithm based on contour feature extraction. J Electron Imag. 2018;27(5):051205. doi:10.1117/1.jei.27.5.051205.

34. Ishmam A, Hasan M, Hassan Onim MS, Roy K, Hoque Akif MA, Nyeem H. Modelling lips state detection using CNN for non-verbal communications. In: Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021. Singapore: Springer Nature; 2022. p. 59–70. doi:10.1007/978-981-19-2445-3_5.
35. Hung CF, Hsu M, Liu HY, Horng MF, Yang CC, Guo LY. Development of a non-contact screening approach for identifying oral function high-risk older adults using jaw movement and diadochokinetic performance. BMC Oral Health. 2025;25(1):1568. doi:10.1186/s12903-025-06725-5.