Journal of Computer Science and Cybernetics, V.42, N.2 (2026), 183–197
DOI: 10.15625/1813-9663/24199

DIGITAL FAST: AN AI-DRIVEN MULTIMODAL FRAMEWORK FOR RAPID AND EARLY STROKE SCREENING

NGOC-KHAI HOANG[1], THI-NHU-MAI-NGUYEN[2], HUY-HIEU PHAM[1,3,*]

[1] *VinUni-Illinois Smart Health Center, VinUniversity, Vinhomes Ocean Park, Gia Lam Ward, Ha Noi, Viet Nam*
[2] *Study Program Medical Informatics, Universität zu Lübeck, Germany*
[3] *College of Engineering and Computer Science, VinUniversity, Vinhomes Ocean Park, Gia Lam Ward, Ha Noi, Viet Nam*

Abstract. Early identification of stroke symptoms is essential for enabling timely intervention and improving patient outcomes, particularly in prehospital settings. This study presents a fast, non-invasive multimodal deep learning framework for automatic binary stroke screening based on data collected during the F.A.S.T. assessment. The proposed approach integrates complementary information from facial expressions, speech signals, and upper-body movements to enhance diagnostic robustness. Facial dynamics are represented using landmark based features and modeled with a Transformer architecture to capture temporal dependencies. Speech signals are converted into mel spectrograms and processed using an Audio Spectrogram Transformer, while upper-body pose sequences are analyzed with an MLP-Mixer network to model spatiotemporal motion patterns. The extracted modality specific representations are combined through an attention-based fusion mechanism to effectively learn cross modal interactions. Experiments conducted on a self-collected dataset of 222 videos from 37 subjects demonstrate that the proposed multimodal model consistently outperforms unimodal baselines, achieving 95.83% accuracy and a 96.00% F1-score. The model attains a strong balance between sensitivity and specificity and successfully detects all stroke cases in the test set. These results highlight the potential of multimodal learning and transfer learning for early stroke screening, while emphasizing the need for larger, clinically representative datasets to support reliable real-world deployment.

Keywords. Stroke detection, multimodal learning, pretrained model, attention fusion.

1. INTRODUCTION

Stroke refers to a group of cerebrovascular disorders caused by blocked or ruptured blood vessels supplying the brain. Among these conditions, acute ischemic stroke is one of the leading causes of long-term disability and mortality worldwide [1, 2], with epidemiological studies estimating that approximately one in four individuals will experience a stroke during their lifetime [3, 4, 5]. Consequently, timely and appropriate intervention is of critical importance [6], and early detection is essential to enable prompt treatment.

*Corresponding author.
E-mail address: hoangngockhai000@gmail.com (N.K. Hoang), thi01.nguyen@student.uni-luebeck.de (T.N.M. Nguyen), hieu.ph@vinuni.edu.vn (H.H. Pham).

Notably, when individuals exhibit early signs of stroke, bystanders often lack sufficient medical knowledge to recognize these symptoms as accurately as healthcare professionals [7]. This gap highlights the urgent need for automated methods capable of supporting stroke risk detection in real-world settings, particularly for non-expert users.

In the context of rapid digitalization, artificial intelligence (AI) has significantly impacted healthcare applications [8], including automated healthcare systems, clinical decision support, and medical diagnosis, where AI has, in certain cases, achieved performance superior to that of human specialists [9, 10, 11]. Multimodal learning approaches, which integrate information from multiple data sources, have demonstrated superior performance compared to models relying on a single modality [12]. However, despite recent advances in multimodal biomedical AI [13], multimodal stroke detection, particularly under subject independent and prehospital settings, has not yet been sufficiently explored [14, 15]. Addressing this gap requires effective multimodal modeling strategies as well as datasets and experimental settings that reflect real-world constraints.

Building upon this motivation, this study aims to advance multimodal stroke screening by jointly exploiting complementary cues from visual, acoustic, and motion based signals collected during the F.A.S.T. assessment [16]. In particular, we focus on designing a practical and non-invasive solution suitable for prehospital scenarios, where rapid assessment, robustness to limited data, and reliable generalization across subjects are critical. Our main contributions are summarized as follows:

- We construct a self-collected multimodal dataset for prehospital stroke screening, consisting of temporally synchronized facial videos, speech recordings, and upper-body pose data acquired during the F.A.S.T. assessment.
- We develop a multimodal deep learning framework that integrates Transformer-based encoders for facial and speech modalities with an MLP-Mixer architecture for pose representation, combined through an attention-based fusion mechanism to capture cross-modal dependencies.
- We conduct extensive experimental evaluations and ablation studies under subject independent settings, demonstrating that the proposed multimodal approach outperforms unimodal and pairwise fusion baselines while maintaining a robust balance between sensitivity and specificity.

The remainder of this paper is organized as follows. Section 2 reviews related work on stroke detection and multimodal deep learning approaches. Section 3 presents the proposed multimodal framework, including the modality specific encoders and the fusion strategy. Section 4 describes the dataset and data acquisition process in detail. Section 5 reports the experimental setup and quantitative performance results. Section 6 provides ablation studies to analyze the contribution of individual modalities and fusion strategies. Section 7 discusses the experimental findings, limitations, and practical implications of the proposed method. Finally, Section 8 concludes the paper and outlines directions for future work.

2. RELATED WORK

Recent advances in artificial intelligence have led to increasing interest in applying machine learning and deep learning techniques to stroke diagnosis and prognosis. A compre-

hensive systematic review by Shurrab et al. [14] analyzed multimodal machine learning approaches for stroke-related tasks, showing that most existing studies focus on neuroimaging modalities, such as computed tomography (CT) and magnetic resonance imaging (MRI), as well as electronic health records and clinical data for stroke detection, tissue segmentation, risk estimation, and outcome prediction. Despite the growing popularity of multimodal learning, the review also highlighted the limited availability of diverse multimodal datasets and the need for methods tailored to practical clinical and prehospital scenarios.

Beyond imaging-based approaches, several studies have investigated non-imaging indicators relevant to stroke detection, including facial abnormalities, speech impairments, and motor dysfunction. Wang et al. [17] proposed an ensemble convolutional neural network for detecting acute ischemic stroke from two-dimensional facial images, demonstrating the feasibility of facial image-based stroke screening before CT examination. To address privacy concerns associated with raw facial videos, Cai et al. [18] introduced SafeTriage, a privacy-preserving stroke triage framework that de-identifies facial videos while retaining diagnostically relevant motion patterns. Speech impairments, particularly dysarthria, have also been explored as neurological indicators. Mahum et al. [19] proposed a Swin Transformer-based framework for dysarthria recognition using mel spectrogram representations, while Yang et al. [20] introduced a contrastive learning-based feature extraction method for distinguishing healthy and impaired speech.

More recently, multimodal approaches combining visual, speech, and movement information have been investigated for stroke triage in acute and prehospital settings. Cai et al. [21] proposed M3 Stroke, a mobile multimodal AI system integrating audio-visual data for emergency triage of mild to moderate stroke patients. Similarly, Ou et al. [22] developed a multimodal deep learning framework based on the F.A.S.T. assessment, incorporating facial expressions, speech recordings, and limb movement data collected from emergency room patients. Yu et al. [23] proposed a multimodal framework for rapid stroke diagnosis by emulating clinical screening protocols such as the Cincinnati Prehospital Stroke Scale (CPSS) and the Face Arm Speech Test (FAST). Building upon this direction, Cai et al. [24] introduced DeepStroke, a multimodal adversarial deep learning framework that jointly analyzes facial muscle coordination and speech impairments for emergency room stroke screening.

Although these studies demonstrate the effectiveness of multimodal learning for stroke assessment, most existing systems are designed for controlled clinical environments and often rely on raw facial video data, which may raise privacy concerns and limit their applicability in practical prehospital scenarios. Motivated by these limitations, the present study aims to develop a privacy-aware multimodal deep learning system that integrates facial landmarks, speech signals, and upper-body movement information for early stroke screening in prehospital environments.

3. PROPOSED METHOD

We propose a multimodal deep learning framework that jointly leverages facial expressions, upper-body movements, and speech signals for automated stroke screening in prehospital settings, as illustrated in Fig. 1. The system takes facial videos, upper-body videos, and speech signals collected during the F.A.S.T. assessment as input. These data streams are processed by modality-specific encoders and then integrated at the feature representation level for binary stroke prediction.

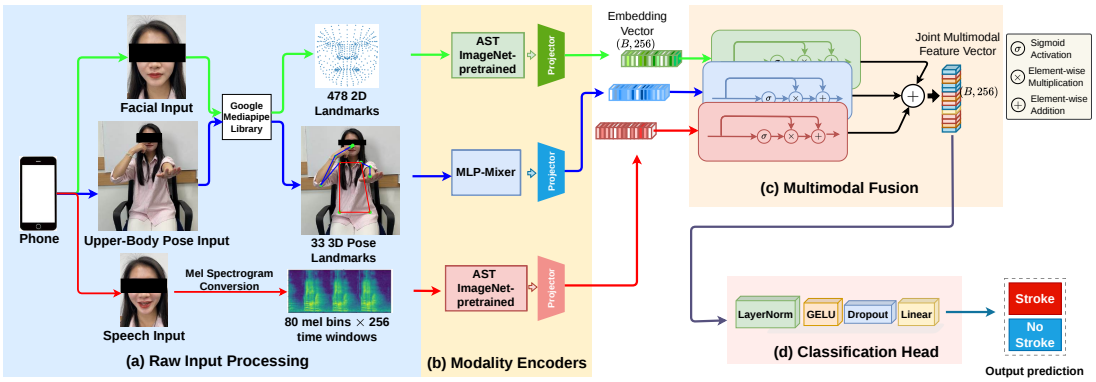

Figure 1: The proposed framework comprises four stages: (a) Raw Input Processing, which transforms raw visual and audio data into structured representations; (b) Modality Encoders, which extract high-level embeddings using modality-specific models; (c) Multimodal Fusion, which integrates the embeddings through a learnable fusion mechanism; and (d) Classification Head, which produces the final binary stroke prediction.

3.1. Raw input processing

As illustrated in Fig. 1(a), the system processes three types of raw inputs: facial videos, upper-body videos, and speech signals. In the visual branch, facial and upper-body videos are processed using Google MediaPipe to extract landmark representations. Facial landmarks are used to capture facial asymmetry, while upper-body pose landmarks represent motion patterns related to arm weakness and coordination deficits.

Landmarks are extracted frame by frame and organized into temporal sequences. Each frame is treated as a token in the sequence. For the face modality, each token is a flattened 2D landmark vector of dimension $D_f = 478 \times 2 = 956$; for the pose modality, each token is a flattened 3D vector of dimension $D_p = 33 \times 3 = 99$. A linear projection maps each token into the encoder embedding space, and positional embeddings are added to preserve temporal order. Sequences are padded or center-cropped along the temporal dimension to ensure consistent tensor shapes.

For the auditory modality, speech signals are converted into mel-spectrogram representations using a short-time Fourier transform followed by mel-scale mapping. These spectrograms are used as the input to the speech encoder.

3.2. Modality encoders

As shown in Fig. 1(b), each modality is processed using a dedicated encoder. The facial landmark sequences are encoded using a Transformer-based encoder initialized from an Audio Spectrogram Transformer (AST) backbone to capture temporal dependencies in facial motion. The upper-body pose sequences are processed using an MLP-Mixer architecture, which models spatial and temporal interactions among pose landmarks with relatively low computational complexity. The speech branch uses a pretrained AST model to extract high-level acoustic features from mel-spectrogram inputs.

The outputs of all modality encoders are mapped through projector layers to a shared embedding space with a unified dimensionality of 256. Formally, let X_f, X_p, and X_s denote the input sequences for the face, pose, and speech modalities, respectively. The modality-specific representations are computed as

$$H_f = \text{AST}(X_f),\ H_p = \text{MLPMixer}(X_p),\ H_s = \text{AST}(X_s). \tag{1}$$

The extracted features are then projected into a shared embedding space

$$z_i = \text{Projector}(H_i),\ i \in \{f, p, s\}, \tag{2}$$

where z_f, z_p, and z_s denote the projected embeddings of the face, pose, and speech modalities, respectively.

3.3. Multimodal fusion

After obtaining the three modality embeddings, the system performs multimodal fusion using a sigmoid-based gating mechanism, as illustrated in Fig. 1(c). The gating operation adaptively enhances informative modality-specific features while suppressing less relevant signals

$$z'_f = z_f \odot \sigma(z_f) + z_f,\ z'_p = z_p \odot \sigma(z_p) + z_p,\ z'_s = z_s \odot \sigma(z_s) + z_s, \tag{3}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function and $\odot$ represents element-wise multiplication.

The final multimodal representation is obtained by aggregating the gated embeddings

$$z' = z'_f + z'_p + z'_s, \tag{4}$$

where z' denotes the fused multimodal feature vector used for downstream classification. This fusion mechanism allows the model to adaptively combine complementary cues from facial motion, body movement, and speech, improving robustness when the reliability of individual modalities varies across samples.

3.4. Classification head

Finally, as shown in Fig. 1(d), the fused representation is passed through a lightweight classification head consisting of Layer Normalization, GELU activation, Dropout, and a Linear layer followed by sigmoid activation

$$\hat{y} = \text{Sigmoid}\left(\text{Linear}\left(\text{Dropout}\left(\text{GELU}\left(\text{LayerNorm}(z')\right)\right)\right)\right). \tag{5}$$

The output $\hat{y}$ represents the predicted probability of the stroke class.

4. DATASET

4.1. Data acquisition

Data collection was conducted using the rear camera of an iPhone 12 mounted on a simple tripod, positioned at a height aligned with the subject's face while seated. An overview of the recording setup and task-specific framing are illustrated in Fig. 2. To ensure consistency and minimize environmental noise, all recordings were performed in a quiet indoor environment against a neutral background. Videos were captured in vertical orientation at a resolution of 1080p and a frame rate of 30 fps, with High Dynamic Range (HDR) disabled and mono audio recording enabled.

For the facial expression and speech tasks, the camera was positioned sufficiently close such that the face occupied approximately 80% of the video frame, enabling detailed capture of facial movements. During the upper-body task, the camera was placed at a greater distance to ensure that both arms and hand movements remained fully visible throughout the recording, while minimizing excessive background space.

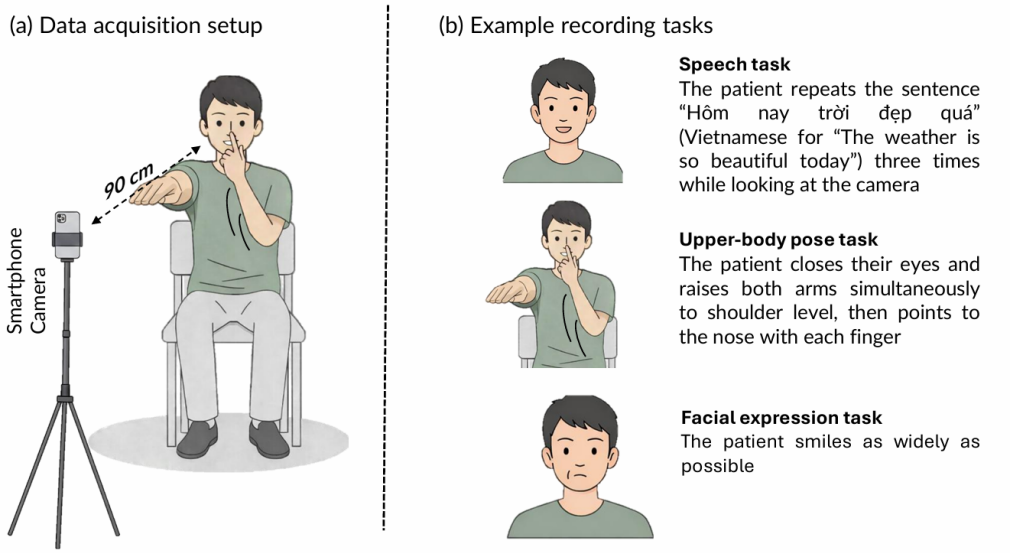

Figure 2: (a) Standardized data acquisition setup using a rear iPhone 12 camera mounted on a tripod at face height while the subject was seated. (b) Recording tasks including speech, upper-body pose, and facial expression assessment.

The dataset comprises recordings collected from 37 participants under a controlled pre-hospital setting. We define one multimodal sample (trial) as a triplet of recordings (Face, Speech, and Pose) captured for the same repetition. Each participant performed 6 repetitions (3 healthy + 3 simulated stroke), resulting in 6 multimodal samples per participant and 222 multimodal samples in total. Each multimodal sample contains three modality-specific videos, yielding 666 raw video clips (222×3) before quality filtering. All participants performed a predefined set of tasks based on the F.A.S.T. assessment protocol. Detailed demographic information and data acquisition statistics are summarized in Table 1.

As summarized in Table 2, each subject was instructed to perform three tasks corresponding to the components of the F.A.S.T. assessment. For the facial task, subjects were asked to smile as widely as possible under normal conditions. To simulate facial paralysis, they were instructed to smile using only one side of the face.

For the speech task, participants were required to repeat the sentence "Hôm nay trời đẹp quá", which translates to "Today is such a nice day". This sentence was selected due to its inclusion of multiple plosive phonemes, which are known to be particularly affected by dysarthria - a common speech impairment caused by disrupted neural control of oral muscles in stroke patients [25]. When simulating stroke symptoms, subjects were asked to articulate the sentence more slowly and unclearly.

For the upper-body task, participants were instructed to close their eyes, raise both arms to shoulder height, and subsequently touch the tip of their nose with each arm in turn. Although the original F.A.S.T. test does not include the nose touching component, this task was incorporated to assess motor coordination abilities, following the protocol proposed by Rodrigues et al. [26]. To simulate stroke related motor deficits, subjects were instructed to mimic unilateral arm weakness, accompanied by intentional hand tremors and imprecise nose touching movements.

Table 1: Participant demographics and dataset statistics

Category	Description
Participants	
Total participants	37
Female / Male	19 (51.35%) / 18 (48.65%)
Age (18–30) / (30–40)	35 / 2
Protocol and modalities	
Assessment protocol	F.A.S.T.-inspired tasks
Task types	3 (Face, Speech, Arm movement)
Modalities per sample	3 (Face, Speech, Pose)
Recording counts	
Repetitions per condition (per participant)	3 healthy + 3 simulated stroke
Total repetitions (per participant)	6
Multimodal samples (per participant)	6
Total multimodal samples	222
Raw video clips (per participant)	18 (6 samples × 3 modalities)
Total raw video clips	666 (222 samples × 3 modalities)

4.2. Data preprocessing

In this study, facial landmarks were utilized to capture positional cues associated with stroke related symptoms, including facial drooping and asymmetrical mouth movements, while preserving participant privacy because raw video frames were not processed directly. Facial landmark representations have also been explored in other facial behavior analysis tasks, including personality assessment from interview videos, demonstrating their ability to capture discriminative facial movement patterns without relying directly on raw facial appearance [27]. Facial landmarks were extracted on a frame-by-frame basis using Google's MediaPipe library [28] with the Face Landmarker model, which outputs 478 landmarks defined by three dimensional coordinates. To specifically capture horizontal and vertical asymmetries, only the x and y coordinates were retained.

Beyond facial features, body landmarks were extracted to characterize upper-body motion and skeletal dynamics, as body landmark representations have been shown to effectively capture motion patterns in video-based analysis of neurological conditions [29]. The MediaPipe Pose Landmarker [30] model was employed, providing 33 landmarks with x, y, and z coordinates. In contrast to facial landmarks, all three axes were preserved for pose representation to retain depth information, which is important for movements involving spatial displacement, such as finger to nose gestures. Collectively, these preprocessing pipelines transform each video into a temporal sequence of landmark coordinates suitable for sequential modeling.

Audio signals were represented using mel spectrograms, which effectively capture time-frequency characteristics relevant to speech analysis [31]. Each audio recording was first downsampled to 16 kHz and transformed using the Short-Time Fourier Transform, followed by mel-scale mapping. To ensure uniform input dimensions, all mel spectrograms were generated with 80 mel bins and temporally padded or truncated to a fixed length of 256 frames.

Table 2: Tasks used in the F.A.S.T. assessment and their corresponding stroke-related impairments.

Task	Expected behavior of stroke-affected patients	Physical / vocal impairment
The patient smiles as widely as possible	Facial movement is uneven between the two sides of the face	Facial drooping
The patient repeats the sentence "Hôm nay trời đẹp quá" (Vietnamese for "The weather is so beautiful today") three times while looking at the camera	Speech becomes slowed and pronunciation is unclear	Slurred speech
The patient closes their eyes and raises both arms simultaneously to shoulder level, then points to the nose with each finger	One arm does not lift as quickly or as high as the other. When pointing to the nose, the finger drifts slightly and may not reach the nose on the first attempt	Arm weakness and coordination problems

5. EXPERIMENTS

5.1. Implementation details

The proposed system consists of multiple modality-specific branches designed to process heterogeneous input signals. Facial landmark sequences were modeled using a Transformer encoder initialized from a pretrained Audio Spectrogram Transformer (AST) backbone with a base configuration. Speech signals were processed using an AST model with a tiny configuration to extract discriminative acoustic representations from mel spectrograms. For upper-body pose data, an MLP-Mixer architecture was employed and trained from scratch to capture temporal and spatial interactions among pose landmarks.

All models were implemented in PyTorch and trained for 200 epochs using the Adam optimizer with the Binary Cross-Entropy loss function. A decision threshold of 0.5 was applied to classify samples as stroke or non-stroke. The experiments were conducted on an NVIDIA GeForce RTX 3090 GPU.

After data cleaning, the dataset consisted of 222 valid video samples and was split in a subject-independent manner into training, validation, and test sets. Specifically, 29 subjects (116 videos) were used for training, while 4 subjects (24 videos) were allocated to each of the validation and test sets, ensuring no subject overlap across splits. Model performance was evaluated using accuracy, Area Under the Receiver Operating Characteristic Curve (AUC), and the balanced F1-score. Table 3 summarizes the main model configurations and training hyperparameters used in our experiments.

5.2. Computational efficiency and inference cost

To address practical deployment considerations, we report the model size and inference efficiency of the proposed multimodal framework. Table 4 summarizes the parameter breakdown across modality-specific encoders and the fusion head. The overall model contains

Table 3: Model configuration and training hyperparameters

Parameter	Value
Face encoder	AST (base)
Speech encoder	AST (tiny)
Pose encoder	MLP-Mixer
Embedding dimension	256
Optimizer	Adam
Learning rate	1×10^{-4}
Batch size	8
Training epochs	200
Loss function	Binary Cross-Entropy
Decision threshold	0.5
Framework	PyTorch
GPU	NVIDIA RTX 3090

Table 4: Efficiency summary of the proposed model

Item	Value
Face encoder params	85.791 M
Audio encoder params	5.809 M
Pose encoder params	0.643 M
Fusion + head params	0.247 M
Total params	92.490 M
Model size (params only)	352.82 MB
Inference time	7.19 ms / sample

92.490M parameters, corresponding to 352.82 MB when stored in 32-bit floating point. The end-to-end inference latency of the full multimodal pipeline is 7.19 ms per sample, measured with batch size = 4. These results indicate that the proposed architecture can support real-time screening requirements in prehospital scenarios.

5.3. Performance comparison across model variants

The results in Table 5 show a clear performance gap between unimodal and multimodal models. Among unimodal approaches, the pose-only model obtains the lowest performance, with an accuracy of 0.6667 and an F1-score of 0.6364, indicating that upper-body pose alone provides limited discriminative information. The voice-only model improves substantially, achieving an accuracy and F1-score of 0.8333, while the face-only model performs best among unimodal variants, with an accuracy of 0.8750, an AUC of 0.9306, and a specificity of 1.0000. However, its sensitivity remains at 0.7500, suggesting that relying on a single modality may still miss stroke cases.

Pairwise fusion further improves performance, confirming the complementarity among modalities. In particular, the Face + Voice configuration achieves the best pairwise results, with an accuracy of 0.9167, an AUC of 1.0000, and an F1-score of 0.9091. The Late Fusion baseline achieves comparable performance, with an accuracy of 0.9167 and an F1-score of

Table 5: Performance comparison of different models. The best results are highlighted in bold.

Model	Accuracy	AUC	F1-score	Sensitivity	Specificity
Pose only	0.6667	0.8125	0.6364	0.5833	0.7500
Voice only	0.8333	0.8819	0.8333	0.8333	0.8333
Face only	0.8750	0.9306	0.8750	0.7500	1.0000
Face + Pose	0.8750	0.9375	0.8571	0.7500	1.0000
Voice + Pose	0.8750	0.9735	0.8696	0.8333	0.9167
Face + Voice	0.9167	1.0000	0.9091	0.8333	1.0000
Late fusion	0.9167	0.9861	0.9091	0.8333	1.0000
Concat + MLP	0.8750	0.9583	0.8571	0.8333	0.9167
Fusion (Frozen weight)	0.8750	0.9792	0.8571	0.7500	**1.0000**
Fusion (Fine-tuned) (Ours)	**0.9583**	**1.0000**	**0.9600**	**1.0000**	0.9167

0.9091, while the Concat + MLP baseline reaches an accuracy of 0.8750 and an F1-score of 0.8571. These results indicate that combining modalities is beneficial, but simple prediction averaging or feature concatenation may not fully exploit cross-modal relationships.

The best overall performance is achieved by the proposed fusion model with fine-tuned pretrained weights, which obtains an accuracy of 0.9583, a perfect AUC of 1.0000, and an F1-score of 0.9600. Importantly, this model reaches a sensitivity of 1.0000, indicating that all stroke cases in the test set are correctly identified, while maintaining a high specificity of 0.9167. Overall, these findings demonstrate the advantage of fine-tuned multimodal fusion for capturing complementary facial, speech, and motion cues in prehospital stroke screening.

6. ABLATION STUDY

6.1. Performance between different fusion strategies

Table 6 compares different fusion strategies to evaluate their impact within the multimodal framework. Overall, learnable fusion methods outperform simple aggregation approaches, suggesting that modeling inter-modal relationships is beneficial for prehospital stroke screening.

Concat achieves an accuracy of 0.8750 and an F1-score of 0.8696 with an AUC of 0.9792, while Sum attains the same accuracy (0.8750) and a perfect AUC of 1.0000 but lower sensitivity (0.7500), indicating limited flexibility in capturing stroke-related cues. Learnable Weighted Sum improves performance to an accuracy of 0.9167 and an F1-score of 0.9091 while maintaining an AUC of 1.0000.

Attention fusion achieves the best overall results, with an accuracy of 0.9583, an F1-score of 0.9600, and a sensitivity of 1.0000. These findings demonstrate that attention-based fusion more effectively captures cross-modal interactions and improves screening reliability.

6.2. Performance under controlled modality corruptions

To assess robustness under challenging prehospital recording conditions, we conduct a controlled *test-time* corruption study by corrupting each modality independently and in

Table 6: Performance comparison of different fusion strategies

Fusion Strategy	Accuracy	AUC	F1-score	Sensitivity	Specificity
Concat	0.8750	0.9792	0.8696	0.8333	0.9167
Sum	0.8750	1.0000	0.8571	0.7500	**1.0000**
Learnable Weighted Sum	0.9167	1.0000	0.9091	0.8333	**1.0000**
Attention Fusion (Ours)	**0.9583**	**1.0000**	**0.9600**	**1.0000**	0.9167

Table 7: Performance under different corruption settings. A checkmark indicates the modality is corrupted.

Corrupted Modality			Performance				
Face	Pose	Speech	Acc	AUROC	F1	Sen	Spec
			0.9583	1.0000	0.9600	1.0000	0.9167
✓			0.9167	0.9861	0.9231	1.0000	0.8333
	✓		0.9583	1.0000	0.9600	1.0000	0.9167
		✓	0.8750	0.9722	0.8889	1.0000	0.7500
✓	✓		0.9167	0.9861	0.9231	1.0000	0.8333
	✓	✓	0.8750	0.9722	0.8889	1.0000	0.7500
✓		✓	0.8750	0.9583	0.8889	1.0000	0.7500
✓	✓	✓	0.8333	0.9514	0.8571	1.0000	0.6667

combination. For Face and Pose, visual corruptions are applied at the raw video-frame level, including motion perturbation, illumination/color changes, and blur, followed by landmark re-extraction using the same MediaPipe pipeline. For Speech, additive noise is injected into the raw waveform at a fixed SNR of 10dB before mel-spectrogram computation.

As reported in Table 7, the proposed multimodal model preserves perfect sensitivity across all corruption settings (Sen = 1.0000), which is critical for screening-oriented applications. Among single-modality corruptions, Speech corruption causes the largest degradation (Acc = 0.8750, F1 = 0.8889), while Pose corruption has the smallest impact (Acc = 0.9583, F1 = 0.9600). Under the most challenging Face+Pose+Speech corruption, performance decreases to Acc = 0.8333 and F1 = 0.8571, mainly due to reduced specificity (Spec = 0.6667) rather than missed stroke cases.

6.3. Pose branch temporal-segment ablation

To further analyze the pose modality, we conduct a temporal-segment ablation on the upper-body pose landmark sequences. For each sample, we extract a fixed-length window of $T = 200$ frames using three strategies: early (first T frames), center (a centered window), and late (last T frames). All other settings are kept unchanged.

Table 8 shows that the temporal window selection affects the pose-only performance. The center window yields the most balanced results, achieving Acc = 66.67% and F1 = 0.6364 with Sen = 0.5833 and Spec = 0.7500. The early window produces higher specificity (Spec = 0.9167) but substantially lower sensitivity (Sen = 0.2500), indicating more missed positive cases when only the initial segment is used. The late window achieves perfect specificity (Spec

Table 8: Pose-only performance under different temporal window selections (window length $T = 200$ frames).

Window	Acc	AUROC	F1	Sen	Spec
Early	0.5833	0.7778	0.3750	0.2500	0.9167
Center	**0.6667**	**0.8125**	**0.6364**	**0.5833**	**0.7500**
Late	0.6667	0.8333	0.5000	0.3333	1.0000

= 1.0000) but lower sensitivity (Sen = 0.3333), suggesting that while late-stage movements can be discriminative, relying only on the final segment is insufficient for consistent positive detection. Overall, these results indicate that coordination cues in the nose-touching task are not uniformly distributed over time, and the pose modality is sensitive to temporal cropping.

7. DISCUSSION

Overall, the experimental results highlight the potential of multimodal learning combined with transfer learning for prehospital stroke screening under limited-data conditions. The fine-tuned fusion model achieves a sensitivity of 1.0000, which is particularly important for screening applications where missed stroke cases should be minimized. At the same time, the model maintains high specificity, suggesting its potential as a screening-oriented decision-support prototype.

Despite these promising results, several limitations remain. First, the dataset is relatively small, consisting of 222 samples from a limited number of participants, which may limit generalization to more diverse prehospital scenarios and increase the risk of overfitting. Second, the data were collected from simulated stroke symptoms rather than real stroke patients; therefore, the current study should be regarded as a proof-of-concept rather than a clinically validated deployment-ready system. Third, the cohort is skewed toward younger participants, with 35/37 participants aged 18–30, whereas stroke incidence is higher in older populations. Future work will focus on collecting larger and more clinically representative datasets, including older participants and real stroke patients, and evaluating the proposed framework on independent external datasets.

8. CONCLUSION

In this study, we propose a multimodal learning based approach for prehospital stroke screening that jointly exploits information from facial appearance, speech, and upper-body movements acquired during the F.A.S.T. assessment. The system is designed to be non-invasive, easy to deploy, and well suited to prehospital scenarios, where early detection and rapid decision support are critical.

Experimental results demonstrate that the proposed multimodal model consistently outperforms unimodal baselines across most evaluation metrics, particularly in achieving a favorable balance between sensitivity and specificity. Leveraging pretrained models in combination with an attention-based fusion strategy substantially improves overall performance while maintaining training stability under limited data conditions. Notably, the proposed method achieves complete detection of stroke cases in the test set, highlighting its potential

for practical deployment in early stroke screening.

Despite limitations related to dataset size and representativeness, the obtained results confirm the feasibility of applying multimodal deep learning to prehospital stroke screening. Future work will focus on expanding the dataset with recordings from real stroke patients, diversifying data acquisition scenarios, and evaluating the proposed system on independent datasets to further enhance its generalization capability and reliability.

ACKNOWLEDGMENT

The authors would like to sincerely thank the Vietnam Young Talent Support Fund, Tan Hiep Phat Trading – Service Co., Ltd., and the Ben Dam Me Award Fund for their valuable support and encouragement of this work.

REFERENCES

[1] S. C. Johnston, S. Mendis, and C. D. Mathers, "Global variation in stroke burden and mortality: estimates from monitoring, surveillance, and modelling," *The Lancet Neurology*, vol. 8, no. 4, pp. 345–354, 2009.

[2] World Health Organization, "Who guidelines for indoor air quality: Household fuel combustion," World Health Organization, Geneva, Switzerland, Tech. Rep., 2014. [Online]. Available: https://iris.who.int/handle/10665/141496

[3] V. L. Feigin, M. Brainin, B. Norrving, S. O. Martins, J. Pandian, P. Lindsay, M. F Grupper, and I. Rautalin, "World stroke organization: global stroke fact sheet 2025," *International Journal of Stroke*, vol. 20, no. 2, pp. 132–144, 2025.

[4] V. L. Feigin, B. A. Stark, C. O. Johnson, and et al., "Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the global burden of disease study 2019," *The Lancet Neurology*, vol. 20, no. 10, pp. 795–820, 2021.

[5] C. O. Johnson, M. Nguyen, G. A. Roth, E. Nichols, T. Alam, D. Abate, F. Abd-Allah, A. Abdelalim, H. N. Abraha, N. M. Abu-Rmeileh, and et al., "Global, regional, and national burden of stroke, 1990–2016: a systematic analysis for the global burden of disease study 2016," *The Lancet Neurology*, vol. 18, no. 5, pp. 439–458, 2019.

[6] J. L. Saver, "Time is brain - quantified," *Stroke*, vol. 37, no. 1, pp. 263–266, 2006.

[7] J. Harbison, O. Hossain, D. Jenkinson, J. Davis, S. J. Louw, and G. A. Ford, "Diagnostic accuracy of stroke referrals from primary care, emergency room physicians, and ambulance staff using the face arm speech test," *Stroke*, vol. 34, no. 1, pp. 71–76, 2003.

[8] G. Litjens, F. Ciompi, J. M. Wolterink, B. D. de Vos, T. Leiner, J. Teuwen, and I. Išgum, "State-of-the-art deep learning in cardiovascular image analysis," *JACC: Cardiovascular imaging*, vol. 12, no. 8 Part 1, pp. 1549–1565, 2019.

[9] S. A. Alowais, S. S. Alghamdi, N. Alsuhebany, T. Alqahtani, A. I. Alshaya, S. N. Almohareb, A. Aldairem, M. Alrashed, K. B. Saleh, H. A. Badreldin, and et al., "Revolutionizing healthcare: the role of artificial intelligence in clinical practice," *BMC Medical Education*, vol. 23, no. 1, p. 689, 2023.

[10] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[11] D. P. Thanh, D. N. Van, T. T. Tan, and V. A. Nguyen, "Spatio-temporal graph learning with epidemiological factors for hiv epidemic short-term prediction," *Journal of Computer Science and Cybernetics*, vol. 40, no. 4, pp. 363–380, 2024.

[12] A. Benani, S. Ohayon, F. Laleye, P. Bauvin, E. Messas, S. Bodard, and X. Tannier, "Is multimodal better? a systematic review of multimodal versus unimodal machine learning in clinical decision-making," *medRxiv*, pp. 2025–03, 2025.

[13] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical AI," *Nature Medicine*, vol. 28, no. 9, pp. 1773–1784, 2022.

[14] S. Shurrab, A. Guerra-Manzanares, A. Magid, B. Piechowski-Jozwiak, S. F. Atashzar, and F. E. Shamout, "Multimodal machine learning for stroke prognosis and diagnosis: A systematic review," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[15] J. Heo, J. G. Yoon, H. Park, Y. D. Kim, H. S. Nam, and J. H. Heo, "Machine learning–based model for prediction of outcomes in acute stroke," *Stroke*, vol. 50, no. 5, pp. 1263–1265, 2019.

[16] D. O. Kleindorfer, R. Miller, C. J. Moomaw, K. Alwell, J. P. Broderick, J. Khoury, D. Woo, M. L. Flaherty, T. Zakaria, and B. M. Kissela, "Designing a message for public education regarding stroke: does fast capture enough stroke?" *Stroke*, vol. 38, no. 10, pp. 2864–2868, 2007.

[17] Y. Wang, Y. Ye, S. Shi, K. Mao, H. Zheng, X. Chen, H. Yan, Y. Lu, Y. Zhou, W. Ye, and et al., "Prediagnosis recognition of acute ischemic stroke by artificial intelligence from facial images," *Aging Cell*, vol. 23, no. 8, p. e14196, 2024.

[18] T. Cai, H. Ni, W. Ma, Y. Xue, Q. Ma, R. Leicht, K. Wong, J. Volpi, S. T. Wong, J. Z. Wang, and et al., "Safetriage: facial video de-identification for privacy-preserving stroke triage," in *International Conference on Information Processing in Medical Imaging*. Springer, 2025, pp. 390–404.

[19] R. Mahum, I. Ganiyu, L. Hidri, A. M. El-Sherbeeny, and H. Hassan, "A novel swin transformer based framework for speech recognition for dysarthria," *Scientific Reports*, vol. 15, no. 1, p. 20070, 2025.

[20] Y. Yang, X. Wu, X. Liu, J. Liu, J. Zhou, R. Wang, X. Wang, R. Su, N. Yan, and L. Wang, "Feature extraction method based on contrastive learning for dysarthria detection," in *International Conference on Social Robotics*. Springer, 2024, pp. 272–281.

[21] T. Cai, K. Wong, J. Z. Wang, S. Huang, X. Yu, J. J. Volpi, and S. T. Wong, "M 3 stroke: Multi-modal mobile ai for emergency triage of mild to moderate acute strokes," in *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2024, pp. 1–8.

[22] Z. Ou, H. Wang, B. Zhang, H. Liang, B. Hu, L. Ren, Y. Liu, Y. Zhang, C. Dai, H. Wu, and et al., "Early identification of stroke through deep learning with multi-modal human speech and movement data," *Neural Regeneration Research*, vol. 20, no. 1, pp. 234–241, 2025.

[23] M. Yu, T. Cai, X. Huang, K. Wong, J. Volpi, J. Z. Wang, and S. T. Wong, "Toward rapid stroke diagnosis with multimodal deep learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 616–626.

[24] T. Cai, H. Ni, M. Yu, X. Huang, K. Wong, J. Volpi, J. Z. Wang, and S. T. Wong, "Deepstroke: An efficient stroke screening framework for emergency rooms with multimodal adversarial deep learning," *Medical Image Analysis*, vol. 80, p. 102522, 2022.

[25] American Speech-Language-Hearing Association, "Dysarthria in adults," https://www.asha.org/practice-portal/clinical-topics/dysarthria-in-adults/.

[26] M. R. Rodrigues, M. Slimovitch, G. Chilingaryan, and M. F. Levin, "Does the finger-to-nose test measure upper limb coordination in chronic stroke?" *Journal of Neuroengineering and Rehabilitation*, vol. 14, no. 1, p. 6, 2017.

[27] D. T. Nguyen, M. N. Nguyen, A. T. Le, N. D. D. Kien, and Q. D. Minh, "Analyzing and evaluating personality and human behavior based on facial index and big five model," *Journal of Computer Science and Cybernetics*, vol. 40, no. 3, pp. 249–265, 2024.

[28] Google, "Face landmarker - mediapipe solutions (vision)," https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker.

[29] H. Fleyeh and J. Westin, "Extracting body landmarks from videos for parkinson gait analysis," in *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2019, pp. 379–384.

[30] Google, "Pose landmarker - mediapipe solutions (vision)," https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker?hl=vi.

[31] W. Zhu and M. Omar, "Multiscale audio spectrogram transformer for efficient audio classification," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

Received on February 11, 2026
Accepted on March 24, 2026