Enhancing AI-assisted Stroke Emergency Triage with Adaptive Uncertainty Estimation

Shuhua Yang[1]*, Tongan Cai[1]*, Haomiao Ni[2]*, Wenchao Ma[1],
Yuan Xue[3], Kelvin Wong[4], John Volpi[4], James Z. Wang[1],
Sharon X. Huang[1], and Stephen T.C. Wong[4]

[1] The Pennsylvania State University, University Park, Pennsylvania, USA
[2] The University of Memphis, Memphis, Tennessee, USA
[3] The Ohio State University, Columbus, Ohio, USA
[4] Houston Methodist Hospital, Houston, Texas, USA
stwong@houstonmethodist.org

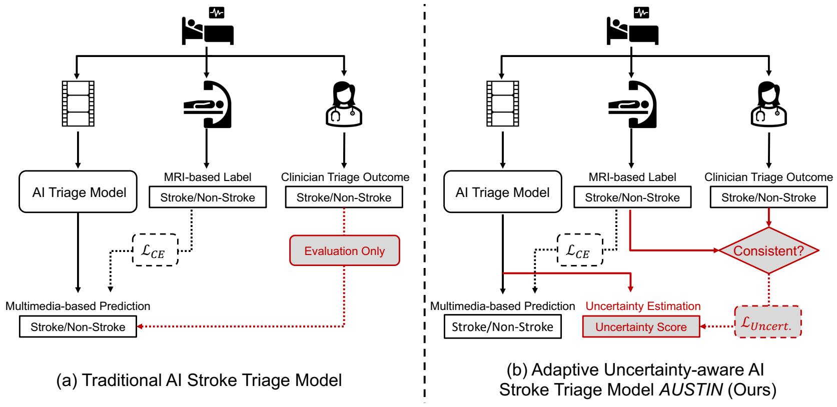
Abstract. Stroke diagnosis in emergency rooms (ERs) is challenging due to limited access to MRI scans and a shortage of neurologists. Although AI-assisted triage has shown promise, existing methods typically use MRI-derived training labels, which may not align with stroke patterns in patient multimedia data. To address this mismatch, we propose an Adaptive Uncertainty-aware Stroke TrIage Network (AUSTIN)[1], that leverages inconsistencies between clinician triage decisions and MRI-derived labels to enhance AI-driven stroke triage. This approach mitigates overfitting to clinician-MRI disagreement cases during training, significantly improving test accuracy. Additionally, it identifies high-uncertainty samples during inference, prompting further imaging or expert review. Evaluated on a clinical stroke patient dataset collected in an ER setting, AUSTIN achieves over 20% performance gain over human triage and a 13% improvement over a prior state-of-the-art method. The learned uncertainty scores also show strong alignment with discrepancies in clinical assessments, highlighting the framework's potential to enhance the reliability of AI-assisted stroke triage.

Keywords: Stroke Triage · Uncertainty Estimation · Multimedia

1 Introduction

Stroke is a leading cause of disability and mortality worldwide [11]. Early diagnosis and intervention significantly improve survival outcomes and post-stroke quality of life. However, delays due to misdiagnosis and underdiagnosis are common, and limited treatment options often arise during stroke presentation, evaluation, diagnosis, and management [1,21]. The gold standard for stroke diagnosis is advanced neuroimaging such as diffusion-weighted MRI, known for its high sensitivity and specificity in detecting brain infarcts. Despite its accuracy,

[*] These authors contributed equally to this work.
[1] Source code for the framework is at https://github.com/shuashua0608/AUSTIN.



(a) Traditional AI Stroke Triage Model

(b) Adaptive Uncertainty-aware AI Stroke Triage Model *AUSTIN* (Ours)

Fig. 1: Illustration of our motivation. Traditional AI-based stroke triage methods [2,24] are trained solely using MRI-based labels, where clinician triage decisions are used only for performance comparison. In contrast, AUSTIN leverages the (in)consistency between MRI-based labels and clinician triage outcomes to *adaptively* estimate the uncertainty of each training instance. This helps reduce overfitting to ambiguous cases—where clinician and MRI labels disagree—during training, while enabling the identification of high-uncertainty samples at inference time for further imaging or expert review.

MRI accessibility in emergency rooms (ERs) is limited due to scarce availability and high operational costs. Therefore, in ER triage, clinicians often rely on standardized assessments such as the National Institutes of Health Stroke Scale (NIHSS) [18], which evaluate unilateral facial droop, arm drift, and speech impairment. However, the shortage of experienced neurologists [14] and the subtle presentation of these symptoms [24] can compromise the accuracy of stroke triage in critical cases.

Recent advances in machine intelligence have shown promise in identifying neurological disorders through multimedia analysis. Cai *et al.* [2] introduced *DeepStroke*, an effective stroke triage framework tailored for ER environments, utilizing face video frames and speech data for stroke detection. Ou *et al.* [19] proposed a multimedia framework that utilizes patient motion video and speech spectrograms, while Yu *et al.* [24] developed a method for evaluating facial imagery and speech transcripts. More recent efforts have expanded to deploying such frameworks on mobile platforms [3]. Nevertheless, existing multimedia-based stroke triage methods typically derive their training labels directly from diffusion-weighted MRI scans and then attempt to predict stroke presence in unseen test data (Fig. 1). This process assumes a direct correlation between MRI findings and stroke symptoms observable in patient videos. This assumption may be flawed since MRI captures structural brain abnormalities that may not consistently manifest as clear motor or behavioral deficits, potentially introducing discrepancies between training data and model predictions.

Relying solely on MRI-based labels to train a multimedia-based stroke triage model may be suboptimal, as it forces the AI model to assess stroke presence with MRI-level accuracy using only multimedia data—despite fundamental differences across clinical protocols. This mismatch, illustrated in Fig. 1, motivates our approach: we integrate clinician triage decisions as complementary supervision during training and enhance the model with uncertainty estimation alongside stroke presence prediction, providing more reliable decision support for clinicians in ER setting. Specifically, we propose AUSTIN, the **A**daptive **U**ncertainty-aware **S**troke **T**r**I**age **N**etwork, a novel model that preserves MRI-based supervision while accounting for higher uncertainty in cases when clinician triage assessments conflict with MRI labels. This design recognizes that some cases are inherently ambiguous based on video data alone—for instance, when an MRI-confirmed stroke lacks clear motor abnormality patterns in videos, or when a non-stroke patient exhibits misleading motor abnormalities.

Allowing for higher uncertainty in ambiguous cases—where clinician triage and MRI labels disagree—offers two key advantages. First, at inference time, it enhances clinical relevance by flagging uncertain cases for further imaging or expert review, aligning with real ER decision-making. Second, during training, it preserves the strengths of MRI-based supervision while introducing flexibility, thereby preventing overconfident predictions in cases with inconclusive evidence. Importantly, when subtle but discriminative visual or auditory patterns are detectable, the model remains capable of confidently aligning its predictions with MRI labels, even when those labels differ from clinician assessments.

We implement AUSTIN within a vision-audio classification framework and evaluate it on an in-house stroke patient dataset collected in an ER setting. Comprehensive experiments demonstrate that our method significantly enhances stroke triage performance, with an over 20% performance gain compared to human triage and a 13% improvement over the prior leading method. The learned uncertainty scores align closely with discrepancies in clinical assessments, enabling a grounded risk estimation proxy in AI-assisted stroke triage.

2 Methods

Dataset. The clinical dataset adopted in this study focuses on multimedia-based AI triage for mild to moderate acute strokes. Participants include patients admitted to the ER with neurological symptoms[2]. Each patient was video-recorded using a mobile phone while describing the "Cookie Theft" picture, as instructed by the NIH Stroke Scale [18], to assess their cognitive and speech abilities. For each video, the discharge binary **ground truth**— referred to as the MRI label— is derived from diffusion-weighted MRI, indicating the confirmed presence or absence of stroke. We also introduce a binary label, referred to as the triage label, which captures the clinician **triage outcomes**, representing the nurse's impression of the presence of stroke-related symptoms.

[2] This study was approved by Institutional Review Boards (IRBs) of Houston Methodist (Protocol No. PRO00020577) and Penn State (Site No. SITE00000562).

The final cohort consists of 249 participants, including 171 positive cases and 78 negative cases for stroke based on MRI-confirmed ground truth. The cohort reflects diversity in races, ages, and genders. To promote generalizability of the proposed methods, we use temporal holdout as a proxy for prospective testing during model development and evaluation. Specifically, 170 participants are allocated for training, 36 for validation, and 43 for testing.

Multimedia Stroke Screening Model. The acquired multimedia data undergoes preprocessing before encoding. Video frames are processed using face detection, tracking, and motion estimation to extract near-frontal views that exhibit meaningful motion, following current best practices [2]. The extracted frame segments are assembled into N video clips, each of fixed length L frames. Corresponding audio is extracted, trimmed of silent intervals, and evenly sliced to align with the N video clips. These sliced audio segments are then converted into log-mel spectrograms.

Capturing subtle stroke-related features from various modalities remains a significant challenge. As demonstrated in Fig. 2, we propose a 3-path encoder for multimedia data encoding, comprising a frame pathway, a local audio pathway, and a global audio pathway. For the facial video, we adopt an image encoder E_f, pre-trained on common face image benchmarks, and encode each processed face frame into feature space. Regarding the speech audio, it is important to note that patients typically present similar speech content, and embedding methods that prioritize downstream transcription tasks are not appropriate. Meanwhile, clinicians tend to focus more on global speech patterns such as pace and slurs. Therefore, we adopt a global audio encoder E_a to encode the whole speech audio. To enhance frame-level facial video features, we introduce a Siamese network E_s which integrates temporally aligned, fine-grained local audio features from the local audio pathway into the frame pathway at multiple stages. E_s is an image encoder that follows the same structure as E_f but operates on audio spectrograms. Frame-level features from the frame and local audio pathways are temporally aggregated with a trainable, parameter-efficient state-space modeling layer as a clip-level feature, which is concatenated with the global audio feature in a late-fusion scheme. We denote the final feature as h.

Adaptive Uncertainty-Aware Loss. To handle uncertainty in stroke diagnosis, especially considering the cases where MRI and triage labels are inconsistent, we propose to utilize both labels to enhance the overall model performance. We denote the ground-truth MRI labels and triage labels as y_{MRI} and y_{triage}, respectively. Specifically, we propose an adaptive uncertainty-aware loss into our stroke diagnosis model, which dynamically adjusts the model's learning focus and enables further evaluation with the learned uncertainty scores. This novel loss function allows the model to efficiently learn from high-confidence cases—where y_{MRI} and y_{triage} agree—while capturing uncertainty in cases of disagreement. The uncertainty-aware loss function also provides potential explanations for the uncertainty associated with each patient case, offering deeper insights into clin-

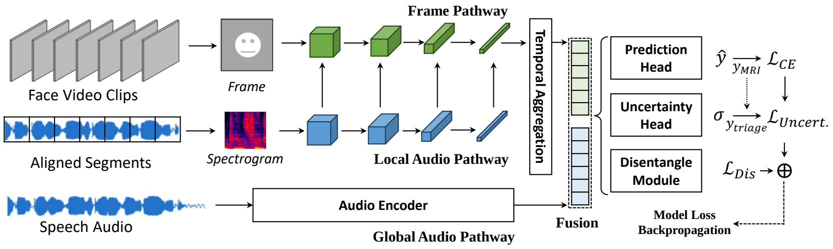


Fig. 2: Overview of the proposed AUSTIN framework. This model comprises three pathways: the frame pathway and a local audio pathway form a Siamese-like network, enabling multi-level feature fusion, with temporal aggregation capturing temporal dynamics. A global audio pathway is implemented via an audio encoder with global attention. Additionally, the model encoder is enhanced with an uncertainty head incorporating our proposed adaptive uncertainty loss function and a disentangling module for adversarial training.

ical diagnosis. Inspired by [13], this loss function is formulated as follows:

$$\mathcal{L}_{\text{Uncert.}} = \frac{1}{2\sigma^2}\mathcal{L}_{\text{CE}} + w\log(\sigma+\epsilon)\ , \quad w = \exp(-\alpha|y_{\text{MRI}} - y_{\text{triage}}|), \tag{1}$$

where $\mathcal{L}_{\text{CE}}$ is cross-entropy loss between y_{MRI} and the predicted output $\hat{y}$ from the prediction head, w is a weight adaptively adjusted based on agreement between y_{MRI} and y_{triage}, and σ is the output logits from the uncertainty head that represents the learned uncertainty score for each patient case. The constant ϵ (set to 1 in our experiments) ensures numerical stability and guarantees that the second term in the loss function remains positive. The first term in Eq. 1, $\mathcal{L}_{\text{CE}}/2\sigma^2$, penalizes prediction errors relative to y_{MRI}; the uncertainty score σ scales the penalty, forcing the model to assign higher uncertainty to ambiguous patient cases where y_{MRI} and y_{triage} are inconsistent. The second term, $w\log(\sigma+\epsilon)$, regularizes the uncertainty score σ, with w capturing the differences between MRI and triage labels. When y_{MRI} and y_{triage} are consistent, i.e., $w{=}1$, the model is encouraged to reduce σ, reflecting higher confidence in making final predictions. Conversely, when these two labels disagree, the second term scales to $e^{-\alpha}$, allowing σ to remain high, indicating potential difficulty in predicting 'uncertain' patient cases.

The learned uncertainty score σ potentially provides meaningful insights into the model's confidence in clinical diagnosis and indicates the consistency of clinical assessments. Ideally, lower σ values indicate high confidence in prediction, generally associated with cases where triage and MRI labels agree or where decisions are easier to make. In contrast, higher σ values reflect greater uncertainty, corresponding to cases with label disagreement or complex patient presentations that may require further clinical evaluation. As demonstrated in Sec. 3.2, analysis of the distribution of σ values enables explainable assessment of the diagnosis model. This provides clinicians with valuable cues for identifying patterns that contribute to misalignment between triage assessment and MRI-confirmed diagnoses.

Overall Training Objectives. Following previous practice [2], we also adopt a discriminative network with adversarial loss to perform identity disentanglement, aiming to generate identity-free audio-visual features. Given a pair of video frames, which may come from the same or different subjects, we encode their features as h_i and h_j. A discriminator, denoted as D, then predicts whether h_i and h_j belong to the same person. The training loss for D takes the form of Mean Squared Error (MSE), as LS-GAN adopts [16]:

$$\mathcal{L}_{\text{Dis}} = \sum_{i,j} \|\delta_{ij} - \text{D}(h_i, h_j)\|_2 \ , \qquad \mathcal{L}_{\text{adv,E}} = -\sum_{i,j} \|0.5 - \text{D}(h_i, h_j)\|_2 \ , \tag{2}$$

where δ is the Kronecker delta function, defined as $\delta_{ij} = 1$ when $i = j$ and 0 otherwise. The final model encoder training loss $\mathcal{L}_{\text{E}}$ is composed of two components. The first component is the aforementioned adaptive uncertainty-aware loss, $\mathcal{L}_{\text{Uncert.}}$. The second component is an adversarial loss imposed on the encoder $\mathcal{L}_{\text{adv,E}}$ to adversarially promote uncertainty in the output of D, thereby enhancing the robustness of the model against overfitting and improving its ability to generalize to unseen data. The total training loss is formulated as $\mathcal{L}_{\text{E}} = \mathcal{L}_{\text{Uncert.}} + \lambda \mathcal{L}_{\text{adv,E}}$ with tunable λ. We train E and D iteratively like GAN [7] that alternates parameter update and freezing between E and D.

3 Experiments

3.1 Setup and Implementation

We perform face preprocessing with the PLFD model [8]. We set $N = 7$ and $L = 64$ during preprocessing. We configured $M = 128$ audio log-mel bins and employed a hidden dimension of 600. For the backbone of the frame pathway, we have assessed the effectiveness of face-pretrained transformer models [5] and VideoMAE [22], but neither surpasses a face-pretrained ResNet [9]. Performance comparisons are detailed in the ablation studies. The multimedia frame-level encoder consisted of a ResNet-50 face frame encoder pretrained on the FairFace dataset [12] and a ResNet-18 local spectrogram encoder pretrained on the ESC-50 dataset [20]. We adopt the Structure State-Space Model (S4) [10] for temporal feature aggregation. We leverage One-Peace [23], a top-performing audio transformer, for extracting global audio features. The discriminator D employed a fully convolutional network (FCN) [15] with three layers. During training, all encoders, including ResNets and One-Peace, had parameters frozen. The temporal module S4, classification head, uncertainty head, and module D received gradient updates. We trained the model using a batch size of 32, with a learning rate of $1e^{-4}$ and a dropout ratio of 0.2. The best-performing model on the validation set was retained. Training for 100 epochs took about six hours on an NVIDIA V100 GPU.

Table 1: Main Performance Comparison. Results are reported on the temporal holdout test data. Due to the imbalance ratio in the dataset, we use AUC as the benchmark (not computed for triage performance due to binary label).

Model	Accuracy	Specificity	Sensitivity	AUC
Clinician Triage Performance	0.5349	0.5385	0.5333	-
DeepStroke (SoTA)	0.6977	0.6154	0.7333	0.6564
Proposed Encoder w/o $\mathcal{L}_{\text{Uncert.}}$	0.6047	0.6923	0.7000	0.6658
+ $\mathcal{L}_{\text{Uncert.}}$ w/ Fix $w = 1$	0.6976	0.6154	0.7333	0.7128
+ Adaptive w (**AUSTIN**)	**0.7442**	**0.7692**	**0.7333**	**0.7897**

3.2 Stroke Diagnosis Model Performance

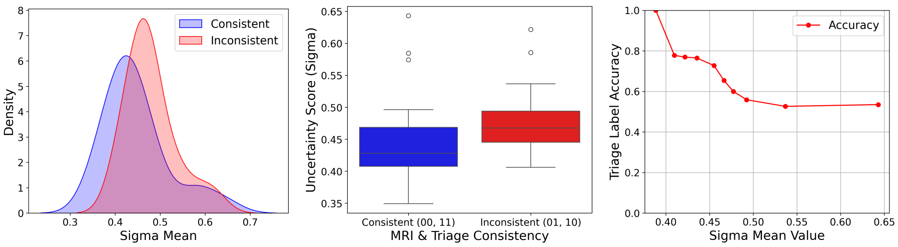
To evaluate the proposed model AUSTIN, we carried out extensive experiments to demonstrate its effectiveness, including comparison with a state-of-the-art (SoTA) stroke diagnosis model *DeepStroke* [2]. All reported results are based on the aforementioned temporal holdout test set with 43 patients. We measured model performance using Accuracy, Specificity, Sensitivity, and AUC. The main performance comparison is presented in Table 1. AUSTIN significantly outperforms clinician triage, with gains of 21%, 23%, and 20% in accuracy, specificity, and sensitivity, respectively. It also surpasses the *DeepStroke* framework with 13% AUC improvements, establishing a new SoTA for AI-assisted stroke triage. Breaking down these gains, the proposed 3-path encoder with a simple CE loss improves AUC by 3% over *DeepStroke*. Incorporating uncertainty estimation into the framework further enhances performance, with the weight-fixed uncertainty loss contributing an additional 3% improvement. With the help of adaptive weighting, the final proposed AUSTIN model achieves over 13% performance margin in terms of AUC, validating the effectiveness of integrating adaptive uncertainty-aware loss into our model.

Ablation Study. Theoretically, the proposed 3-path encoder can take arbitrary backbones and can be easily generalized to other related tasks. We demonstrate the impact of different encoders on model performance. We evaluated several SoTA vision backbones specifically designed for facial feature encoding, including vision transformer (FaceXFormer) [17] and VideoMAE (MARLIN) [4]. Besides the adopted One-Peace audio transformer, we also evaluated the Audio Spectrogram Transformer (AST) [6]. Note that the uncertainty loss was not included when benchmarking these models. As shown in Table 2, our chosen vision and audio backbones outperform other models in multimedia stroke detection.

Uncertainty Evaluation. To gain deeper insight into the proposed uncertainty-aware loss, we analyzed the estimated uncertainty values σ in relation to the consistency between MRI and triage labels, with results shown in Fig 3. Fig. 3a presents the kernel density estimation (KDE) plot of the mean σ values for each patient case, indicating a distributional shift in uncertainty values between consistent and inconsistent cases: inconsistent cases exhibit higher uncertainty scores. Fig. 3b further categorizes σ distribution based on four possible MRI-

Table 2: Ablation Study on Model Configuration.

Vision	**Global Audio**	**Local Audio**	**Accuracy**	**Specificity**	**Sensitivity**	**AUC**
ResNet50	One-Peace	ResNet18	0.6047	**0.6923**	0.7000	**0.6658**
ResNet50	AST	ResNet18	**0.6977**	0.6154	**0.7217**	0.6564
ResNet50	×	ResNet18	0.5584	0.4846	0.6333	0.5821
×	One-Peace	×	0.5814	0.4615	0.6333	0.5897
FaceXFormer	One-Peace	×	0.6279	0.6154	0.6333	0.6051
MARLIN	One-Peace	×	0.6279	0.5846	0.6667	0.6129



(a) KDE plot

(b) Uncertainty Scores

(c) Consistency

Fig. 3: Comparison of σ distributions. (a) KDE plot of mean σ values for consistent vs. inconsistent cases; (b) Boxplot for each MRI-triage label combination: 0=positive; 1=negative; (c) Consistency curve showing MRI-triage agreement (measured as triage label accuracy) at different σ thresholds.

triage label combinations, (0,0), (1,1), (0,1), and (1,0). Notably, cases labeled as (1,0) demonstrate the highest median σ values, suggesting these are among the most diagnostically challenging—patients with MRI-confirmed stroke but no clear outward symptoms—thus requiring further clinical assessment. We also plot MRI-triage consistency (measured as triage label accuracy) against varying thresholds of σ in Fig. 3c, where accuracy declines with higher σ values, indicating a correlation between σ values and uncertainty level for patient cases. These findings confirm that the learned uncertainty parameter serves as an informative proxy for case difficulty, offering valuable decision support in stroke diagnosis.

4 Conclusion

This paper presents AUSTIN, an adaptive uncertainty-aware framework for multimedia-based stroke triage that addresses the fundamental challenge of label inconsistency in clinical AI systems. By incorporating uncertainty estimation sensitive to clinician-MRI label discrepancies, AUSTIN not only achieves SoTA performance but also produces interpretable confidence measures that reflect real-world diagnostic difficulty. The resulting uncertainty scores serve as valuable decision-support signals, helping to identify cases that warrant further expert review—a particularly important feature in resource-constrained emergency settings. Beyond stroke triage, the uncertainty-aware adaptive training paradigm holds promise for broader application in other medical domains where multimodal input may yield conflicting diagnostic labels.

Acknowledgments. Research reported in this publication was supported primarily by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under award number R01NS140292, the T.T. and W.F. Chao Foundation, and the John S. Dunn Foundation. This work used cluster computers at the National Center for Supercomputing Applications through an allocation from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program supported by NSF Grants 2138259, 2138286, 2138307, 2137603, and 2138296. The work also used high-performance computing servers at the Chao Center for BRAIN and Penn State. We also thank Amber Criswell for her assistance with patient enrollment.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cai, T., Ni, H., Ma, W., Xue, Y., Ma, Q., Leicht, R., Wong, K., Volpi, J., Wong, S.T., Wang, J.Z., Huang, S.X.: SafeTriage: Facial video de-identification for privacy-preserving stroke triage. In: Proceedings of the International Conference on Information Processing in Medical Imaging (2025)
2. Cai, T., Ni, H., Yu, M., Huang, X., Wong, K., Volpi, J., Wang, J.Z., Wong, S.T.: DeepStroke: An efficient stroke screening framework for emergency rooms with multimodal adversarial deep learning. Medical Image Analysis **80**, 102522 (2022)
3. Cai, T., Wong, K., Wang, J.Z., Huang, S., Yu, X., Volpi, J.J., Wong, S.T.: M^{3}Stroke: Multi-modal mobile ai for emergency triage of mild to moderate acute strokes. In: IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). pp. 1–8 (2024). https://doi.org/10.1109/BHI62660.2024.10913652
4. Cai, Z., Ghosh, S., Stefanov, K., Dhall, A., Cai, J., Rezatofighi, H., Haffari, R., Hayat, M.: Marlin: Masked autoencoder for facial video representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1493–1504 (2023)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
6. Gong, Y., Chung, Y.A., Glass, J.: AST: Audio Spectrogram Transformer. In: Proc. Interspeech. pp. 571–575 (2021). https://doi.org/10.21437/Interspeech.2021-698
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
8. Guo, X., Li, S., Yu, J., Zhang, J., Ma, J., Ma, L., Liu, W., Ling, H.: Pfld: A practical facial landmark detector. arXiv preprint arXiv:1902.10859 (2019)

9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
10. Islam, M.M., Bertasius, G.: Long movie clip classification with state-space video models. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 87–104. Springer (2022)
11. Johnson, W., Onuma, O., Owolabi, M., Sachdev, S.: Stroke: A global response is needed. Bulletin of the World Health Organization **94**(9), 634 (2016)
12. Karkkainen, K., Joo, J.: 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 (2021)
13. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? Advances in Neural Information Processing Systems **30** (2017)
14. Leira, E.C., Kaskie, B., Froehler, M.T., Adams Jr, H.P.: The growing shortage of vascular neurologists in the era of health reform: Planning is brain! Stroke **44**(3), 822–827 (2013)
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440 (2015)
16. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2794–2802 (2017)
17. Narayan, K., VS, V., Chellappa, R., Patel, V.M.: Facexformer: A unified transformer for facial analysis. arXiv preprint arXiv:2403.12960 (2024)
18. NIH: NIH Stroke Scale. National Institutes of Health. <`https://www.stroke.nih.gov/resources/scale.htm`> (2003), (accessed 25-Feb-2025)
19. Ou, Z., Wang, H., Zhang, B., Liang, H., Hu, B., Ren, L., Liu, Y., Zhang, Y., Dai, C., Wu, H., Li, W., Li, X.: Early identification of stroke through deep learning with multi-modal human speech and movement data. Neural Regeneration Research **20**(1), 234–241 (2025)
20. Piczak, K.J.: ESC: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM International Conference on Multimedia. pp. 1015–1018 (2015)
21. Rafay, M.F., Pontigon, A.M., Chiang, J., Adams, M., Jarvis, D.A., Silver, F., MacGregor, D., deVeber, G.A.: Delay to diagnosis in acute pediatric arterial ischemic stroke. Stroke **40**(1), 58–64 (2009)
22. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in Neural Information Processing Systems **35**, 10078–10093 (2022)
23. Wang, P., Wang, S., Lin, J., Bai, S., Zhou, X., Zhou, J., Wang, X., Zhou, C.: One-peace: Exploring one general representation model toward unlimited modalities. arXiv preprint arXiv:2305.11172 (2023)
24. Yu, M., Cai, T., Huang, X., Wong, K., Volpi, J., Wang, J.Z., Wong, S.T.: Toward rapid stroke diagnosis with multimodal deep learning. In: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, Part III. pp. 616–626. Springer (2020)