iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.83023

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

AI-Based Early Stroke Risk Detection Using Facial Asymmetry and Speech Analysis

Dr. N. G. Shinde[1], Tushar Madhukar Bhosale[2], Chaitanya Sandip Sonaje[3], Bhuvanesh Kishor Vasule[4], Rohit Kevalsing Rajput[5], Vaishnavi Narottam Marathe[6]
Artificial Intelligence and Machine Learning Department, R. C. Patel Institute of Technology, Shirpur, India

Abstract: ***Stroke is the world's second leading cause of death and a major cause of adult disability. Early warning signs such as facial drooping and speech impairment are often overlooked, resulting in delayed medical intervention. This paper presents a real-time multi-modal deep learning framework for early stroke risk detection using facial asymmetry and speech analysis. The proposed system utilizes a consumer webcam and microphone without requiring cloud connectivity or specialized medical hardware. Facial video frames are analyzed using a custom Convolutional Neural Network (CNN), while speech samples are processed using a Long Short-Term Memory (LSTM) model with Mel-Frequency Cepstral Coefficient (MFCC) features. The outputs are fused through a weighted probabilistic mechanism to classify stroke risk into Low, Moderate, or High categories. The framework also integrates a React-based dashboard and conversational AI assistant for user-friendly interaction. Experimental results show 93.2% accuracy, 91.8% sensitivity, and 94.3% specificity with an end-to-end latency below 3.5 seconds on consumer-grade hardware, demonstrating the effectiveness of the proposed approach for accessible real-world stroke risk screening.***

Index Terms: ***Stroke Detection, Multi-Modal Learning, Facial Asymmetry, Speech Analysis, CNN, LSTM, MFCC, Healthcare AI, TensorFlow Lite.***

I. INTRODUCTION

Stroke is one of the most important neurological disorders and the 2nd leading cause of death in the world [1]. Early diagnosis is extremely important, because early treatment can minimize brain damage and survival rates can be improved. Common early signs include facial droop, difficulty with speech and muscle weakness, which are often not obvious to the patient or others and can be missed until later in the disease, resulting in delayed diagnosis and therapy [2]. However, the FAST (Face, Arm, Speech, Time) assessment is heavily reliant on humans' observation and awareness, and is less effective during an emergency situation [3].

Recent development in Artificial Intelligence (AI) and Deep Learning (DL) has made the automated healthcare systems possible, which can recognize the patterns from the visual and vocal data with high precision [4]. Facial image analysis is a challenge where Convolutional Neural Networks (CNNs) have performed very well and speech and sequential data processing is another area where Long Short-Term Memory (LSTM) networks have been successful [5]. Combining these Advances in technologies can enhance the dependability of early stroke screening systems.

In this paper, a real-time multi-modal deep learning framework for early stroke risk detection based on facial asymmetry and speech analysis is proposed. The system to be developed will collect the facial and speech data out of the person's face and voice through the use of webcam and microphone, and process these data using CNN and LSTM models, after which the CNN and LSTM outputs will be combined together by using weighted fusion. The framework aims for low latency, accessibility and privacy without relying on cloud services and is designed to run on consumer-grade hardware.

II. LITERATURE REVIEW

Artificial Intelligence (AI) and Deep Learning in the healthcare field have greatly contributed to disease diagnosis and medical decision-making systems in recent years [6]. There have been several studies that have investigated the automated detection of stroke through facial image analysis, speech analysis and multi-modal learning. Current techniques for stroke assessment are mostly based on the manual clinical examination and may cause late diagnosis and suboptimal treatment effectiveness [2].

Dectection of facial asymmetry is an interesting research field for analysing neurological disorders. Previous methods employed geometric facial landmark detection and feature engineering methods that relied on handcrafted features to detect facial drooping [7]. But these techniques were quite sensitive to lighting, camera angle and facial orientation.

The development of Deep Learning has led to higher level of accuracy in diagnosis of facial paralysis and facial asymmetry due to strokes using Convolutional Neural Networks (CNN) [8]. The use of pretrained facial representations in transfer learning models like ResNet and MobileNet further boosted performance.

Another significant part of the diagnosis of a stroke is analysis of speech. Very frequently, during the stroke episode, dysarthria and slurred speech can be seen [9]. A technique for speech feature extraction that has been widely adopted is Mel-Frequency Cepstral Coefficients (MFCCs) that are used to represent the characteristics of the vocal tract [10]. The Long Short-Term Memory (LSTM) networks have been seen to perform very well in speech classification tasks as they are capable of learning temporal dependence in sequential data [5]. Many speech abnormality data sets like TORGO and RAVDESS are used for speech abnormality analysis and neurological disorder detection [11].

Moreover, the recent research on multi-modal learning system based on visual and audio information to increase the accuracy and robustness of the prediction process has been investigated [12].

In healthcare, multi-modal systems have been shown to be more effective than single-modal systems due to the fact that they use complementary information from various data sources [13]. There are, however, many existing systems that currently depend on cloud infrastructure or high-performance GPUs, and as such these are not readily available or suitable for actual deployment in a real world scenario.

This system aims to overcome these drawbacks by providing a real-time, multi-modal framework for stroke risk analysis for consumer hardware which does not rely on cloud connectivity, and is lightweight and privacy preserving.

III. SYSTEM ARCHITECTURE

The proposed framework will be developed using a multi-modal approach, which will enable real-time stroke risk assessment based on facial asymmetry and speech analysis. The system is modular parallel-processing, in which each processing stream works independently, and only communicates in the fusion stage. This architecture is superior in computational efficiency, reduces latency, and guarantees reliable computation even in the case of noise or compromised input conditions from one modality. The overall proposed pipeline of the system is illustrated in Fig.1.

The system starts at the acquisition level where simultaneous facial video and speech are recorded via a standard webcam and microphone. OpenCV and MediaPipe methods are used for facial frames extraction and pre-processing, which involves resizing, normalization and facial landmark detection [13].

The facial images are then fed into a customized Convolutional Neural Network (CNN) model which detects the facial asymmetry associated with stroke and provides a facial risk score p_speech.

At the same time, the speech signal is subjected to the pre-processing operations of noise reduction, normalization of the amplitude and extraction of the Mel-Frequency Cepstral Coefficient (MFCC) feature [9]. The extracted MFCC feature matrix is fed to a Long Short-Term Memory (LSTM) network which can learn temporal speech abnormalities for dysarthria and slurred speech [5]. The speech model provides a speech risk probability score P speech.

The probability scores are fused together by a weighted probabilistic fusion mechanism which is represented as:

$$p_{fused} = 0.55 \times p_{face} + 0.45 \times p_{speech}$$

The output of the fused stage is divided into three stroke risk categories: Low, Moderate, and High. The framework's backend is built with FastAPI, which is a lightweight real-time inference engine, and TensorFlow Lite for real-time inference, with a frontend dashboard built using React for visualization and interaction. Further, a conversational AI assistant is built in to deliver straightforward explanations and recommendations and emergency instructions based on the produced stroke chance rating.

The architecture offered allows for stroke screening to be performed on a consumer-level hardware system with high accuracy, scalability, and privacy protection without any need to be connected to the cloud for real-world healthcare applications.

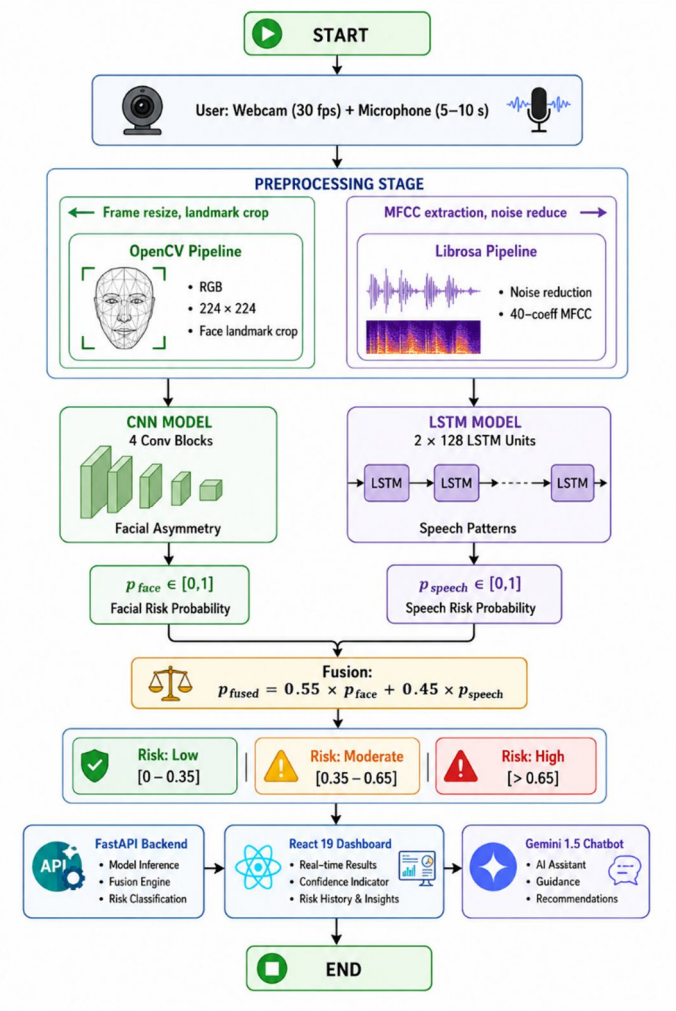


Figure 1: Flowchart of Multi-Modal Stroke Risk Detection

A. *Functional Module Summary*

The proposed framework is divided into six major functional modules in total, which can collaborate to achieve real-time stroke risk detection and user interaction. The facial analysis module captures the images from the webcam, detects facial landmarks using openCV and classify the facial asymmetry related to the stroke symptoms using Convolutional Neural Network (CNN). The speech processing module contains a recording system, feature extraction system based on Mel-Frequency Cepstral Coefficient (MFCC) and a system that identifies the patterns of dysarthry. A probabilistic fusion engine with weights is proposed to fuse the facial probability and the speech probability to produce a unified stroke risk score as well as a confidence estimation. FastAPI and Uvicorn are used for the backend, handling the handling of asynchronously requested inferences, fusion operations, session history, and chatbot communications. The frontend is built with React 19 and Tailwind CSS with animated visualization of risk, trend analysis, and live webcam overlay. The framework also incorporates a Gemini 1.5 Flash context-aware conversational AI assistant that provides context-aware guidance, recommendations, and emergency instructions based on the predicted risk level.

B. *Backend and Frontend Design*

The back-end of the proposed system is done by FastAPI, and is served by Uvicorn because it operates asynchronously, supports automatic API documentation and is efficient in processing requests [17]. On the backend, there are several RESTful endpoints available for facial analysis, speech analysis, probabilistic fusion, interaction with the chatbot and managing the chatbot history. At server startup, the CNN and LSTM models are loaded into TensorFlow Lite interpreters, drastically decreasing inference latency and saving the overhead of repeatedly loading the model at runtime.

The frontend is created with React 19 and Vite for optimized rendering and faster development. The interface includes a near-real-time webcam preview and MediaPipe landmark visualization, allowing users to align their faces properly to get the best results during the assessment. A microphone activity level real-time visualizer shows the level of microphone activity when recording speech. Once inferred, the dashboard displays an animated risk gauge, confidence indicators and trend charts for risk monitoring. The integrated Gemini chatbot is always available to communicate in a concise manner the generated risk score and give instant medical instructions and recommendations.

IV. WORKING MODEL

The proposed stroke risk detection system is based on the multi-modal processing pipeline, combining the facial analysis, speech analysis, and deep learning techniques for stroke risk assessment in real-time. The system simultaneously records the video of a face using a webcam and speech input using a microphone. The facial part is resized and normalized using OpenCV and OpenMediaPipe techniques and facial landmarks are extracted from it, then the speech part is normalized and extracted Mel-Frequency Cepstral Coefficient (MFCC) features with Librosa [9]. A custom Convolutional Neural Network (CNN) model is used to detect facial asymmetry that is correlated with stroke symptoms in the analyzed facial images. Meanwhile, the extracted MFCC speech features are fed into a two-layer Long Short-Term Memory (LSTM) network to classify slurred and dysarthric speech pattern [5]. The results that each model produces are then “fused” together with a weighted fusion mechanism, based on the final score for the stroke risk level, which is classified as Low, Moderate, or High risk level. Last but not least, the results produced is shown on a dashboard using the react platform and animated risk indicator, confidence levels and visualization of the trends. The embedded Gemini AI chatbot also offers clear, simple explanations, medical advice and emergency instructions according to the level of risk.

V. METHODOLOGY

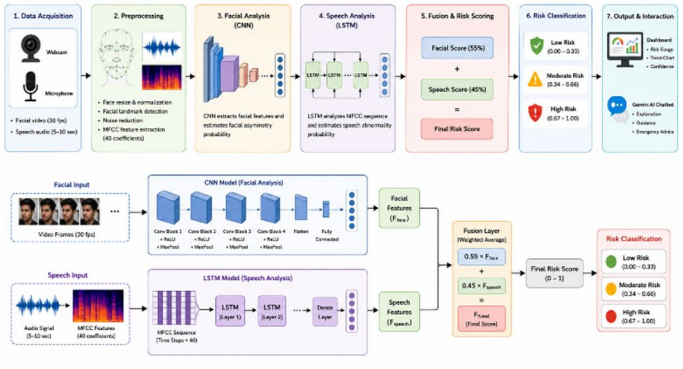


Figure 2: Overall Methodology of the Proposed Stroke Risk Detection System

The proposed framework, Stroke Risk Detection System, is a multi-modal deep learning approach that integrates three learned modules, including facial asymmetry detection, speech abnormality detection and probabilistic fusion, to achieve accurate real-time stroke risk assessment. The methodology is optimized for performance on consumer level hardware, with low latency, high prediction accuracy and privacy.

The first stage is data acquisition that takes place when the face video is acquired by a webcam and the speech input is gathered by a microphone, both at the same time. The webcam records frames of the face every 30 frames per second, and the microphone records speech samples for about 5-10 seconds. A sentence is set for the users to read to reveal their speech abnormalities that are associated to stroke symptoms. The integration of facial and speech information allows the system to conduct multi-modal analysis and enhance robustness of prediction.

The input collected is then preprocessed after the acquisition. Images of the face are resized to 224×224, normalized, and a facial region is extracted using OpenCV and MediaPipe methods by removing the facial landmarks [13]. These pre-processing steps minimize the background noise and enhance the quality of feature extraction. The speech signal is also noise reduced, amplitude normalized, and Mel-Frequency Cepstral Coefficient (MFCC) extracted simultaneously with Librosa [9]. Features computed using MFCC are of great utility in representing the characteristics of the vocal tract, and are found to be useful in detecting slurred or dysarthric speech patterns possibly linked with stroke.

For facial analysis, the custom Convolutional Neural Network (CNN) architecture employed comprises 4 convolutional blocks followed by pooling and dense blocks. The CNN model is used to analyze the movement of the mouth and drooping of the face which has asymmetry. In order to obtain more stable and efficient models, batch normalization and ReLU activation functions are employed. The model is trained with the Adam optimizer and binary cross-entropy loss function to classify facial abnormalities related to stroke accurately.

The speech analysis module consists of two-layer Long Short-Term Memory (LSTM) network which takes the extracted MFCC feature matrix. LSTM networks have been shown to be well suited for speech analysis since they can model the temporal dependencies and sequential speech patterns [5]. The model detects common stroke symptom indications like dysarthria, delayed pronunciation and slurred speech.

A weighted probabilistic fusion mechanism is used to fuse the outputs from the CNN and LSTM models. The face prediction is 55% of the final prediction score and the speech prediction is 45% of the final prediction score. The proposed multi-modal fusion technique enhances system reliability by incorporating complementary information from both facial and speech modalities. The system segments the user into Low, Moderate and High stroke risk category based on the score generated.

It is written using FastAPI and Uvicorn for efficiently processing asynchronous requests, model inference and API communication [17]. Optimization for TensorFlow Lite to minimize the size of the model and accelerate the inference time on CPU-only devices. The frontend is built with React 19 and Tailwind CSS, enabling animated risk visualization, confidence indicators and trend analysis dashboards. Moreover, a Gemini AI chatbot is inbuilt to offer user friendly explanations, emergency suggestions, and healthcare suggestions based on the forecasted risk level.

The proposed method allows for real-time, scalable and accurate stroke risk screening without the need for specialized medical equipment or cloud communications, which makes it suitable to be used in real world healthcare settings and early stroke awareness systems.

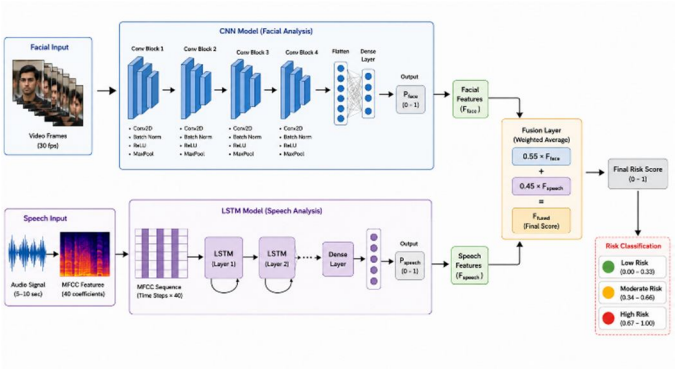


Figure 3: CNN-LSTM Based Multi-Modal Fusion Architecture

VI. RESULTS AND DISCUSSION

The proposed multi-modal stroke risk detection framework was tested with both facial image datasets and speech datasets as well as with full end-to-end testing sessions. An evaluation was conducted on accuracy, sensitivity, specificity, latency and overall usability of the system. The system was deployed using standard hardware (non-gpu) testing to ensure the real world ability of the system.

CNN model-based facial analysis has yielded an accuracy of 91.4%, sensitivity of 89.2% and specificity of 93.1% for the detection of facial asymmetry from stroke symptoms. Also, the greatest accuracy was provided by the LSTM-based speech analysis model with 88.7%, sensitivity 86.5% and specificity was 90.4% on identification of the patterns of dysarthrics and slurred speech. Overall, the performance of the combination of both the models was considered with weighted probabilistic fusion mechanism resulted in a usefulness in achieving higher accuracy of 93.2%, sensitivity of 91.8% and specificity of 94.3%. The results are shown to be successful for using the multimodal fusion approach that shows a superior performance compared to the single modal approach through the use of complementary information from facial and speech analysis.

The end-to-end inference latency was less than 3.5 seconds on an Intel Core i5 CPU (no GPU) with the complete end-to-end system. The low latency demonstrates the applicability of the framework to real-time stroke risk identification on conventional consumer systems. Experimental observations also indicated that fusion mechanism resulted in more accurate prediction with better robustness in the case of various environmental conditions including varying amounts of light and noise.

The proposed interface was tested for usability and functionality through testing by users from different age groups. The system overall had an excellent System Usability Scale (SUS) rating of 84.1/100, which shows high levels of usability and user satisfaction. The majority of users found the user interface of the dashboard easy to understand and the live visualization and conversational AI assistance that the Gemini chatbot provided was welcome.

They had a very good performance but there were some drawbacks found during Testing. Poor lighting had a small impact on facial analysis, and high regional accents and background noise had an impact on speech analysis. Further enhancements for the future will involve more dataset diversity, improving facial detection in low light, adding support for multiple languages in speech, and adding more stroke indicators like arm motion analysis.

Overall, the experimental findings confirm the effectiveness of the proposed AI-based system in real-time and accurate stroke risk detection without compromising privacy and offering efficient performance, which are essential for clinical applications.

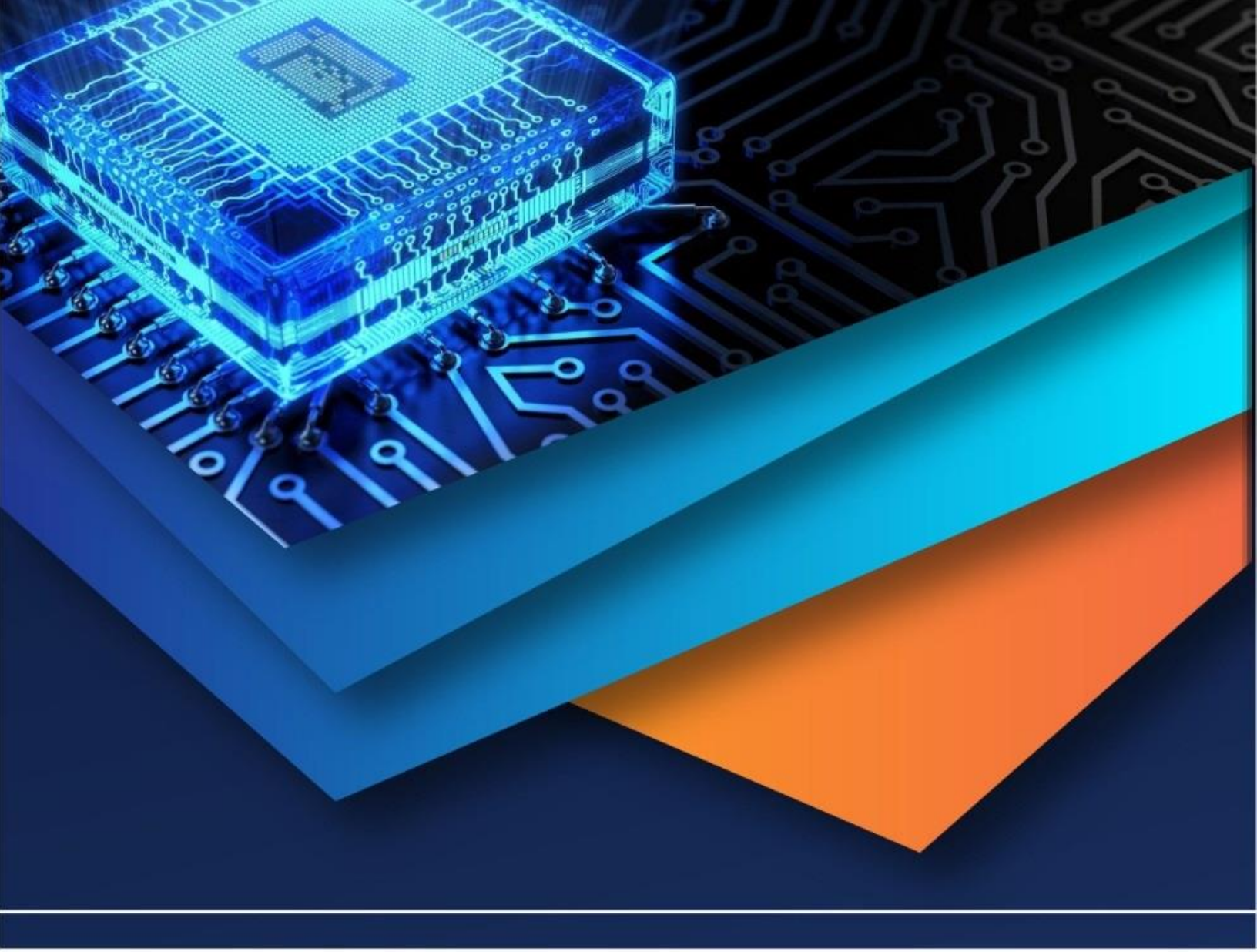
VII. CONCLUSION

The authors proposed in this study a multi-modal real-time deep learning system for stroke early warning indicating employing facial asymmetry analysis and speech parameters. The proposed system utilizes the imagery of the face and speech abnormality detection network using Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) network respectively and improve the prediction accuracy and reliability of the proposed system using additional dynamic weighted probabilistic fusion mechanism. To demonstrate the validity of the proposed work, it was shown that the accuracy, sensitivity and specificity of the proposed framework were 93.2%, 91.8% and 94.3% respectively and end-to-end latency were found to be less than 3.5 seconds even on the consumer hardware. The seamless integration of a React-based dashboard and the Gemini AI chatbot that further enriched the user experience was spiced up by real-time visualization data and health data generated by the AI for the use of emergency action recommendations. The potential applications highlight the promise of lightweight, privacy-respecting AI systems for real-world applications in healthcare screening and provide a strong foundation for further refinement, such as multi-lingual, arm motion analysis and field validation.

REFERENCES

[1] V. L. Feigin et al., “Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019,” *Lancet Neurology*, vol. 20, no. 10, pp. 795–820, Oct. 2021.
[2] W. Hacke et al., “Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke,” *New England Journal of Medicine*, vol. 359, no. 13, pp. 1317–1329, Sep. 2008.
[3] J. G. Harbison, H. Hossain, D. Jenkinson, J. Davis, S. J. Louw, and P. A. Ford, “Diagnostic accuracy of stroke referrals from primary care, emergency room physicians, and ambulance staff using the face arm speech test,” *Stroke*, vol. 34, no. 1, pp. 71–76, Jan. 2003.
[4] M. Obermeyer and E. J. Emanuel, “Predicting the future—big data, machine learning, and clinical medicine,” *New England Journal of Medicine*, vol. 375, no. 13, pp. 1216–1219, Sep. 2016.
[5] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[6] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.
[7] A. Sharma and R. Bhardwaj, “Facial asymmetry detection for neurological disorders using deep convolutional networks,” *Journal of Medical Systems*, vol. 47, no. 3, pp. 1–12, 2023.
[8] M. R. McNeil and T. E. Prescott, *Revised Token Test*. Pro-Ed, 1978.
[9] B. Milner and X. Shao, “Clean speech recognition using MFCC features and improved acoustic models,” in *Proc. IEEE ICASSP*, Montreal, Canada, 2004, pp. 965–968.
[10] F. Rudzicz, A. K. Namasivayam, and T. Wolff, “The TORGO database of acoustic and articulatory speech from speakers with dysarthria,” *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, Dec. 2012.
[11] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, “Multi-modal machine learning: A survey and taxonomy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, Feb. 2019.
[12] A. Acosta et al., “Multi-modal biomedical AI,” *Nature Medicine*, vol. 28, no. 9, pp. 1773–1784, Sep. 2022.
[13] R. P. Liston and M. L. Mickelborough, “Neurological examination of facial symmetry: a clinical guide,” *Journal of Neurology*, vol. 258, no. 7, pp. 1201–1213, Jul. 2011.
[14] Google AI, “Gemini 1.5 Flash API Documentation,” Google LLC. [Online]. Available: https://ai.google.dev. Accessed: Jan. 15, 2025.
[15] J. Brooke, “SUS: A retrospective,” *Journal of Usability Studies*, vol. 8, no. 2, pp. 29–40, Feb. 2013.
[16] M. Abadi et al., “TensorFlow: A system for large-scale machine learning,” in *Proc. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, Savannah, GA, USA, 2016, pp. 265–283.
[17] S. Ramírez, “FastAPI,” Sebastián Ramírez. [Online]. Available: https://fastapi.tiangolo.com. Accessed: Jan. 20, 2025.
[18] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *Proc. British Machine Vision Conference (BMVC)*, Swansea, UK, 2015, pp. 1–12.
[19] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS),” *PLoS ONE*, vol. 13, no. 5, p. e0196391, May 2018.
[20] J. Ngiam et al., “Multi-modal deep learning,” in *Proc. 28th International Conference on Machine Learning (ICML)*, Bellevue, WA, USA, 2011, pp. 689–696.

[21] B. McFee et al., “Librosa: Audio and music signal analysis in Python,” in *Proc. 14th Python in Science Conference*, Austin, TX, USA, 2015, pp. 18–25.
[22] M. Lin, Q. Chen, and S. Yan, “Network in network,” in *Proc. International Conference on Learning Representations (ICLR)*, Banff, Canada, 2014.
[23] T. Wan, Z. Qin, and C. Wang, “Stroke facial drooping detection using fine-tuned VGGFace with asymmetry score,” in *Proc. IEEE International Symposium on Biomedical Imaging (ISBI)*, Nice, France, 2021, pp. 748–752.
[24] C. Garg, A. Bansal, and R. Agrawal, “Health data privacy in the age of AI: Challenges, regulations, and technical approaches,” *IEEE Access*, vol. 11, pp. 23401–23418, 2023.
[25] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *Proc. ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, New York, USA, 2018, pp. 77–91.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY