Medical Image Analysis 80 (2022) 102522

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

DeepStroke: An efficient stroke screening framework for emergency rooms with multimodal adversarial deep learning

Tongan Cai[a,1,*], Haomiao Ni[a,1], Mingli Yu[b], Xiaolei Huang[a], Kelvin Wong[c], John Volpi[d], James Z. Wang[a,*], Stephen T.C. Wong[c,*]

[a] College of Information Sciences and Technology, The Pennsylvania State University, University Park, Pennsylvania 16803, USA
[b] Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16803, USA
[c] T.T. and W.F. Chao Center for BRAIN & Houston Methodist Cancer Center, Houston Methodist Hospital, Houston, Texas 77030, USA
[d] Eddy Scurlock Comprehensive Stroke Center, Department of Neurology, Houston Methodist Hospital, Houston, Texas 77030, USA

ARTICLE INFO

Article history:
Received 8 September 2021
Revised 2 April 2022
Accepted 24 June 2022
Available online 25 June 2022

MSC:
68T10
68T45
92C55

Keywords:
Stroke
Multi-modal
Computer vision

ABSTRACT

In an emergency room (ER) setting, stroke triage or screening is a common challenge. A quick CT is usually done instead of MRI due to MRI's slow throughput and high cost. Clinical tests are commonly referred to during the process, but the misdiagnosis rate remains high. We propose a novel multimodal deep learning framework, *DeepStroke*, to achieve computer-aided stroke presence assessment by recognizing patterns of minor facial muscles incoordination and speech inability for patients with suspicion of stroke in an acute setting. Our proposed *DeepStroke* takes one-minute facial video data and audio data readily available during stroke triage for local facial paralysis detection and global speech disorder analysis. Transfer learning was adopted to reduce face-attribute biases and improve generalizability. We leverage a multi-modal lateral fusion to combine the low- and high-level features and provide mutual regularization for joint training. Novel adversarial training is introduced to obtain identity-free and stroke-discriminative features. Experiments on our video-audio dataset with actual ER patients show that *DeepStroke* outperforms state-of-the-art models and achieves better performance than both a triage team and ER doctors, attaining a 10.94% higher sensitivity and maintaining 7.37% higher accuracy than traditional stroke triage when specificity is aligned. Meanwhile, each assessment can be completed in less than six minutes, demonstrating the framework's great potential for clinical translation.

1. Introduction

Stroke is a common cerebrovascular disease that can cause lasting brain damage, long-term disability, or even death (Centers for Disease Control and Prevention, 2005). It is the second leading cause of death and the third leading cause of disability worldwide (Johnson et al., 2016). According to the Centers for Disease Control and Prevention (2020), someone in the United States has a stroke every forty seconds and someone dies of a stroke every four minutes. In acute ischemic stroke where brain tissue lacks blood supply, the shortage of oxygen needed for cellular metabolism quickly causes long-lasting tissue damage. If identified and treated in time, many interventions are available and an acute ischemic stroke patient will have a greater chance of survival and subsequently a better quality of life.

However, treatment delays related to misdiagnosis and underdiagnosis are common during triage (Rafay et al., 2009). A major problem in acute ischemic stroke is misdiagnosis. Misdiagnosis in stroke leads to undertreatment, overtreatment, and the potential for mental and physical disability (Crichton et al., 2016). Approximately 22% of all ischemic stroke patients are missed during the pre-hospital triage screening, and the problem is more severe in community hospitals than in academic hospitals (Arch et al., 2016).

Rapid diagnosis of acute ischemic stroke relies on clinical diagnosis and imaging as there is no point-of-care test available. Currently, the gold standard test for stroke is advanced neuro-imaging including diffusion-weighted MRI scan (DWI) that detects brain infarct with high sensitivity and specificity. Although accurate, DWI is usually not accessible in the emergency room (ER) due to limited equipment availability and high operating cost. Even in advanced centers with ER availability of MRI, the turnaround time for patient transport, testing, and results adds 30–60 minutes, which is too in-

* Corresponding authors.
E-mail addresses: cta@psu.edu (T. Cai), jwang@psu.edu (J.Z. Wang), stwong@houstonmethodist.org (S.T.C. Wong).
[1] T. Cai and H. Ni made equal contributions.

https://doi.org/10.1016/j.media.2022.102522
1361-8415/

efficient for the stroke triage screening process. The healthcare system, instead, relies on rapid imaging such as CT and clinical judgment of nurses and physicians to detect neurological symptoms during emergency triage. In the actual ER scenario, clinicians commonly adopt the following three tests: the Cincinnati Pre-hospital Stroke Scale (CPSS) (Kothari et al., 1999), the Face Arm Speech Test (FAST) (Harbison et al., 2003), and the National Institutes of Health Stroke Scale (NIHSS) (NIH, 2003). All these methods assess the presence of any unilateral facial droop, arm drift, and speech disorder. The patient is requested to repeat a specific sentence (CPSS) or have a conversation with the doctor (FAST), and abnormalities arise when the patient slurs, fails to organize his speech, or is unable to speak. For NIHSS, face and limb palsy conditions are also evaluated. However, the scarcity of neurologists (Leira et al., 2013) makes such tests difficult to be timely and effectively conducted in all stroke emergencies. The evaluation may also fail to detect stroke cases where only very subtle facial motion deficits exist—that clinicians are unable to observe.

Recently, with the help of machine intelligence, researchers have proposed more and more accurate detection and evaluation methods for neurological disorders. Besides working on computer-aided medical image understanding for CT and MRI images (Xue et al., 2018; Yao et al., 2018; Akkus et al., 2017), many researchers are now focusing on alternative contactless, efficient, and economic ways for the analysis of various neurological conditions. One of the most popular domains is the detection of facial paralysis with computer vision, *i.e.*, letting machines detect the anomalies in the subject's face. However, the clinical scenario is always overlooked where obvious stroke patients are readily identified without treatment delay and subtle/non-obvious strokes are easily missed. Obvious strokes can be identified by key-point methods (Parra-Dominguez et al., 2021; Zhuang et al., 2021) but the same strategy will not be effective on subtle strokes.

Moreover, the majority of work neglects the readily available and indicative speech audio features (Thevenot et al., 2017), which can be an important source of information in stroke diagnosis. Also, current methods ignore the spatiotemporal continuity of facial motions (He et al., 2008; Hakata et al., 2013; Flores-Mondragón et al., 2015) and fail to tackle the problem of static/natural asymmetry. Common video classification frameworks like I3D (Carreira and Zisserman, 2017) and SlowFast (Feichtenhofer et al., 2019) also fail to serve the stroke pattern recognition purpose due to the lack of training data and quick overfitting as "subject-remembering" effect.

Worse still, few datasets of high quality have been constructed in the stroke diagnosis domain. The current clinical datasets (Perveen, 2019; Kihara et al., 2011; Hsu et al., 2018; Greene et al., 2020) are small (with hundreds of images or dozens of videos) and unable to comprehensively represent the diversity in stroke patients in terms of gender, race/ethnicity, and age. Also, they either evaluate normal subjects *versus* those with clear signs of a stroke (Bandini et al., 2016; Ngo et al., 2016) or deal with full synthetic data (*i.e.*, healthy people pretend to have palsy facial patterns) (Zhuang et al., 2018; Storey and Jiang, 2018); some others establish experimental settings with hard constraints on the patient's head (Bevilacqua et al., 2011; Hamm et al., 2011; Wang et al., 2014). All these shortcomings will hinder their clinical implementation for ER screening or patient self-assessment.

In this paper, we propose a novel deep learning framework, named *DeepStroke*, to assist the stroke triage team in accurately and efficiently analyzing the presence of stroke in patients admitted to the Emergency Rooms (ER) with a subtle sign of stroke. The problem is formulated as a binary classification task, *i.e.*, stroke *vs.* non-stroke. Instead of taking a single-modality input, the core network in *DeepStroke* consists of two temporal-aware branches, the video branch for local facial motion analysis and the audio branch for global vocal speech analysis, to collaboratively detect the presence of stroke patterns. A novel lateral fusion scheme between these two branches is introduced to combine the low- and high-level features and provide mutual regularization for joint training. To mitigate the "subject-remembering" effect, *DeepStroke* also adopts adversarial learning to extract identity-free and stroke-discriminative features, as well as transfer learning to reduce facial-attribute biases and improve the network generalizability.

To better illustrate the use case of our proposed method, we plot the anticipated clinical workflow of *DeepStroke* in helping stroke diagnosis in ER, as shown in Fig. 1. If the incoming patient has a clear indication of a stroke, he/she should be directly transferred to the stroke team for evaluation and treatment. For non-obvious cases, *DeepStroke* will be applied for the reference of the triage team and ER doctors to reduce misdiagnosis.

To evaluate *DeepStroke*, we construct a stroke patient video/audio dataset that records the facial motions of the patients during their process of performing a set of vocal speech tests when they visit the ER. The recruited participants are all showing some level of neurological conditions with suspicion of stroke when visiting the ER. This is closer to the real ER scenarios and *much more challenging* than distinguishing stroke patients from healthy people. Our dataset includes diverse patients of different genders, races/ethnicity, ages, and at different levels of stroke conditions; the subjects are free of motion constraints and are in arbitrary body positions, illumination conditions, and background scenarios, which can be regarded as "in-the-wild." Experiments on our dataset show that the *DeepStroke* framework achieves higher performance than the typical stroke triage team and even outperform trained ER clinicians for stroke (who has more clinical information available) while maintaining a manageable computation workload.

A preliminary version of this work was presented in our earlier publication (Yu et al., 2020). In this paper, we substantially extend our prior work in the following aspects. First, we substitute the text transcript inputs with the spectrogram input and replace the text LSTM with ResNet-18 to maintain the most of speech information, reduce errors induced by the transcription process as well as ease the later feature fusion with video module. Secondly, we introduce a lateral connection to the two branches to resolve the unstable convergence of the network caused by different training dynamics of branches and share low-/high-level features for a better combination of the global context (audio) and local representation (video). Thirdly, we adopt a transfer learning pipeline and introduce state-of-the-art fairness-aware face image pre-training and image classification pre-training to reduce facial feature biases, mitigate network overfitting and improve network generalizability. Fourthly, we introduce an adversarial training scheme aiming to remove the subject identity features from the input to alleviate the "subject-remembering" effect of the deep neural networks. Lastly, we significantly enlarge the dataset with newly collected data to support extensive new cross-validation experiments and hold-out study on our proposed new framework.

The **main contributions** of this work are summarized below.

1. We constructed a real ER patient facial video and vocal audio dataset for stroke screening with diverse participants. The videos are collected "in the wild," with unconstrained patient body positions, environment illumination conditions, or background scenarios.
2. We propose a deep-learning multimodal framework, *DeepStroke*, that highlights its video-audio multi-level feature fusion scheme that combines global context to local representation, the adversarial training that extracts identity-free stroke features, the transfer learning that improves generalizability, and

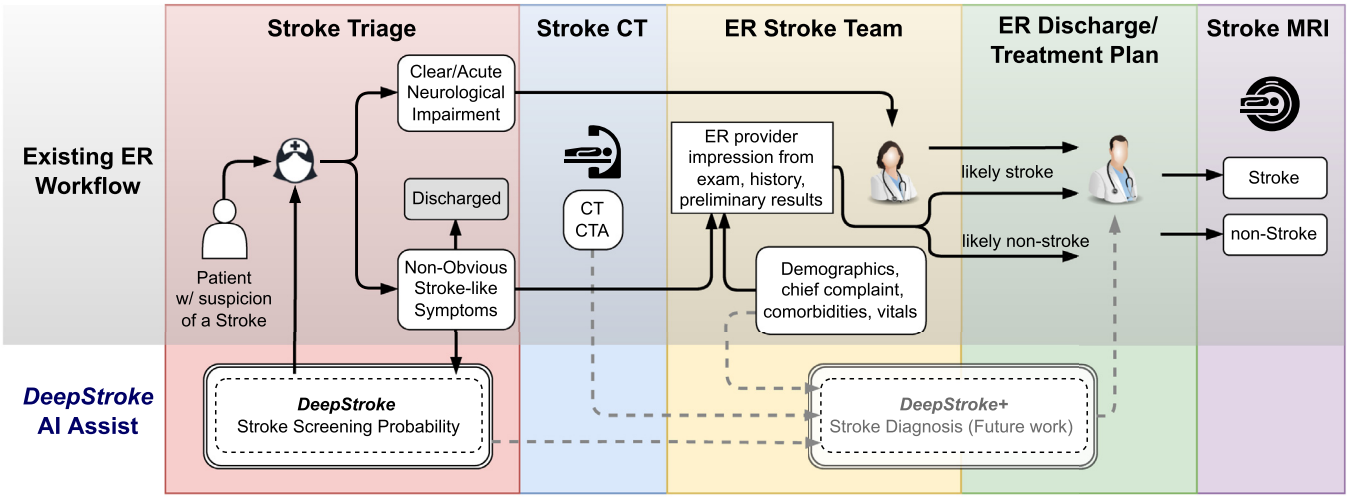


Fig. 1. The anticipated workflow of *DeepStroke* in assisting stroke diagnosis in the ERs.

the spatiotemporal proposal mechanism for frontal human facial motion sequences.

3. The proposed multi-modal method achieves high diagnostic performance and efficiency on our proposed challenging dataset and outperforms the clinicians, demonstrating its high clinical value to be deployed for real ER use.

2. Related work

2.1. Facial motion analysis for medical diagnosis

In the past two decades, researchers have been striving for more and more novel facial motion analysis methods to achieve accurate detection and evaluation of facial paralysis and other medical conditions. Early work focused on static image analysis. Rogers et al. (2007) aimed to detect the improvement in abnormal facial movements after treatment with botulinum toxin A; Dong et al. (2008) introduced a "face nerve index" based on coordinates of image feature points and achieved an accurate classification of facial paralysis levels. Some others utilized multiple images in a sequence for improved analysis, *e.g.*, He et al. (2007) and Green and Guan (2004). With video data, researchers can capture between-frame motions of patient faces and temporal patterns in facial abnormalities. For example, Frey et al. (2008) analyzed the video trajectories of facial regions for face paralysis detection and evaluation; Bandini et al. (2017) extracted articulatory movements of patient faces in a video to identify patterns of Parkinson's disease. A small video dataset for facial paralysis has been offered by Greene et al. (2020) along with the eFACE (Banks et al., 2015), House-Brackmann (House and Brackmann, 1985), and Sunnybrook (Ross et al., 1996) metrics. There is not a publicly-available image or video database specifically for the evaluation of stroke within a patient due to the concern of personal identity leakage and IRB constraints; if no real patient case is available, synthetic datasets like CK+ (Lucey et al., 2010) have to be used, where only healthy participants are involved. Biases in genders, ages, and races will arise as the majority of the prior datasets only have a limited number of subjects.

With recent developments in computer vision and especially face alignment, the facial landmark is becoming a popular representation of facial patterns. Anping et al. (2017) measured asymmetry for facial movement images based on facial landmarks; Zhuang et al. (2018) and Szczapa et al. (2019) also introduced similar facial landmark-based methods on video or image data. Wang et al. (2016) and Zhuang et al. (2019) stratified the face of a patient into multiple regions and considered regional information with landmarks. While easy to obtain, the landmarks are not robust and suffer greatly from illumination, occlusion, and motion. Some more recent work designs deep learning methods for the task. Song et al. (2018) developed an Inception-DeepID-FNP neural network for paralysis classification, Storey and Jiang (2018) considered a multi-task network for both face detection and facial asymmetry evaluation, and Guo et al. (2017) introduced a deep learning framework to classify facial paralysis with regard to the H-B standard. However, all these deep learning frameworks suffer from non-standard head poses or non-facial motions. To address the alignment issues, the majority of the prior work sets up experimental settings with uniform illuminations and puts up hard constraints on the subjects' heads to avoid alignment issues. Such simplification of scenarios greatly hinders the clinical value of these proposed methods, especially in an emergency setting.

Upon reviewing the above literature, we concluded that existing approaches either evaluate their methods between subjects that are normal *versus* those with clear signs of a stroke, deal with only synthetic data, fail in capturing the spatiotemporal details of facial muscular motions, or rely on experimental settings with hard constraints. Without considering challenges in real emergencies such as illumination variations, pose changes, skin color, and subtleties of patterns, such methods are not practical in real clinical practice or for patient self-assessment.

2.2. Disentangle representation learning

Our work is also closely related to disentangle representation learning. Tran et al. (2017) developed DR-GAN to learn a generative and discriminative representation for pose-invariant face recognition. Liu et al. (2018) proposed the D^2AE framework to adversarially learn the identity-free features for identity verification and the identity-dispelled features to fool the verification system. Zhang et al. (2019) proposed an AutoEncoder framework to explicitly disentangle pose and appearance features from RGB imagery and the LSTM-based integration of pose features over time produces the gait feature. Lu et al. (2019) developed an unsupervised domain-specific image deblurring framework by disentangling the content and blur features. Lee et al. (2018) proposed an image-to-image translation framework by disentangling image features into a domain-invariant content space and a domain-specific attribute space and combining them to produce diverse output.

Especially, learning identity-free stroke patterns for diagnosis is a similar process to extracting identity-free features for facial expression recognition (FER), which also aims to detect patterns of facial movements. Meng et al. (2017) designed identity-contrasting loss that goes in parallel with the expression-related loss to achieve identity-invariant expression recognition. Cai et al. (2019) proposed a conditional generative model to transform an average neutral face into an average expressive face with the same expression as the input image that is naturally identity-free. Wang et al. (2019) constructed a pose discriminator and a subject discriminator to classify the pose and the subject from the extracted feature representations, respectively, to make them robust to poses and subjects. In this work, we adopt a similar approach and set up identity-related discriminators, and use an adversarial training scheme to extract identity-free features. To our knowledge, our work is the first attempt that introduces adversarial training loss to stroke diagnosis tasks.

2.3. Multimodal deep learning

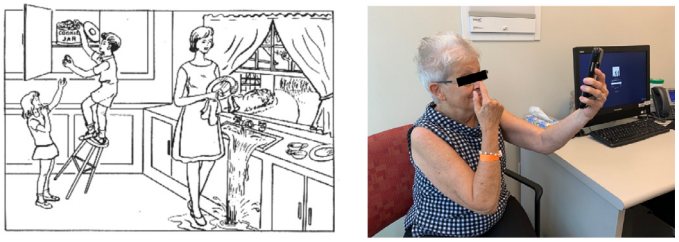
Deep networks that learn features over multiple modalities via shared representation learning (Ngiam et al., 2011) have achieved outstanding performance on various multimedia tasks. Intuitively, when more than one form of data is available, a deep network that utilizes most of them and takes aligned information from different sources, where each modality tends to complement each other and incorporate richer information, will usually result in higher task performance than using these information sources individually. The idea of multimodal deep learning has been making great improvements to domains including emotion recognition (Kahou et al., 2016), disease diagnosis (Xu et al., 2016), object detection (Eitel et al., 2015), etc. Common modalities include text, audio, and video, and they complement each other within a multimodal neural network.

Specifically for disease diagnosis, multimodal methods have been proven useful to integrate patient data from multiple platforms or in different forms. Li and Shen (2018) proposed a multi-view deep learning framework (MvNet) with three branches to segment multimodal brain images from different view-points, *i.e.* slices along x-, y-, z-axes; Menegotto et al. (2020) developed a multimodal deep machine learning architecture for hepatocarcinoma diagnosis with computed tomography images, laboratory test results, anthropometric and sociodemographic data as input; to model the difficulties to start or to stop movements for patients with Parkinson's disease, Vásquez-Correa et al. (2019) conducted multimodal deep learning over information from speech, handwriting, and gait; Liang et al. (2015) proposed multimodal deep belief network (DBN) to cluster cancer patients observation data from different platforms. The fusion of medical imaging and electronic health records using deep learning has been another heated topic in recent years (Huang et al., 2020), which will make pixel-based models and contextual data from electronic health records (EHR) work cooperatively for the diagnostic purpose.

Stroke diagnosis can also benefit from multimodal deep learning. To diagnose a stroke case, neurologists detect the facial deficits on the patient's face or the disorder in speech—the two modalities tend to work jointly. In this work, we adopt the multimodal idea to allow the facial video and vocal audio to work together for a final diagnosis.

3. Dataset

The clinical dataset for this study was acquired in the ERs of the Houston Methodist Hospital in Texas by the physicians and caregivers from the Eddy Scurlock Stroke Center at the Hospital under an IRB-approved study.[1] It has taken more than a year for us to recruit a large pool of patients in various stroke-related emergencies, and the cohort is still expanding. The subjects enrolled are patients with suspicion of stroke while visiting the ER while obvious stroke cases with severe symptoms are excluded. To help preserve the patients' personal information, we only transmitted limited information that we think is sufficient for this study between institutes to avoid any identifiable information to be collected and dispensed. The patients are only assessed when they're in a relatively stable condition so the collection of data will not impose extra risk on high-emergency cases. Note that the gender ratio of our dataset is relatively balanced without intervention, and the race/ethnicity or age distributions are not manually controlled, which would roughly represent the real distribution of the incoming ER patients. Potential bias will be discussed in Section 5.



(a) The "Cookie Theft" picture **(b)** Setup of the recording protocol

Fig. 2. The speech task and the sample recording protocol.

Clinically, the ability of speech is an important and efficient indicator of the presence of stroke and is the preferable measurement doctors will use to make initial clinical impressions; if a potential patient slurs, mumbles, or even fails to speak, he or she will have a very high chance of stroke (Harbison et al., 2003; Kothari et al., 1999). During the evaluation and recording stage, we follow the NIH Stroke Scale (2003)) and perform the following speech tasks on each subject: (1) We ask the patient to repeat the sentence "it is nice to see people from my hometown," and (2) we ask the patient to describe the "Cookie Theft" picture as shown in Fig. 2a.

The "Cookie Theft" task requires the subject to retrieve, think, organize, and express the information, which will evaluate the patient's speech ability from both motor and cognitive aspects. It has been making great success in identifying patients with Alzheimer's-related dementia, aphasia, and some other cognitive-communication impairments (Giles et al., 1996), and would be suitable for our stroke screening purpose.

The subjects are video recorded when they are performing the two tasks. The collection protocol is set up with an Apple iPhone X, as shown in Fig. 2b. Note that a phone rack and an auxiliary microphone will be used if the patient is too weak to hold the device stably. Each video is accompanied by metadata information on both clinical impressions by the ER physician (indicating the doctor's initial judgment on whether the patient has a stroke or not from his/her speech and facial muscular conditions) and ground truth from the diffusion-weighted MRI (including the presence of acute ischemic stroke, transient ischemic attack (TIA), etc.). Samples of such metadata information are shown in Table 1.

Note that in this dataset, the Clinical Impression is given after the ER screening by the doctors who usually have access to emergency imaging reports, vitals, and other information in the Electronic Health Records (EHR). Some early-enrolled patients are

[1] This study is conducted under Houston Methodist IRB protocol No. Pro00020577, Penn State IRB site No. SITE00000562.

Table 1
Metadata info for subjects enrolled. (PE: Patient Example #).

PE	Complaint	Clinical Impression	Discharge Diagnoses	Details
1	Weakness	DR: Stroke / TIA	1, Acute ischemic stroke 2, Dyslipidemia 3, Type 2 diabetes A1c	L cortical infarct (precentral gyrus) and additional R corpus callosum (subacute)
2	Other - Neurologic problem	DR: Stroke / TIA	1, TIA 2, possible TIA *versus* seizure 3, acute metabolic encephalopathy 4, UTI 5, Obesity 6, HLD 7, HTN 8, Pre-diabetes	right vertebral artery, V3, V4 Unable to obtain MRI due to metal artifact
3	Tremor, Stuttering	Aphasia, Ataxia, Unspecified Tremor, CVA	1, Tremors	Neurology gave pt DDx of acute ischemic stroke, TIA, and toxic--metabolic. Workup ruled out CVA and indicated breakthrough
4	...	...	...	...

Table 2
Summary of subjects' demographic information in the dataset.

Attribute	Group	Stroke	Non-Stroke	Total
Gender	Male	76	32	108
	Female	83	30	113
Age	$\leq$ 65 y/o	80	36	116
	$\geq$ 65 y/o	79	26	105
Ethnicity	Hispanic	12	4	16
	Non-Hispanic	146	52	198
	Opt-out	1	6	7
Race	African American	54	18	72
	Other	105	44	149
All subjects		159	62	221

not included later in this study due to missing clinical impression data resulting from being transferred from another healthcare provider.

We have been recruiting new subjects to expand the dataset in the last few years. Up to the time of completing the manuscript, 108 males and 113 females have been recruited, non-specific of age, race/ethnicity, or the seriousness of stroke. Among the 221 individuals, 159 are patients diagnosed with stroke using MRI, 62 are patients who do not have a stroke but are diagnosed with other clinical conditions. A summary of demographic information is shown in Table 2. In this work, we formulate the diagnosing process as a binary classification task and only attempt to identify stroke/TIA cases from non-stroke cases. Though there are varieties of stroke subtypes, binary output has been sufficient to function as a screening decision in ER.

Our dataset is unique, as compared to existing ones (Guo et al., 2017; He et al., 2007), because our cohort consists of actual patients visiting the ERs and the videos are collected under unconstrained, or "in-the-wild" conditions. Existing work generally setup experimental settings before collecting the image or video data, which will result in uniform illumination conditions and minimum background noise. In our dataset, the patients can be in bed, sitting, or standing, where the background and illumination are usually not under ideal control conditions. Furthermore, previous work often enforced rigid constraints over the subjects' head motions, which sidesteps the alignment challenges and makes the unrealistic assumption of having stable face poses. We only ask patients to focus on the instructions, without rigidly restricting their motions. We use intelligent video-processing methods to accommodate for the "in-the-wild" conditions. The acquisition of facial data in natural settings allows comprehensive evaluation of the robustness and practicability of our work for real-world clinical use, remote diagnosis, and self-assessment in most settings.

4. Methods

Fig. 3 shows the training framework of *DeepStroke*. We first preprocess the i-th input raw video to obtain the facial-motion-only, near-frontal face sequence $\mathcal{F}_i$ and its corresponding audio spectrogram $\mathcal{M}_i$. Then we extract the audio-visual feature e_i from $\mathcal{F}_i$ and $\mathcal{M}_i$ by a lateral-connected dual-branch Encoder E, which includes a video module Γ_v for local visual pattern recognition, and an audio module Γ_a for global audio feature analysis. A subject Discriminator D is also employed to help E learn features that are insensitive to subject identity difference but are sensitive to distinguishing stroke from non-stroke. When training, we use the case-level label as a pseudo label for each video frame and train the *DeepStroke* network as a frame-level binary classification model. We also sample the intermediate output feature maps from different videos to train D and E adversarially. During inference, we first perform frame-level classification and then calculate the case-level predictions by averaging over all frames' probabilities to mitigate frame-level prediction noise. More details are described as follows.

4.1. Data preprocessing

For each raw video, we propose a spatiotemporal proposal mechanism to extract frontal-face sequences from the raw video. For each audio extracted from the raw video, we transform the soundtrack into a spectrogram that represents the amplitude at each frequency level over time.

Spatiotemporal proposal of facial action video: In facial motion analysis, one challenge is to achieve good face alignment. As our data are collected "in the wild", we introduce a pipeline to extract frame sequences with near-frontal facial pose and minimum non-facial motions. First, we detect and track the patient's face with a rigid, square bounding box and estimate the poses. Frame sequences 1) estimated with significant roll, yaw, or pitch, 2) showing continuously changing pose metrics, or 3) having excessive head translation or too little change estimated with optical flow magnitude with the previous frame, are excluded (detailed criteria are presented in Section 5.1). A video stabilizer with a sliding window over the trajectory of between-frame affine transformations smooths out pixel-level vibrations on the sequences before they are passed to the Encoder E.

Speech spectrum analysis: Instead of transcribing the vocal audio to text corpus (Yu et al., 2020), which may suffer from translation errors, we turned to spectrograms for speech analysis due to the following reasons: (1) Spectrogram is the complete representation of an audio file since the amplitude over frequency bands are captured over time. Recent deep learning-based transcription mod-

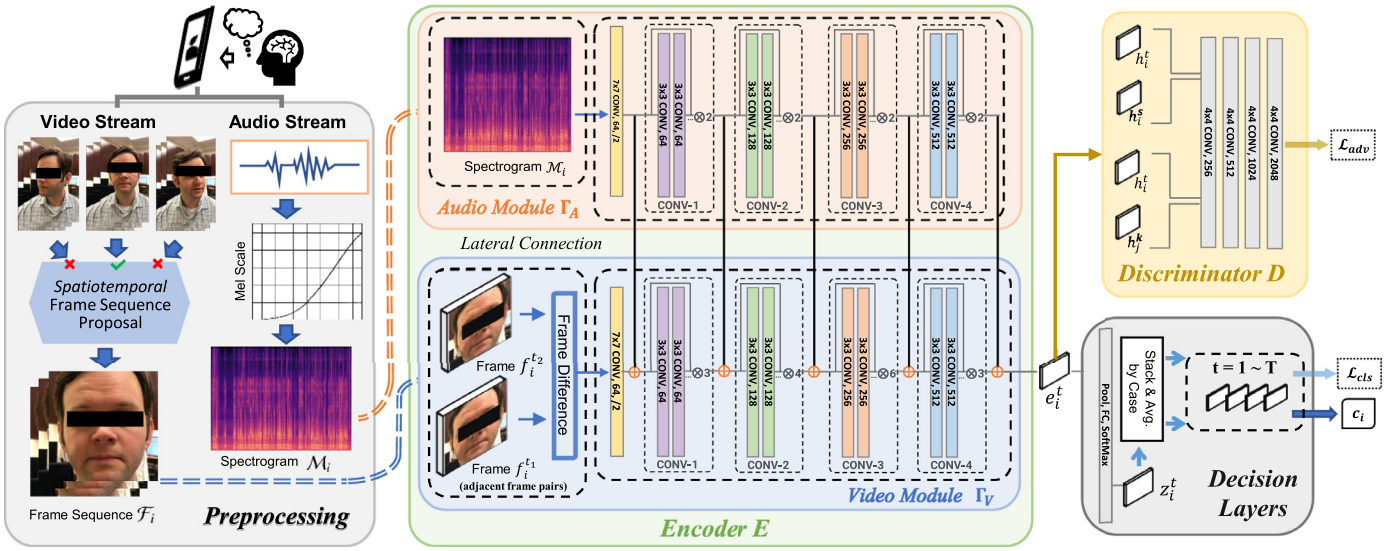


Fig. 3. The training framework of *DeepStroke*. The preprocessing part involves a spatiotemporal facial frame sequence proposal and a transformation of audio features to spectrograms. In the encoder E, at any timestamp t, the input frame pairs $f_i^{t_1}$ and $f_i^{t_2}$ are guaranteed to be adjacent pairs, and the total spectrogram is duplicated and appended as the audio features for time t. The symbol $\oplus$ indicates the concatenation with the lateral connection.

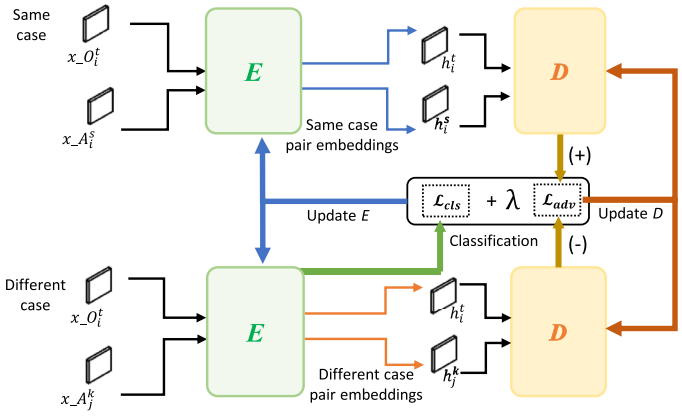


Fig. 4. Workflow of the discriminator D of *DeepStroke*. $\mathcal{L}_{\text{cls}}$ will be calculated first using $x\_O_i^t$, the original input (including $f_i^{t_1}$ and $\mathcal{M}_i$), and generate feature embedding h_i^t. Then the encoder will freeze parameters and generate h_i^s and h_j^k for $x\_A_i^s$ and $x\_A_j^k$, the adversarial inputs from the same/different cases. Positive label $(+)$ is assigned to same-case pairs and negative $(-)$ to different-case pairs. $\mathcal{L}_{\text{adv}}$ is then calculated to update D, and a weighted sum of $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{adv}}$ is used to update E.

els commonly transform soundtracks into spectrograms for their classification models (Hannun et al., 2014; Amodei et al., 2016). (2) By choosing spectrogram, which is image-like input, we can adapt similar networks for the two branches and ensure they have rather similar training dynamics to converge at a similar pace, which will make them cooperate better. In this work, we use the Mel Scale (Stevens et al., 1937) and Fast Fourier Transform (FFT) to transform the audio signal before plotting the spectrogram.

4.2. Model design

As Fig. 3 shows, after preprocessing the raw input videos, we further extract stroke-discriminative and identity-free features from the input video and audio via the feature encoder E and the subject discriminator D.

Feature encoder: Let $\mathcal{F}_i = (f_i^1, \cdots, f_i^T)$ denotes a sequence of T temporally-ordered frames from the i-th input video and its corresponding spectrogram is $\mathcal{M}_i$. We extract their features through feature encoder E, which includes one video module Γ_{v} and one audio module Γ_{a}, fused by lateral connection.

▷ *Video Module.* To extract temporal visual features from input $\mathcal{F}_i$, a pair of adjacent frames from $\mathcal{F}_i$ is forwarded to the video module Γ_{v}. Due to the frame-proposal process, the original frame sequence sometimes has long gaps between two nearby frames, which will result in large, non-facial differences being captured. We tackle this by keeping track of the frame index and only sample a specific number of real adjacent frame pairs in $\mathcal{F}_i$ (frame $f_i^{t_1}$ and $f_i^{t_2}$) to extract local visual information. Instead of directly inputting a pair of frames, for better capturing subtle facial motions between adjacent frames, we compute the image difference between $f_i^{t_1}$ and $f_i^{t_2}$ and then pass it through the network as the feature for the frame pair f_i^t.

▷ *Audio Module.* To extract disease patterns from the input audio spectrogram $\mathcal{M}_i$, we feed $\mathcal{M}_i$ to the audio module Γ_{a}. Since $\mathcal{M}_i$ contains the whole temporal dynamics of the input audio sequence, we append $\mathcal{M}_i$ to each frame pair x_i^t and x_i^{t+1} to provide a global context for the frame-level stroke classification.

▷ *Lateral Connection.* To effectively combine features of video $\mathcal{F}_i$ and audio $\mathcal{M}_i$ at different levels, we also introduce lateral connection (Xiao et al., 2020) between the convolutional blocks of the video module Γ_{v} and the audio module Γ_{a}. To ensure the features are aligned when being appended, we perform 1×1 convolution (Lin et al., 2013) to project the global audio feature to the same embedding space as the local frame features and then sum them up. Compared with the late fusion of two branches used in our prior work (Yu et al., 2020), lateral connections-based fusion not only combines more levels of features but also enables different branch dynamics to stay similar, which will maintain the convergence rate of each branch to be relatively closer and help the global context better complement the local context during the training stage.

Subject discriminator: Due to the relatively small number of available videos, it is easy and tempting for the encoder E to memorize the facial and audio features of each subject and just match testing subjects with training subjects based on similarity in appearance and voice when performing the inference. To avoid this issue of classification based on subject-dependent features, we further design a subject discriminator D with an adversarial learning idea to help encoder E learn identity-free features. The discrimi-

nator D is designed to simply distinguish whether the input pair of intermediate features from encoder E are from the same subject or not and will be used to adversarially train the encoder E. In the implementation, when we feed the (original) input data to the network, we will take another data batch (denoted as the adversarial data) from either the same case or a different case with equal probability. When training D, both original and adversarial data will be processed by the encoder E with parameters frozen for feature embeddings. If the original and adversarial data came from the same case, we assign a positive label to the embedding pair (and a negative label otherwise). The generated feature embedding will include both information from the audio and video (thanks to the lateral connection in the encoder E) and are used to train D. An adversarial loss is given to update D. The adversarial loss is also added to the classification loss for the later update of encoder E. Details of the loss calculation is introduced in Section 4.3.2

In theory, the network of *DeepStroke* can employ various networks as its backbone. Here we choose ResNet-34 for Γ_v and ResNet-18 for Γ_a to accommodate the relatively small size of the video dataset and the simplicity of the spectrogram, and to reduce the computational cost of the framework. For the discriminator D, we follow the design of DCGAN (Radford et al., 2015) with four convolution layers and binary output. For the intermediate feature input to the discriminator, we take the frame-level feature e_i^t (output feature from the decoder) for case i at time t. The choice of intermediate feature is ablated in Section 5.3.1.

Decision layers: During inference, because our model is trained with pseudo frame-level labels, to mitigate frame-level prediction noise, we first perform frame-level classification using encoder E into final fused features e_i^t. e_i^t is then passed through the fully-connected and softmax layers to generate the frame-level class probability score z_i^t. The case-level prediction c_i is then obtained by stacking and averaging the frame-level predictions (stroke probability ranging from 0 to 1) and compare with the decision threshold.

4.3. Model training

4.3.1. Training with transfer learning

Due to the relatively small number of samples available, the proposed deep neural network can overfit on the training samples and result in a "face-remembering" effect. Besides employing the subject discriminator to alleviate this effect, we adopt a transfer-learning framework to train *DeepStroke* by starting with pre-trained network backbones and freezing the parameters. Only the final output layers and the discriminator module are fine-tuned. For the video module network, we employ state-of-the-art fairness-aware face image pre-training, the FairFace (Karkkainen and Joo, 2021) ResNet-34 model that was pre-trained on a face image dataset that has evenly distributed race, age, and ethnicity attributes. This aims to reduce the common facial-attribute biases and prevent network overfitting on irrelevant features. For the audio module, we apply an ImageNet-pre-trained ResNet-18 backbone as the transfer learning starting point to mitigate overfitting. The transfer learning improves the generalizability of the network and will be demonstrated in Section 5.3.

4.3.2. Loss functions

Classification loss: To help E learn stroke-discriminative features, we use a standard binary cross-entropy loss between the prediction z_i^t and video label y_i for all the training videos and their T frames:

$$\mathcal{L}_{\text{cls}}(E) = -\sum_i \sum_t \left(y_i \log z_i^t + (1 - y_i)(1 - \log z_i^t) \right) \ . \tag{1}$$

Adversarial loss: To encourage Encoder E to learn identity-free features, we introduce a novel adversarial loss to ensure that the output feature map h_i^t does not carry any subject-related information. We impose this via an adversarial framework between the subject discriminator D and feature encoder E, as shown in Fig. 3. The latter, E, will provide a pair of feature maps, either computed from the same subject video (h_i^t, h_i^s) at time t and s, or from different subject videos (h_i^t, h_j^k) at time t and k, where time s and k can be randomly chosen. Discriminator D then attempts to classify the pair as being from the same/different subject video using an l_2 loss, as LS-GAN (Mao et al., 2017) adopts:

$$\mathcal{L}_{\text{adv}}(D) = -\sum_i \sum_t \left(\left\| D(h_i^t, h_i^s) \right\|_2 + \left\| 1 - D(h_i^t, h_j^k) \right\|_2 \right) \ . \tag{2}$$

The adversarial framework further imposes a loss function on the feature encoder E that tries to maximize the uncertainty of the discriminator D output on the pair of frames:

$$\mathcal{L}_{\text{adv}}(E) = -\sum_i \sum_t \left(\left\| \frac{1}{2} - D(h_i^t, h_i^s) \right\|_2 + \left\| \frac{1}{2} - D(h_i^t, h_j^k) \right\|_2 \right) \ . \tag{3}$$

Thus the encoder E is encouraged to produce features that the discriminator D is unable to classify if they come from the same subject or not. In so doing, the features h cannot carry information about subject identity, thus avoiding the model to perform inference based on subject-dependent appearance/voice features. Note that our model is different from classic adversarial training used in GANs (Goodfellow et al., 2014) because we only focus on classification and there is no generator network in our framework.

Overall training objective: During training, we minimize the sum of the above losses:

$$l = \mathcal{L}_{\text{cls}}(E) + \lambda(\mathcal{L}_{\text{adv}}(E) + \mathcal{L}_{\text{adv}}(D)) \ , \tag{4}$$

where λ is the balancing parameter. The first two terms can be jointly optimized, but the discriminator D is updated while the encoder E is held constant.

5. Experiment

To better present the details of our proposed method, we first introduce the setup and implementation details and then show the comparative study with baselines. We also ablate the power of model components and structures to validate our design.

5.1. Setup and implementation

The whole framework is running on `Python` 3.7 with `Pytorch` 1.1, `OpenCV` 3.4, `CUDA` 9.0, and `Dlib` 19. Both ResNet models are pre-trained on ImageNet (Deng et al., 2009). Each.mov file from the iPhone will first be separated as frame sequences as batches of.png files and one global.wav audio file.

For the frame sequences, we detect the location of the patient's face as a square bounding box with `Dlib`'s face detector (King, 2009) and track it using the `Python` implementation of the ECO tracker (Danelljan et al., 2017). We perform pose estimation by solving a direct linear transformation from the 2D landmarks predicted by `Dlib` to a 3D ground truth facial landmarks of an average face, which resulted in three values corresponding to the angular magnitudes over three axes—pitch, roll, and yaw. To tolerate estimation errors, we regard those with angular motions less than a threshold β_1 as frontal faces. A 5-frame sliding window records the between-frame changes in the pose. If the total changes (three axes) sum up to more than a threshold β_2, we abandon the frames starting from the first position of the sliding window. In the meantime, the between-frame changes are measured by optical flow magnitude. If the total estimated change is smaller than β_l (no motion) or larger than β_h (non-facial motions),

we also exclude the frame. We empirically set $\beta_1 = 5^\circ$, $\beta_2 = 20^\circ$, $\beta_l = 0.01$, and $\beta_h = 150$. After manipulation, we crop the size of each frame to $224 \times 224 \times 3$ to align with the ImageNet dataset. The real frame numbers are kept to ensure that only adjacent frame pairs are loaded to the network.

For the audio files, we use `librosa` to load and trim the soundwave, and plot the Log-Mel spectrogram, where the horizontal axis is the temporal line, the vertical axis is the frequency bands, and each pixel shows the amplitude of the soundwave at the specific frequency and time. The output spectrogram is also set to the size of $224 \times 224 \times 3$.

The entire pipeline is trained on a server computer with a hex-core CPU, 64GB RAM, and an NVIDIA GPU with 24GB VRAM. To accommodate for the class imbalance inside the dataset and ER setting, a higher class weight of 2.0 is assigned to the non-stroke class. The default learning rate is set to 1e-7 with decay after every five epochs, and we early stop at epoch 20 due to the quick convergence of the network. The batch size is set to be 64 and λ in (4) is set to be 10 for loss scaling purposes. Model parameters and learning rates are then tuned separately for each model/baseline in each experiment.

5.2. Baselines and experiments

We construct baseline models for both video and audio tasks. For each case (video/audio), the ground truth for comparison is the binary diagnosis result obtained through the MRI scan. Our chosen baselines are introduced as follows:

- **Audio module Γ_a:** The first corresponding baseline is the strip audio module from the proposed method that takes the spectrograms as input. We use the same setup to train the audio module and obtain binary classification results on the same data splits. The separate audio module takes the same transfer-learning setup as the full model.
- **Video module Γ_v:** The other baseline is the strip video module from the proposed method that takes the preprocessed frame sequences as input. We use the same adversarial training scheme and transfer-learning setup to train the video module and obtain binary classification results on the same data splits.
- **I3D:** The Two-Stream Inflated 3D ConvNet (I3D) (Carreira and Zisserman, 2017) expands filters and pooling kernels of 2D image classification ConvNets into 3D to learn spatiotemporal features from videos. It was the state-of-the-art model years ago. For our task, I3D can be inferior because the calculation of optical flow can be time-consuming and result in more noise.
- **SlowFast:** The SlowFast (Feichtenhofer et al., 2019) network is a video recognition network proposed by Facebook that involves a Slow pathway to capture spatial semantics and a Fast pathway to capture motion at fine temporal resolution. SlowFast achieves strong performance on action recognition in video and has been a powerful state-of-the-art model in recent years.
- **MMDL:** Our prior work MMDL (Yu et al., 2020) is a preliminary version of the proposed two-branch method that takes similar preprocessed frame sequences for the video branch, but text transcripts for the audio branch. The video branch uses feature difference instead of image difference (which we will ablate later), and the audio branch was an LSTM that performs text classification. Due to drastically different network structures, the two branches only have connections in the final layer using a "late-fusion" scheme.

In evaluation, we conducted both a 5-fold cross-validation experiment and a time-cutoff hold-out experiment. In the cross-validation experiment, the full dataset is split randomly and evenly into 5 folds. Each fold takes four folds for training and the rest for testing, and the model will be reset for each fold. We assess the model performance by the accuracy, specificity, sensitivity, and area under the ROC curve (AUC). In the time-cutoff hold-out experiment, we set the first recruited 168 cases as the training/validation set and the later 53 cases as the testing set. We perform a similar 5-fold cross-validation over the training/validation set and use the saved five top models from each training fold to predict the cases in the testing set. We report the mean, standard deviation, and range of the AUC as both the model performance and stability benchmark.

Because our objective is to perform stroke screening for incoming patients, the proposed methods and the baselines are compared to the triage team's performance in stroke pre-hospital screening and the ER doctors' clinical impression we obtained with the metadata. As the real stroke triage in ER often happens when there is stroke onset of the patient, the triage result was not able to be fully collected like the ER doctor's clinical impression in our dataset, due to the high risk of delayed diagnosis and IRB restrictions. The performance measurement of the triage team is from another internal study based on a much larger cohort of patients over a longer period of time.

5.2.1. Cross-Validation experiment

In the cross-validation experiment, 80% data are used for training and 20% data are used for testing. We first select one fold as the test set and train a model based on the four other folds. After the experiment using one fold as test set is completed, another folder is used as the test set and a new model is trained using the remaining four folds; note that, the new model is re-initialized when using a new test fold, to ensure the testing data is never seen by the new model. The model that achieves the best result among the first 20 epochs on the current test fold is saved, and the overall performance of the current experiment is calculated with all the five folds' testing results (which covers the whole dataset). Hyperparameters of the experimental setup are tuned based on the overall cross-validation performance on the whole dataset.

Besides contrasting the model performance with the triage team and ER doctors, we further examine the effectiveness of our proposed methods in reducing false negatives and improving stroke screening sensitivity by aligning the specificity of each method to be the same as the triage performance by changing the threshold for binary cutoffs, while checking and comparing for other measurements. The results are shown in Table 3. For better comparison, the ROC curves for the proposed model and baselines are also plotted in Fig. 5, together with the triage performance and clinical impression performance. For reference, we include the performance of a "dummy" classifier that predicts all cases to be stroke/positive.

From both Table 3 and Fig. 5, one can see that our proposed *DeepStroke* outperforms several state-of-the-art methods as well as its single module variants. Both the Audio Module Γ_a and the Video Module Γ_v achieve an AUC level of around 0.6, showing very high sensitivity while suffering from low specificity. Comparing the Video Module Γ_v with two state-of-the-art video recognition models, we can also see a much higher model performance. When specificity is aligned, Video Module Γ_v achieves 11.77% higher accuracy, 16.35% higher sensitivity, and 0.0917 higher AUC than I3D and 16.29% higher accuracy, 22.64% higher sensitivity, and 0.1357 higher AUC than SlowFast. When Γ_a and Γ_v collaborate, a decent performance gain is seen in both our previous work Multimodal Deep Learning (MMDL) and our proposed *DeepStroke*. The MMDL method achieves 0.6414 AUC; when specificity is aligned with Triage performance at 45.16%, a sensitivity of 72.96% is reported, which is 2.78% higher than the Triage performance. Compared with our prior work MMDL, *DeepStroke* achieves 0.0749 higher AUC. When specificity is aligned to Triage, sensitivity is improved by

Table 3
Results of the cross-validation experiment. Raw: results with decision threshold 0.5 after the final softmax output; aligned: results with the threshold that makes the specificity aligned with the Triage performance. Due to the dataset limitation, exact alignment is not possible. We take and report the closest possible alignment value. The best performance of each metrics among the model/baselines is highlighted in bold.

	Accuracy (%)		Specificity (%)		Sensitivity (%)		AUC
Model/Baseline	Raw	Aligned	Raw	Aligned	Raw	Aligned	–
Dummy	71.95		0.00		100.00		0.5000
Module Γ_a (Audio)	62.89	63.35	24.19	45.16	77.99	70.44	0.5998
Module Γ_v (Video)	68.32	62.90	16.13	45.16	88.68	69.81	0.6042
I3D (Video)	71.04	51.13	17.74	45.16	91.82	53.46	0.5125
SlowFast (Video)	70.13	46.61	8.01	45.16	**94.34**	47.17	0.4685
MMDL (Audio + Video)	63.80	65.16	30.65	45.16	76.73	72.96	0.6414
DeepStroke (Audio + Video)	**72.40**	**71.40**	**32.26**	45.16	88.05	**81.13**	**0.7163**
Triage Stroke Screening (Triage)	64.03		45.71		70.19		–
Clinical Impression (CI)	69.23		**51.61**		76.10		–

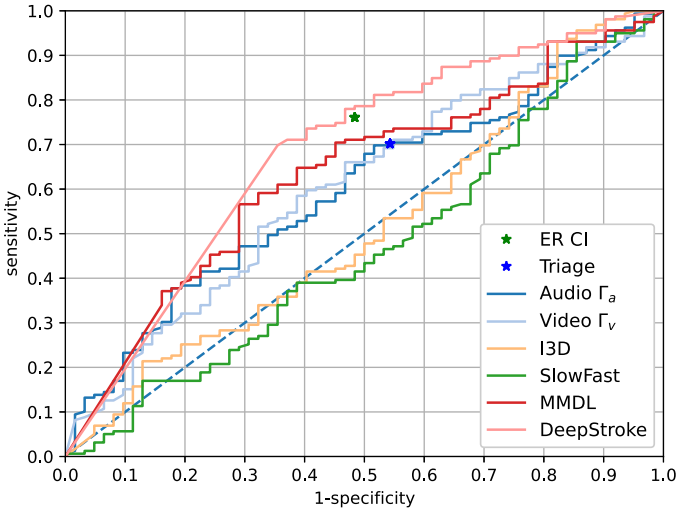


Fig. 5. The ROC curve from 5-fold cross-validation for the proposed *DeepStroke* and other baselines. The blue dot shows the performance of the triage team in stroke screening (Triage), and the green dot shows the performance of clinical impression (CI) from ER physicians. The dashed line is for a "dummy classifier" that predicts all cases to be stroke/positive. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

8.27%. This is a major improvement considering the prior work has already demonstrated good performance on the preliminary version of the dataset.

We believe the improvements came from the following aspects. First, by adopting a different audio representation and introducing the lateral connections, the proposed framework resolves the unstable convergence problem in the prior work caused by different training dynamics of branches, while also sharing low- and high-level features for a better combination of global audio context and local frame features. Secondly, our adversarial training scheme can keep the network from remembering the identity features and extract pure stroke-related features for the network to learn, which also mitigates the overfitting problem. Thirdly, introducing the transfer learning pipeline and adopting fairness-aware face pre-training mitigate the network's overfitting on facial attributes, and improves the generalizability. The three aspects are discussed in the following ablation studies (Section 5.3). Finally, in another aspect, using spectrograms instead of transcripts will maximally preserve the patterns in the original audio files since the soundwave information is fully presented without inference, addition, or deletion.

When compared with the triage team performance, the proposed *DeepStroke* shows 10.94% higher sensitivity, and 7.37% higher

Table 4
Model and baseline performance compared with ER doctor's Clinical Impression (ER CI).

Baselines	Accuracy (%)	Specificity (%)	Sensitivity (%)
Clinical Impression (CI)	69.23	51.61	76.10
DeepStroke (Spec. Aligned)	**71.04**	51.61	**78.61**
DeepStroke (Sens. Aligned)	69.68	**53.22**	76.10

accuracy while achieving the same specificity, illustrating its practicability and effectiveness. The performance is comparable to the ER doctors' clinical impressions (CI). This is also of great significance considering the doctors generally will refer to the patients' CT results, vital signals, and EHR for past stroke history. From Fig. 5, the proposed *DeepStroke* achieves close (and slightly better) performance. For better comparison, we report results when the specificity and sensitivity of *DeepStroke* aligned with ER CI in Table 4.

Summarizing Table 3, Table 4, and Fig. 5, we conclude that the proposed *DeepStroke* is powerful in performing pre-hospital stroke screening and is able to outperform traditional stroke triage. It achieves comparable or better performance with much less available information.

With new patients continually being added to the cohort, our dataset is becoming more and more diverse and even more challenging for the clinicians (for the original dataset, clinicians had 72.94% accuracy, 77.78% specificity, and 70.68% sensitivity). We infer that this is due to the addition of a number of hard cases, where the patterns for stroke are too subtle for the clinicians to capture. Even so, when we align specificity, the proposed method still outperforms the clinicians. ER doctors tend to rely more on the speech abilities of the patients and may have difficulty in cases with too subtle facial motion incoordination. We infer that the video module in our framework can detect those subtle facial motions that doctors can neglect and complement the diagnosis based on speech/audio. The drop in specificity is regarded as permissible compared to the improvements in sensitivity because, in stroke screening, failing to spot a patient with stroke (false-negative) will result in very serious results.

5.2.2. Holdout experiment

Aside from the cross-validation experiments, to ensure the clinical significance of our proposed methods, we conducted a holdout experiment. Cases in our dataset are ordered by their time of data acquisition. To maximally resemble the clinical situation, among the 221 collected cases, we set the first 168 cases for training/validation and the subsequent 53 cases for testing. This keeps the ratio of the testing set to be 25% of the dataset, while we can ensure the testing set has a similar class ratio to the train-

Table 5
Results of the hold-out experiment. The metrics are calculated based on the testing AUC of the five saved models from each fold on the aforementioned testing set. SD: standard deviation. ↑ indicate higher the value, better the performance, and ↓ means the opposite. The best values are in bold.

	Mean↑	SD↓	Range↑
Module Γ_a (Audio)	0.4631	0.1389	[0.2263, 0.6386]
Module Γ_v (Video)	0.5242	**0.0320**	[0.4895, 0.5649]
MMDL	0.6417	0.0475	[0.6018, 0.7333]
DeepStroke	**0.7214**	0.0368	**[0.6772, 0.7895]**

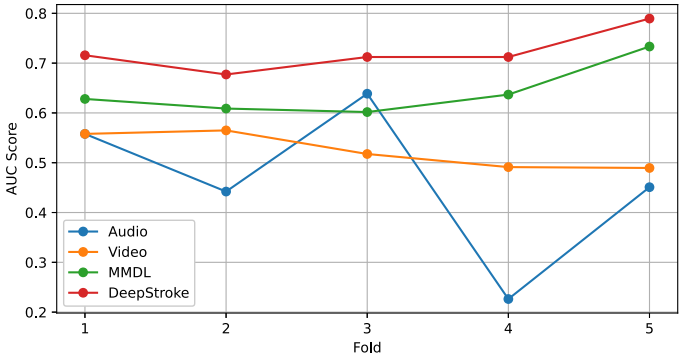


Fig. 6. Cross-fold hold-out testing AUC plot for the proposed *DeepStroke* and other baselines.

ing/validation set. In the training/validation set, we have 121 stroke cases and 47 non-stroke cases resulting in a ratio of 2.57; in the testing set, there are 38 stroke cases and 15 non-stroke cases, giving a ratio of 2.53. In this holdout study, we test the best models acquired from 5-fold cross-validation (with training/validation set) on the *unseen* hold-out testing data. We refer to AUC as the saved model performance benchmark on the testing data. Because each of the five folds will save the best model, we report the mean, standard deviation, and range of the AUC across these models. We also plot the cross-fold AUC for our proposed methods and baselines to demonstrate the stability of our method. Note that due to the apparent low performance of the state-of-the-art video baselines, we omit I3D and SlowFast in the holdout experiment.

As shown in Table. 5, the *DeepStroke* framework can achieve stable and high performance in our stroke screening task. A steady improvement from our previous work (MMDL) is observed. To better present the results, we plot the testing AUC of different models in Fig. 6.

The testing performance from any of the folds generated with the *DeepStroke* is stable and high. With audio alone, the test performance has seen large variations across different fold models. This is not presented in the cross-validation experiment above. In some folds (*i.e.,* fold #4 in Fig. 6), the saved model was performing the best in the validation set but is giving a very low AUC score on the testing set. We infer that merely relying on the patient's speech audio is insufficient to recognize stroke patterns, as the audio input includes a large number of unused patterns and noises. The video data, on the other hand, performs relatively stable yet has been unable to recognize non-stroke cases. The collaboration of the audio module and the video module, as both the previous work and the proposed *DeepStroke* demonstrates, achieves noticeable performance gain. Furthermore, with the proposed improvements over the previous work, *DeepStroke* can perform stroke screening with higher stability and better performance.

Table 6
Ablation study results. The first block of rows ablates the lateral connection we designed over the model components, the second block ablates the other choice of making frame pair difference, and the third block ablates other choices of extracting encoded features for the discriminator. Experiments share the same setup as the cross-validation experiments in Section 5.2.1 and parameters are tuned separately.

Baselines	Accuracy (%)	Specificity (%)	Sensitivity (%)	AUC
w/o transfer	62.44	37.10	72.33	0.5611
w/o lateral	71.49	37.10	84.91	0.6881
w/o adversarial	68.87	17.78	90.57	0.5568
Conv-Diff	69.78	25.80	85.53	0.6680
Feature #1	69.23	30.64	84.28	0.6679
Feature #2	69.23	25.80	86.16	0.6690
Feature #3	69.23	16.13	89.94	0.6678
DeepStroke	72.40	32.26	88.05	**0.7163**

5.3. Ablation study

5.3.1. Model designs

To evaluate the effectiveness of different designs in our proposed model, we perform the following ablation studies. Note that the reported statistics are based on the 5-fold cross-validation results on the same dataset stated above. Here, we denote *DeepStroke* as the full model that introduced lateral connection, adversarial traning, and transfer learning to the network introduced in our previous work.

- **Model without transfer learning (w/o transfer):** We demonstrate the value of our transfer learning setup by contrasting it with a baseline model that takes the same network design as the *DeepStroke* but does not freeze parameters. The baseline model discards 1) FairFace as face video network backbone pre-training and 2) ImageNet as audio network pre-training, and was trained from scratch.
- **Model without lateral fusion (w/o lateral)**: The lateral connection we designed for the two branches was updated from the late fusion scheme in our prior work to connect the two branches at different levels. An experiment was made to only use the late fusion as a comparison to the full proposed model to show the power of the lateral fusion.
- **Model without adversarial training (w/o adversarial)**: The adversarial training scheme aims to extract stroke-related identity-free features. We train the same model without the adversarial scheme to see the improvements.
- **Frame difference (Conv-diff)**: In the proposed method, we set the two consecutive frames to form a difference at the very beginning of the network, but such subtraction can be made elsewhere–in the middle of the network or at the back of the network (*i.e.*, the deep features generated from the framework). An extra experiment is conducted on the frame subtraction scheme that takes the difference of features after one Conv layer as used in our prior work.
- **Discriminator feature input stage (Feature #)**: When adversarial training for distilling the subjects' identities, our network takes the encoded features after all four blocks of ResNet. We have other stages where features are extracted for the adversarial training, (*i.e.*, after the first block, after the third block, and after the third layer).

From Table 6, we could conclude that the lateral connection, transfer learning, and adversarial training have been improving the performance of the framework as expected. The lateral connection improves the overall statistics with 0.0282 in AUC, indicating a better learning outcome that may result from better balanced two-branch learning dynamics. By applying transfer learning, we noticed a clear performance gain in terms of sensitivity of 15.72% and a 0.1552 improvement in AUC. The FairFace pre-training reduces

Table 7
Validation results for African American cases when training the framework on other cases. The experiments are trained on 149 other cases while 72 African American cases are used as validation data. Best test performance on African American test set in 20 epochs is reported.

Experiment	Accuracy (%)	Specificity (%)	Sensitivity (%)	AUC
DeepStroke	70.83	50.00	77.77	0.6967
w/o FairFace	66.66	38.89	75.93	0.6821
w/o Transfer Learning	63.88	61.11	64.81	0.6255

Table 8
Validation results for different attribute groups. The number in the bracket indicates the total number of cases in the group. Some subjects opt-out for ethnicity or race questions and are excluded. Model: *DeepStroke*, CI: Clinical Impression. Values in **red** are considered to yield fairness concerns.

		Accuracy (%)		Specificity (%)		Sensitivity (%)		AUC	
Group	Attribute	Model	CI	Model	CI	Model	CI	Model	CI
Gender	Male (108)	75.00	66.39	34.38	60.00	92.10	69.05	0.6772	–
	Female (113)	70.80	66.11	25.00	48.49	83.15	72.72	0.6905	–
Age	$\leq$ 65 y/o (120)	72.50	67.67	40.00	54.76	83.33	73.63	0.7048	–
	$\geq$ 65 y/o (101)	73.27	64.49	19.23	53.84	92.00	67.90	0.6728	–
Ethnicity	Hispanic (16)	75.00	88.23	40.00	75.00	90.90	92.31	0.8545	–
	Non-Hispanic (198)	72.22	65.27	28.57	55.17	86.58	68.99	0.66.15	–-
Race	African American (72)	69.44	61.84	23.08	45.00	79.66	67.86	**0.5580**	–
	Other (149)	74.50	68.29	32.56	58.33	91.51	72.41	0.7411	–
All subjects (221)		72.40	69.23	32.26	51.61	88.50	76.10	0.7163	–

the potential biases from facial attributes so that the network has less reference to such features when making a decision. The adversarial training also drastically improved the specificity (14.48% as shown) and resulted in a better learning result (0.1483 improvements in AUC), and we believe the features for stroke are more clearly revealed with this training scheme. Our choice of features out of different network layers also demonstrated the effectiveness of using final concatenated features for the use of a discriminator.

5.3.2. Model fairness

We also examined the discrepancy between different genders, races/ethnicity, and age groups in our dataset. Specifically, we examine the following attribute groups in Table 8 and report the validation results within the 5-fold cross-validation experiment result on the proposed model:

- **Gender:** Previous studies point out that gender discrepancy is noticeable for stroke: women show unique patterns that are often neglected (Colsch and Lindseth, 2018). The probability of men getting misdiagnosed with stroke is lower than women (Newman-Toker et al., 2014). To compare how clinicians and our model perform regarding this discrepancy, we report results on different gender groups.
- **Age:** The risk for stroke increases with age. According to national statistics, the majority of patients hospitalized for stroke are over 65 years old (Hall et al., 2012). This prior is not clearly shown in our dataset, partly because some older patients are having more serious conditions and are not enrolled in this study.
- **Ethnicity:** According to Trimble and Morgenstern (2008), the incidence of stroke in Hispanics is significantly higher than in Non-Hispanic whites. Hispanics are also believed to have a higher rate of stroke recurrence (Sheinart et al., 1998). Note that due to the relatively limited Hispanic participants in our current version of the dataset, the comparison may not be indicative enough for the discrepancy.
- **Race:** African Americans have nearly twice as high a risk of first stroke compared to others (Virani et al., 2020) and the highest mortality rate (Centers for Disease Control and Prevention, 2019). African Americans also receive less evidence-based care and have a higher risk of stroke recurrence (Schwamm et al., 2010). Another potential issue is concerned with the biases in current face-related computational models that are mainly trained on image/video samples from non-African-American participants.

Table 8 shows that the validations on most of the attribution groups result in a relatively stable AUC (variations are expected considering the relatively small number of samples). For the attribute of gender and ethnicity, the variation among different groups can be regarded as reasonable. The results on African American subjects raise some concern as the AUC falls to 0.5580, indicating that the model trained on the complete dataset is not picking up the African American cases very well. This is likely due to the larger imbalance of positive/negative cases among African American groups compared to the complete dataset (54 stroke cases, 18 non-stroke cases). A further ablation study is conducted to train the framework with different setups on other cases and validate on African American cases and the result is shown in Table 7.

In this hold-out experiment, we notice that the model performance is respectable when trained with no African American cases but validated only on African American cases. Using the full *DeepStroke*, the validation performance is at 0.6967 AUC. We then replaced the FairFace pre-training with ImageNet pre-training and obtained a slightly worse result with 0.6821 AUC. If trained from scratch, *i.e.,* without transfer learning, the model performance dropped vastly to 0.6255 AUC. From the experimental results in Table 7, we can conclude that, by adopting the transfer learning scheme and FairFace pre-training, *DeepStroke* can better reduce facial attribute-related biases and improve the generalizability of the method.

Another potentially biased attribute group is age. Age is a known factor for the risk of stroke and with higher age, the probability of stroke is greatly increased. Both the model and clinical impression have a rather large discrepancy in specificity between the two age groups, and cases younger than 65 years old are much higher in specificity. To further look into and attempt to understand this discrepancy, we first train on younger cases and validate

Table 9
Ablation study on age attribute. The first row is trained on younger cases ($\leq$ 65 y/o) and tested on older cases ($\geq$ 65 y/o), and the second row is trained on older cases and trained on younger cases.

Experiment	Accuracy (%)	Specificity (%)	Sensitivity (%)	AUC
$\leq$ 65 y/o	68.31	30.77	81.33	0.6203
$\geq$ 65 y/o	68.33	38.89	80.95	0.6257

on older cases, and then do the reverse. The results are shown in Table 9.

The result shows a similar classification accuracy at 0.62 AUC and is lower than training on all data, which partly indicates that the network may be able to extract useful age information from the patients and use it as a feature for prediction. We infer that the relatively lower specificity and high sensitivity among the age group over 65 years old is also a consequence of heavier class imbalance (79 stroke cases, 26 non-stroke cases), and the network prefers to give a positive prediction on patients in the higher age band. This could reduce the risk of missing a real stroke case in the ER and could potentially save the patient.

6. Discussion

We analyze the running time of the proposed approach. The recording runs for a minute, the extraction of audio and generation of spectrograms takes an extra minute, and the video processing is completed in three minutes. The prediction with the deep models can be achieved within half a minute on a desktop with a mid-level GPU (NVIDIA GTX1070). Therefore, the evaluation process takes no more than six minutes per case. More importantly, the process is almost running at zero external cost and would be contactless, not harming the patients with equipment or by radiation. Considering a complete MRI scan will take more than an hour to perform, a specialized device to run, and hundreds of dollars charged, the proposed method is ideal for performing both cost and time-efficient pre-hospital stroke assessments in an emergency setting.

It is worth mentioning that although our proposed method is demonstrated to outperform triage stroke screening performance by a margin and is comparable to or better than ER doctors who have access to richer patient data, its use should be limited to that of an AI assistant during the ER triage stage to help to detect non-obvious strokes and mitigate the risk of misdiagnosing. The CT scan should be taken as the next step to identify the stroke subtypes and estimate damaged regions for the reference of doctors in performing interventions.

We expect that our proposed approach will be clinically relevant and can be deployed effectively on smartphones for fast and accurate assessment of stroke by ER doctors, tele-stroke neurologists, at-risk patients, or caregivers. If the approach is further optimized and deployed onto a smartphone, we can perform the spatiotemporal face frame proposal and speech audio processing on the phone. Cloud computing can be leveraged to perform the prediction in no more than a minute after the frames are compressed and uploaded. In such a case, the total time for one assessment should be within a few minutes. Moreover, with minimal user education required, such a framework can allow for the patients' self-assessments even before the ambulance arrives. With different labeling of data, the pipeline is also valuable in the screening of other oral-facial neurological conditions.

We also noticed that when clinicians are performing stroke screening in the ER, they would consider race, ethnicity, and age as factors, and they may also refer to the patients' EHR for previous stroke records. Such information is missing from our dataset. While we would like to find patterns directly from facial motions and speaking voices, these factors are certainly beneficial to be adopted as auxiliary information that may guide the framework to project different probabilities of stroke for different demographic groups referring to the prior distributions.

Through our experiment, we noticed that the dataset we have been collecting for the pilot study purposes is still preliminary and small, especially for the evaluation of various demographic subgroups. Also, with a limited number of cases available, the dataset could be insufficient to cover all possible patterns of stroke in facial motions and speech. The limitation also leads to the performance discrepancy in the African American cases and among age-related attribute groups. We are working towards a large-scale clinical trial that targets about 1000 cases to evaluate the proposed method and further improve its robustness. When enough data is available, we would also pick a verified set of held-out test data to cover expected variations and use it as the public evaluation standard for this task.

To compensate for the relatively small data size, while we have been trying to expand our dataset, we also considered two data augmentation schemes and attempted to apply them to the framework, including:

- We recruited Amazon MTurk independent contractors to perform the same speech tasks and use these data to augment the non-stroke cases. For future study purposes, we collected more than enough data to balance the minority class. We trained an audio ResNet-18 (audio module) based on spectrograms and noticed higher testing performance and better stability of the augmented audio module in the hold-out experiment. However, when the balanced-data-trained models were set as transfer learning starting points for the audio sub-network, testing AUC of the whole model (including both audio- and video- modules) drastically dropped. We suspect this is because audio pretrained model augmented by normal audio makes the whole model less generalizable.
- We used one of the best-performing facial motion transfer algorithms (Wang et al., 2021) and attempted to transfer facial motions from non-stroke patients to public online faces, and pair the transferred facial movies with the collected MTurk audio to generate synthetic negative cases. However, the addition of such cases did not help improve the network performance. We noticed that the motion transfer constantly fails to preserve stroke-indicative motions and introduces more artifacts, mainly due to wrinkles and other uncommon facial patterns.

We will further investigate other possibilities of data augmentation for improving the framework's performance

We recognize other limitations of our work. First, the spatiotemporal facial frame proposal method is preliminary and is unable to eliminate all non-facial motions. This induces errors in the captured facial motion features. Also, the state-of-the-art 2D facial landmark tracking and pose estimation algorithms still induce considerable error in the pipeline. A plan is to collect 3D data instead of 2D so that we could achieve accurate alignment and reconstruction of subjects' faces. With an accurate mesh representation of the subjects' facial motions, we could model the motions of different facial regions and correlate them to different areas of brain lesions. The representation could also help reduce color bias and naturally remove identity from subjects, which is promising. Second, the computational workload of the whole pipeline is still heavy for mobile deployment. In conditions with low network bandwidth, the transfer of collected data could be a serious bottleneck and hinder the efficiency of the screening process. Future work would need to address these issues and improve the framework for even better clinical performance.

As a future work, we intend to develop *DeepStroke+* that resolves the existing issues of our current framework and functions

as a full stroke diagnosis AI assistant to span the stroke triage till discharge and treatment. *DeepStroke+* will be based on the much larger dataset collected through the clinical trial and adopt cloud/mobile computing for efficient, mobile assessment. The anticipated model will be taking 3D facial data uploaded with mobile portals and linked to the patient's EHR, demographics, vitals, and CT report as multi-modal input. *DeepStroke+* will be developed as a multi-task network to support the additional decision of stroke subtypes and lesion regions and provide detailed reports to the ER clinicians.

7. Conclusion

In this work, we presented a novel multi-modal deep learning framework, *DeepStroke*, for on-site clinical triage of stroke in an ER setting. Our framework can perform accurate and efficient stroke screening based on the abnormalities in the patient's speech ability and facial muscular movements. We constructed a dual branch deep neural network for the classification of patient facial video frame sequences and speech audio as spectrograms to capture subtle stroke patterns from both modalities. Experiments on our collected clinical dataset with real, diverse, "in-the-wild" ER patients demonstrated that the proposed approach not only outperforms the traditional stroke triage with a 10.94% higher sensitivity rate and 7.37% higher accuracy when specificity is aligned but also outperforms well-trained ER clinicians with more information available. Also, ablation studies validated the value of our multi-modal lateral fusion method, transfer learning pipeline, and adversarial training scheme, which collaboratively improve our proposed model from our prior work. The *DeepStroke* has also been verified to be efficient and provide a screening result for the reference of clinicians in minutes.

Before a clinical deployment, we will continue to conduct the clinical trial, enlarge the dataset, improve the efficiency of the pipeline, and develop the current method into a 3D mobile framework. Future efforts may also try to address the potential biases in the current model and apply the method to other neurological conditions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

T. Cai, H. Ni, M. Yu, X. Huang, and J.Z. Wang were supported by the College of Information Sciences and Technology at The Pennsylvania State University. T. Cai, K. Wong, and S.T.C. Wong were supported by the T.T. and W.F. Chao Foundation, the John S. Dunn Research Foundation, and The Scurlock Foundation. The computation was supported by the NVIDIA Corporations GPU Grant Program and the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant number ACI-1548562. The authors would like to thank Amber Criswell for her help in patient enrollment and Xiaohui Yu for mobile app development. The authors would also like to acknowledge the comments and constructive suggestions from the reviewers and the associate editor.

References

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D.L., Erickson, B.J., 2017. Deep learning for brain MRI segmentation: state of the art and future directions . J. Digit. Ima ging 30 (4), 449–459.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al., 2016. Deep speech 2: end–to-end speech recognition in English and Mandarin. In: Proceedings of the International Conference on Machine Learning, pp. 173–182.

Anping, S., Guoliang, X., Xuehai, D., Jiaxin, S., Gang, X., Wu, Z., 2017. Assessment for facial nerve paralysis based on facial asymmetry . Austral. Phys . Eng. Sci. Med. 40 (4), 851–860.

Arch, A.E., Weisman, D.C., Coca, S., Nystrom, K.V., Wira III, C.R., Schindler, J.L., 2016. Missed ischemic stroke diagnosis in the emergency department by emergency medicine and neurology services. Stroke 47 (3), 668–673.

Bandini, A., Orlandi, S., Escalante, H.J., Giovannelli, F., Cincotta, M., Reyes-Garcia, C.A., Vanni, P., Zaccara, G., Manfredi, C., 2017. Analysis of facial expressions in parkins on's disease through video-based automatic methods. J. Neurosci. Methods 281, 7–20.

Bandini, A., Orlandi, S., Giovannelli, F., Felici, A., Cincotta, M., Clemente, D., Vanni, P., Zaccara, G., Manfredi, C., 2016. Markerless analysis of articulatory movements in patients with Parkinson's disease . J. Voice 30 (6), 766. e–766.e11.

Banks, C.A., Bhama, P.K., Park, J., Hadlock, C.R., Hadlock, T.A., 2015. Clinician-graded electronic facial paralysis assessment: The eFACE . Plast. Reconstr. Surg. 136 (2), 223e–230e.

Bevilacqua, V., D'Ambruoso, D., Mandolino, G., Suma, M., 2011. A new tool to support diagnosis of neurological disorders by means of facial expressions. In: Proceedings of the IEEE International Symposium on Medical Measurements and Applications. IEEE, pp. 544–549.

Cai, J., Meng, Z., Khan, A.S., Li, Z., O'Reilly, J., Tong, Y., 2019. Identity-free facial expression recognition using conditional generative adversarial network. arXiv preprint arXiv:1903.08051.

Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308.

Centers for Disease Control and Prevention, 2005. About Stroke. CDC Stroke. Available at: https://www.cdc.gov/stroke/about.htm (Accessed: 12 Mar 2021).

Centers for Disease Control and Prevention, 2019. Underlying Cause of Death, 1999–2018. CDC *WONDER* Online Database. Available at: https://wonder.cdc.gov/ucd-icd10.html (Accessed: 12 Mar 2021).

Centers for Disease Control and Prevention, 2020. Stroke Facts. CDC Stroke. Available at: https://www.cdc.gov/stroke/facts.htm (Accessed: 12 Mar 2021).

Colsch, R., Lindseth, G., 2018. Unique stroke symptoms in women: a review. J. Neurosci. Nurs. 50 (6), 336–342.

Crichton, S.L., Bray, B.D., McKevitt, C., Rudd, A.G., Wolfe, C.D., 2016. Patient outcomes up to 15 years after stroke: survival, disability, quality of life, cognition and mental health . J. Neurol. Neuro surg. Psychiatry 87 (10), 1091–1098.

Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M., 2017. ECO: Efficient convolution operators for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6638–6646.

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, Li Fei-Fei, 2009. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.

Dong, J., Lin, Y., an Liu, L., Ma, L., Wang, S., 2008. An approach to evaluation of degree of facial paralysis based on image processing and pattern recognition. Journal of Information & Computational Science 5 (2), 639–646.

Eitel, A., Springenberg, J.T., Spinello, L., Riedmiller, M., Burgard, W., 2015. Multimodal deep learning for robust RGB-D object recognition. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 681–687.

Feichtenhofer, C., Fan, H., Malik, J., He, K., 2019. SlowFast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211.

Flores-Mondragón, G., Paredes-Espinoza, M.A., Hernández-Campos, N.A., Sánchez-Chapul, L., Paniagua-Pérez, R., Martínez-Canseco, C., Renán-León, S., Araujo-Monsalvo, V.M., Perea-Paz, J.M., Flores-Jacinto, A., 2015. Facial anthropometry: a tool for quantitative evaluation in patients wi th peripheral facial paralysis. Int. J. Sci. Eng. Res. 6, 1657–1660.

Frey, M., Michaelidou, M., Tzou, C.-H., Pona, I., Mittlböck, M., Gerber, H., Stüssi, E., 2008. Three-dimensional video analysis of the paralyzed face reanimated by cross-face nerve grafting and free gracilis muscle transplantation: quantification of the functional outcome. Plast. Reconstr. Surg. 122 (6), 1709–1722.

Giles, E., Patterson, K., Hodges, J.R., 1996. Performance on the Bosto n Cookie Theft picture description task in patients with early dementia of the alzheimer's type: missing information. Aphasiology 10 (4), 395–408.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. arXiv preprint arXiv:1406.2661.

Green, R.D., Guan, L., 2004. Quantifying and recognizing human movement patterns from monocular video im ages-part ii: applications to biometrics. IEEE Trans. Circuits Syst. Video Technol. 14 (2), 191–198.

Greene, J.J., Guarin, D.L., Tavares, J., Fortier, E., Robinson, M., Dusseldorp, J., Quatela, O., Jowett, N., Hadlock, T., 2020. The spectrum of facial palsy : the MEEI facial palsy photo and video standard set. Laryngoscope 130 (1), 32–37.

Guo, Z., Shen, M., Duan, L., Zhou, Y., Xiang, J., Ding, H., Chen, S., Deussen, O., Dan, G., 2017. Deep assessment process: Objective assessment process for unilateral peripheral facial paralysis via deep convolutional neural network. In: Proceedings of the 2017 IEEE International Symposium on Biomedical Imaging, pp. 135–138.

Hakata, M., Seo, M., Chen, Y., Matsushiro, N., 2013. Facial paralysis modeling based on image morphing. In: Proceedings of the 6th International Conference on Biomedical Engineering and Informatics, pp. 806–810.

Hall, M.J., Levant, S., DeFrances, C.J., 2012. Hospitalization for Stroke in US hospitals, 1989–2009. US Department of Health and Human Services, Centers for Disease Control and Prevention.
Hamm, J., Kohler, C.G., Gur, R.C., Verma, R., 2011. Automated facial action coding system for dyn amic analysis of facial expressions in neuropsychiatric disorders. J. Neurosci. Method. 200 (2), 237–256.
Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al., 2014. Deep speech: scaling up end–to-end speech recognition. arXiv preprint arXiv:1412.5567.
Harbison, J., Hossain, O., Jenkinson, D., Davis, J., Louw, S.J., Ford, G.A., 2003. Diagnostic accuracy of stroke referrals from primary care, emergency room physicians, and ambulance staff using the face arm speech test. Stroke 34 (1), 71–76.
He, S., Soraghan, J.J., O'Reilly, B.F., 2007. Automatic motion feature extraction with application to quantitative assessment of facial paralysis. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 1, p. 444.
He, S., Soraghan, J.J., O'Reilly, B.F., 2008. Supervised local linear embedding (SLLE) for facial paralysis image sequence analysis. In: Proceedings of the IEEE International Conference on Multimedia and Expo. IEEE, pp. 49–52.
House, J.W., Brackmann, D.E., 1985. Facial nerve grading system . Otolaryngolog y 93 (2), 146–147.
Hsu, G.J., Huang, W., Kang, J., 2018. Hierarchical network for facial palsy detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 693–699.
Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., Lungren, M.P., 2020. Fusion of medical imaging an d electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digital Medicine 3 (1), 1–9.
Johnson, W., Onuma, O., Owolabi, M., Sachdev, S., 2016. Stroke: a global response is needed. Bull. World Health Organ. 94 (9), 634.
Kahou, S.E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., Jean, S., Froumenty, P., Dauphin, Y., Boulanger-Lewandowski, N., et al., 2016. Emonets: multimodal deep learning approaches for emotion recognition in video. J. Multi. User Interface. 10 (2), 99–111.
Karkkainen, K., Joo, J., 2021. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1548–1558.
Kihara, Y., Duan, G., Nishida, T., Matsushiro, N., Chen, Y.-W., 2011. A dynamic facial expression database for quantitative analysis of facial paralysis. In: Proceedings of the 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT). IEEE, pp. 949–952.
King, D.E., 2009. Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. 10, 1755–1758.
Kothari, R.U., Pancioli, A., Liu, T., Brott, T., Broderick, J., 1999. Cincinnati prehospital stroke scale: repr oducibility and validity. Ann. Emerg. Med. 33 (4), 373–378.
Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., Yang, M.-H., 2018. Diverse image–to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 35–51.
Leira, E.C., Kaskie, B., Froehler, M.T., Adams Jr, H.P., 2013. The growing shortag e of vascular neurologists in the era of health reform: planning is brain!. Stroke 44 (3), 822–827.
Li, Y., Shen, L., 2018. Deep learning based multimodal brain tumor diagnosis. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, Cham, pp. 149–158.
Liang, M., Li, Z., Chen, T., Zeng, J., 2015. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach . IEEE/ACM Tran s. Comput. Biol. Bioinf. 12 (4), 928–937.
Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400.
Liu, Y., Wei, F., Shao, J., Sheng, L., Yan, J., Wang, X., 2018. Exploring disentangled feature representation beyond face identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2080–2089.
Lu, B., Chen, J.-C., Chellappa, R., 2019. Unsupervised domain-specific deblurring via disentangled representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10225–10234.
Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I., 2010. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPRW), pp. 94–101.
Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2794–2802.
Menegotto, A.B., Lopes Becker, C.D., Cazella, S.C., 2020. Computer-aided hepatocarcinoma diagnosis using multimodal deep learning. In: Novais, P., Lloret, J., Chamoso, P., Carneiro, D., Navarro, E., Omatu, S. (Eds.), Ambient Intelligence – Software and Applications –,10th International Symposium on Ambient Intelligence. Springer International Publishing, Cham, pp. 3–10.
Meng, Z., Liu, P., Cai, J., Han, S., Tong, Y., 2017. Identity-aware convolutional neural network for facial expression recognition. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), pp. 558–565.
Newman-Toker, D.E., Moy, E., Valente, E., Coffey, R., Hines, A.L., 2014. Missed diagnosis of stroke in the emergency department: a cross-sectional analysis of a large population-based sample. Diagnosis 1 (2), 155–166.
Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y., 2011. Multimodal deep learning. In: Proceedings of the 28th International Conference on International Conference on Machine Learning, pp. 689–696.
Ngo, T.H., Seo, M., Matsushiro, N., Xiong, W., Chen, Y.-W., 2016. Quantitative analysis of facial paralysis based on limited-orientation modified circular gabor filters. In: Proceedings of the 23rd International Conference on Pattern Recognition (ICPR). IEEE, pp. 349–354.
NIH, 2003. NIH Stroke Scale. Available at: https://www.stroke.nih.gov/resources/scale.htm (Accessed: 12 Mar 2021).
Parra-Dominguez, G.S., Sanchez-Yanez, R.E., Garcia-Capulin, C.H., 2021. Facial paralysis detection on images using key point analysis . Appl. Sci. 1 1 (5), 2435.
Perveen, N., 2019. Facial paralysis dataset. IEEE Dataport. Available at: https://dx.doi.org/10.21227/6dsz-7d76 (Accessed: 12 Mar 2022).
Radford, A., Metz, L., Chintala, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.
Rafay, M.F., Pontigon, A.-M., Chiang, J., Adams, M., Jarvis, D.A., Silver, F., MacGregor, D., deVeber, G.A., 2009. Delay to diagnosis in acute pediatric arterial ischemic stroke. Stroke 40 (1), 58–64.
Rogers, C.R., Schmidt, K.L., VanSwearingen, J.M., Cohn, J.F., Wachtman, G.S., Manders, E.K., Deleyiannis, F.W.-B., 2007. Automated facial image analysis: det ecting improvement in abnormal facial movement after treatment with botulinum toxin a. Ann Plast Surg 58 (1), 39–47.
Ross, B.G., Fradet, G., Nedzelski, J.M., 1996. Development of a sensitive clinical facial grading system . Otolaryngology 114 (3), 380–386.
Schwamm, L.H., Reeves, M.J., Pan, W., Smith, E.E., Frankel, M.R., Olson, D., Zhao, X., Peterson, E., Fonarow, G.C., 2010. Race/ethnicity, quality of care, and outcomes in ischemic stroke. Circulation 121 (13), 1492.
Sheinart, K., Tuhrim, S., Horowitz, D., Weinberger, J., Goldman, M., Godbold, J., 1998. Stroke recurrence is more frequent in Blacks and Hispanics. Neuroepidemiology 17 (4), 188–198.
Song, A., Wu, Z., Ding, X., Hu, Q., Di, X., 2018. Neurologist standard classification of facial nerve paralysis with deep neural networks. Future Internet 10 (11), 111.
Stevens, S.S., Volkmann, J., Newman, E.B., 1937. A scale for the measurement of the psychological magnitude pitch . J. Acoust. S oc. Am. 8 (3), 185–190.
Storey, G., Jiang, R., 2018. Face symmetry analysis using a unified multi-task CNN for medical applications. In: Proceedings of SAI Intelligent Systems Conference. Springer, pp. 451–463.
Szczapa, B., Daoudi, M., Kacem, A., Guerreschi, P., Gebert, L., Alvarez-Paiva, J.C., 2019. 2D landmark-based facial asymmetry assessment in the clinical case of facial paralysis. In: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, pp. 1–5.
Thevenot, J., López, M.B., Hadid, A., 2017. A survey on computer vision for assistive medical diagnosis from faces . IEEE J. Biomed. Health Inform. 22 (5), 1497–1511.
Tran, L., Yin, X., Liu, X., 2017. Disentangled representation learning GAN for pose-invariant face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, pp. 1415–1424.
Trimble, B., Morgenstern, L.B., 2008. Stroke in minorities . Neurol Clin 26 (4), 1177–1190.
Virani, S.S., Alonso, A., Benjamin, E.J., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Delling, F.N., et al., 2020. Heart disease and stroke st atistics–2020 update: areport from the American Heart Association. Circulation 141 (9), e139–e596.
Vásquez-Correa, J.C., Arias-Vergara, T., Orozco-Arroyave, J.R., Eskofier, B., Klucken, J., Nöth, E., 2019. Multimodal assessment of Parkinson's disease: adeep learning app roach. IEEE J. Biomed. Health Inform. 23 (4), 1618–1630.
Wang, C., Wang, S., Liang, G., 2019. Identity- and pose-robust facial expression recognition through adversarial feature learning. In: Proceedings of the 27th ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, pp. 238–246.
Wang, T., Dong, J., Sun, X., Zhang, S., Wang, S., 2014. Automatic recognition of facial movement for paralyzed face . Biomed. Mater. Eng. 24 (6), 2751–2760.
Wang, T., Zhang, S., Dong, J., Liu, L., Yu, H., 2016. Automatic evaluation of the degree of facial nerve paralysis . Multimed. Tool. Ap pl. 75 (19), 11893–11908.
Wang, T.-C., Mallya, A., Liu, M.-Y., 2021. One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition, pp. 10039–10049.
Xiao, F., Lee, Y.J., Grauman, K., Malik, J., Feichtenhofer, C., 2020. Audiovisual slowfast networks for v ideo recognition. arXiv preprint arXiv:2001.08740.
Xu, T., Zhang, H., Huang, X., Zhang, S., Metaxas, D.N., 2016. Multimodal deep learning for cervical dysplasia diagnosis. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 115–123.
Xue, Y., Xu, T., Zhang, H., Long, L.R., Huang, X., 2018. SegAN: adversarial network wi th multi-scale L1 loss for medical image segmentation. Neuroinformatics 16 (3), 383–392.
Yao, J., Xu, Z., Huang, X., Huang, J., 2018. An efficient algorithm for dynamic MRI using low-rank and total variation regularizations . Med. Image Anal. 44, 14–27.
Yu, M., Cai, T., Huang, X., Wong, K., Volpi, J., Wang, J.Z., Wong, S.T.C., 2020. Toward rapid stroke diagnosis with multimodal deep learning. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention. Springer International Publishing, Cham, pp. 616–626.
Zhang, Z., Tran, L., Yin, X., Atoum, Y., Liu, X., Wan, J., Wang, N., 2019. Gait recognition via disentangled representation learning. In: Proceedings of the IEEE/CVF Internal Conference on Computer Vision and Pattern Recognition, pp. 4710–4719.
Zhuang, Y., McDonald, M.M., Aldridge, C.M., Hassan, M.A., Uribe, O., Arteaga, D., Southerland, A.M., Rohde, G.K., 2021. Video-based facial weakness analysis . IEEE Trans. Bio med. Eng. 68 (9), 2698–2705.

Zhuang, Y., Uribe, O., Mcdonald, M., Lin, I., Arteaga, D., Dalrymple, W., Worrall, B., Southerland, A., Rohde, G., 2018. Pathological facial weakness detection using computational image analysis. In: Proceedings of the IEEE 15th International Symposium on Biomedical Imaging. IEEE, pp. 261–264.
Zhuang, Y., Uribe, O., McDonald, M., Yin, X., Parikh, D., Southerland, A., Rohde, G., 2019. F-DIT-V: An automated video classification tool for facial weakness detection. In: Proceedings of the 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). IEEE, pp. 1–4.

Tongan Cai is a Ph.D. candidate in the College of Information Sciences and Technology at the Pennsylvania State University, advised by James Z. Wang. His research interests are medical image analysis, health informatics, computer vision and data mining. Before Penn State, Tongan received his BSE in Data Science from the University of Michigan and BSE in Electrical and Computer Engineering from Shanghai Jiao Tong University. He was a student member of the Michigan integrated Center for Health Analytics and Medical Prediction and worked for the Medical AI Lab for Tencent USA.

Haomiao Ni is currently pursuing the Ph.D. degree in the College of Information Sciences and Technology at the Pennsylvania State University. He received his bachelor's degree in Information Engineering from School of Electronic and Information Engineering, South China University of Technology in 2016, and his master's degree in Computer Science and Technology from Institute of Computing Technology, Chinese Academy of Sciences in 2019. His research interest broadly lies in machine learning and its application on computer vision, especially on medical image analysis and video processing.

Mingli Yu received the B.S. degree in computer science from Tsinghua University and the M.S. degree in computer science from Pennsylvania State University, where he is currently pursuing the Ph.D. degree in computer science, under the supervision of Prof. Thomas La Porta. His research interests include computer networking, network security, and machine learning.

Xiaolei Huang is an Associate professor in the College of Information Sciences and Technology at the Pennsylvania State University. Her research interests lie at the intersection of biomedical image analysis, computer vision, and machine learning, focusing on methods for image segmentation, image synthesis, object recognition, computer-assisted diagnosis, among others. She has over 160 publications and holds 7 patents in these areas. She is an associate editor for the Computer Vision and Image Understanding journal. She received her bachelor's degree in computer science from Tsinghua University, and her master's and doctoral degrees in computer science from Rutgers University.

Kelvin Wong is an Associate Professor of Radiology and Neurological Surgery at Weill Cornell Medical College. He received B.Eng., M.Phil., and Ph.D. in Electrical and Electronic Engineering degrees from the University of Hong Kong. He received his postdoctoral training in neuro-informatics in Brigham and Women's Hospital and Harvard Medical School in 2007. He was a visiting scholar at the Electrical Engineering Department, and a visiting staff associate at the Radiology Department of Columbia University of New York. He joined Houston Methodist as a faculty member since 2007 and was a faculty in Weill Cornell Medical College since 2008.

John Volpi is a Phi Beta Kappa graduate of the University of Texas at Austin, who completed his medical school and adult neurology residency at Baylor College of Medicine where he served as Chief Resident in 2007. He currently serves as the Director of the Houston Methodist Eddy Scurlock Stroke Center and Vice Chair for the Department of Neurology. Dr. Volpi's research centers on innovations in the management and prevention of ischemic stroke, and the critical care of neurological illnesses patients. He has conducted many national clinical trials of novel stroke therapies, from cell-based therapies to surgical devices to robotics.

James Z. Wang is a Distinguished Professor of Information Sciences and Technology at The Pennsylvania State University. He received the bachelor's degree in mathematics from the University of Minnesota, and the MS degree in mathematics, the MS degree in computer science, and the PhD degree in medical information sciences, all from Stanford University. His research interests include image analysis, image modeling, image retrieval, and their applications. He was a visiting professor at the Robotics Institute at Carnegie Mellon University, a lead special section guest editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence, and a program manager at the Office of the Director of the National Science Foundation.

Stephen T.C. Wong has over three decades of experience in healthcare and life sciences. He holds John S. Dunn Sr. Presidential Distinguished Chair and is the Founding Chair of Systems Medicine and Bioengineering Department; Director of T.T. & W.F. Chao Center for BRAIN; Associate Director of Cancer Center at Houston Methodist; and Professor of Weill Cornell Medicine. Previously, he was a Professor at UCSF and Harvard, handling major medical information and imaging system design and implementation. Stephen has also served in executive roles in major technology-driven companies including HP, AT&T Bell Labs, Philips Healthcare and Charles Schwab.