ELDERLY WELLNESS COMPANION WITH VOICE AND VIDEO-BASED HEALTH ANOMALY DETECTION

Dhananjay, Kumar[1]; Mehal Sakthi, Muthusamy Sivaraja[1]; Sowbarnigaa, Kogilavani Shanmugavadivel[1]; Ved P., Kafle[2]

[1]Department of Information Technology, Anna University, MIT Campus, Chennai, India

[2]National Institute of Information and Communications Technology, Tokyo, Japan

ABSTRACT

The elderly healthcare requires an innovative approach to address multifaceted challenges in tracking, monitoring, and reporting in real-time. The proposed solution harnesses the capabilities of voice and video-based anomaly detection systems to offer continuous monitoring, personalized support, and timely intervention for the physical and emotional well-being of elderly individuals. Central to the proposed system is the integration of real-time voice emotion recognition and video-based posture recognition modules, constructed using cutting-edge deep learning and transfer learning models respectively. These modules are deployed on the Raspberry Pi platform, ensuring accessibility and efficiency. Moreover, attention mechanisms are incorporated to boost accuracy and effectiveness in detecting health anomalies, with a particular focus on identifying falls. The proposed elderly companion system implemented on Raspberry Pi achieves a validation accuracy of 96.34% in voice module and 87.91% in video module in delivering comprehensive healthcare for the elderly. The proposed solution demonstrates a potential work for standardization through the ITU/WHO Focus Group on AI for Health (FG-AI4H).

Keywords – Elderly healthcare, voice emotion recognition, video-based fall detection, transfer learning

1. INTRODUCTION

Reports from WHO and United Nations Department of Economic and Social Affairs reveal that the elderly population of 65 years or older, which was 727 million in 2020, would get doubled by 2050 [1-2]. This demographic shift underscores the critical need for innovative health care tracking systems tailored to the unique requirements of elderly people. An early warning system can analyze changes in voice quality, such as breathiness or hoarseness, as an indication of respiratory or cardiovascular issues. Analyzing acoustic features of speech, such as intensity and pitch variability, can provide insights into the respiratory function and cardiovascular health. Furthermore, the healthcare system can sense shifts in speech patterns indicative of a fall or sudden deterioration in health and trigger alerts to caregivers or healthcare providers. Speech analysis can contribute to personalized healthcare by providing objective and quantifiable measures of health status. The system needs to accurately recognize emotions from voice samples across a diverse population, accounting for variations in tone, pitch, modulation, language, slang, and other factors. By monitoring individualized speech profiles over time, healthcare interventions can be tailored to meet the specific needs of each individual. A multi-modal approach such as combining speech and visual data can provide a more comprehensive understanding of an individual's health status.

Traditional approaches based on sensors [3-4], although monitors more health parameters, lack applicability in elderly health care due to their inherent limitations. The existing state-of-the-art healthcare systems like CarePredict [5] are based on wearable sensors, presenting challenges for seniors who are not technologically savvy or comfortable with wearing devices continuously. Additionally, the accuracy of data collected through wearables can be compromised due to device malfunction or improper usage, leading to potential false alarms, or missed health concerns. This necessitates the development of a touchless health monitoring system which works seamlessly integrated with existing healthcare networks. Video based solutions offer better contactless health care requirements, however they infringe the privacy of end users. The concern for privacy can be minimized by capturing audio alone through multiple microphone systems and video can be captured only in highly abnormal cases.

The proposed elderly wellness companion offers a non-intrusive, yet comprehensive monitoring solution while preserving privacy. By integrating a voice module alongside the video component, our system ensures privacy by activating the video module only upon detecting an anomaly, thus minimizing unnecessary surveillance. Moreover, our approach incorporates multi-modal confirmation of anomalies, enhancing the system's reliability and reducing false alarms. This strategic combination of technologies not only addresses privacy concerns but also enhances the effectiveness of anomaly detection.

In line with the ITU-T FG-AI4H [6] and Recommendation Y.4220 [7], the proposed work focuses on developing AI-driven solutions to enhance fall risk assessment and prevention among elderly. By leveraging machine learning techniques and data fusion, models are developed that integrate multi-modal data sources to enable standardized multifactorial assessment of health risk factors. Through a comprehensive evaluation of the implemented modules and testing scenarios, this research demonstrates the effectiveness and feasibility of our approach in providing comprehensive care for the elderly population. Our proposed work not only considers multi-dimensional input but also works on low power edge devices (e.g., Raspberry pi-based system) while producing better precision score. By advancing AI-driven healthcare systems, our work contributes to improving the quality of life and promoting independent living among elderly individuals, addressing critical needs in eldercare, and fostering societal well-being.

The rest of the paper is organized as follows. Section 2 outlines the related works and standards. Section 3 presents the architectural details and algorithm of the proposed system, and Section 4 describes the implementation and experimentation. Results and discussions are presented in Section 5, and Section 6 provides concluding remarks with major findings of the proposed work.

2. RELATED WORKS

The existing works [8-9] on elderly personalized healthcare cover abnormal behaviors such as falls, tumbles, aggression, and wandering, based on wearable sensors or the video feed captured on consumer networked cameras and processed on a local GPU server. In machine learning based techniques, the attention mechanisms and advanced models enhance the accuracy of anomaly detection, particularly in identifying critical events like falls, thereby mitigating the risk of false alarms, or overlooked health concerns. However, the standardized approaches in system development require interoperability with home networks, IP networks, and health service provider network while meeting the critical requirements of healthcare.

The ITU Focus Group on Artificial Intelligence for Health (FG-AI4H) Topic Group on Falls Amongst the Elderly (TG-Falls) [6] addresses the critical issue of preventing falls among elderly, a common health problem with significant repercussions. Falls among community-dwelling adults aged 65 years or older account for a substantial portion of hospitalizations and lead to loss of independence. AI techniques offer a promising solution by generating models that combine various data sources, enabling standardized multifactorial assessment of fall risk factors and facilitating their implementation in clinical practice. Recommendation ITU-T Y.4220 [7] delineates the criteria and functionalities essential for an Abnormal Event Detection System (AEDS) implemented within smart homes, with a primary focus on identifying health-related occurrences such as falls or strokes. Central to its directives is the establishment of standards aimed at enhancing system efficacy, minimizing false alarms, and addressing privacy apprehensions. The AEDS platform assumes a pivotal role in integrating Internet of Things (IoT) devices with emergency contacts and medical practitioners.

3. PROPOSED SYSTEM

The proposed system comprises two interconnected modules, the Real-Time Voice Emotion Recognition System and the Video-Based Anomaly Detection System. The voice emotion recognition module utilizes a deep learning technique namely Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) and a diverse range of datasets (in training/testing system) to accurately detect and interpret health emotional states in real-time. Concurrently, the video-based anomaly detection module employs a lightweight deep neural network called MobileNet for embedded systems to identify abnormal behaviors and trigger timely interventions. As illustrated in Figure 1, these modules together provide a robust architecture for continuous monitoring, personalized support, and prompt response, in a home environment for elderly individuals living independently.

At its core is the smart health device, serving as the central hub for data processing and decision-making. The voice signal captured through microphones from the user's premise initiates the process, providing real-time emotional cues for analysis. This input undergoes preprocessing and chunk formation to facilitate efficient processing and analysis. Subsequently, the Acoustic Feature Extraction Module extracts relevant features from the voice input, capturing nuances essential for health state recognition. Following feature extraction, dimensionality reduction techniques are applied to streamline the data, enhancing computational efficiency without sacrificing accuracy. The processed voice data then enters the voice-based state recognition model CNN-LSTM, which is a deep learning-based algorithm trained to interpret health states. Predicted health states are generated, indicating the severity of detected emotions. If the severity is low, the system continues to monitor voice input, maintaining vigilance for any significant changes. However, if a high-severity emotion is detected, the system triggers the video capturing module, transitioning to visual analysis for obtaining additional context.

Upon activation, the video input undergoes preprocessing to enhance image quality and reduce noise, preparing it for analysis. The Person Localization Model identifies and tracks individuals within the video feed, facilitating targeted analysis. Concurrently, the Voice-based Emotion Recognition Model operates in tandem, correlating voice cues with visual observations to refine emotion predictions. The overall algorithm of the proposed system is described in Algorithm 1. The YOLO-based segmentation and localization on video feed is used to continuously monitor health-related activities over extended periods, providing valuable insights into patient behavior, movement patterns, and activity levels. This information aids in assessing patient well-being and identifying any deviations from normal behavior. The MobileNet-based enhanced transfer learning model is used on video data to classify postures (e.g.,

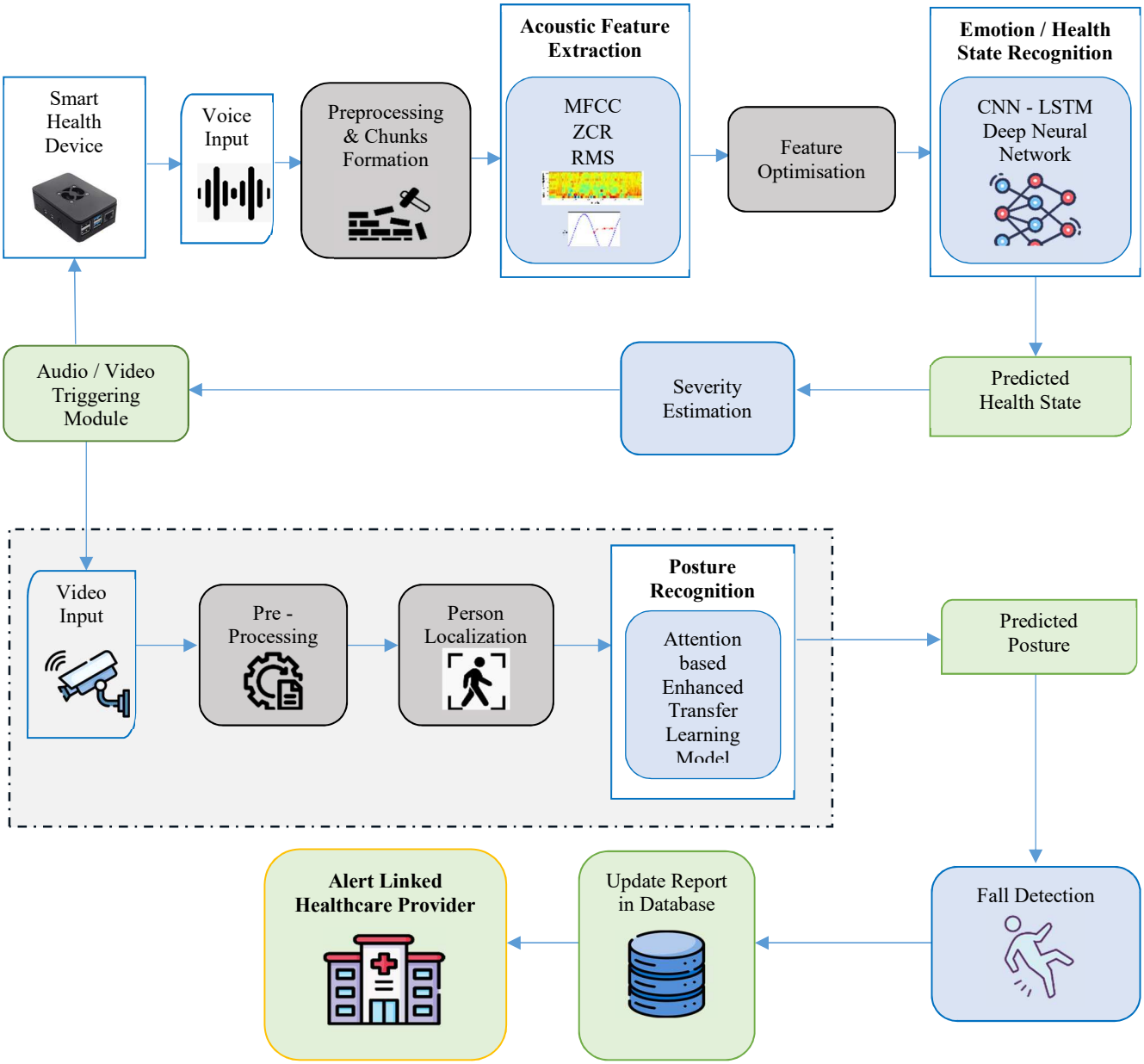

Figure 1 - Proposed system architecture of voice and video-based health anomaly detection

Standing, Sitting, or Fallen) and further analysis. Predicted postures are evaluated for signs of potential health issues, enabling timely anomaly detection and intervention. The developed algorithm is tailored to run on an edge device connected to an IP network system. A Raspberry Pi-based single-board computer is used to implement and test the proposed system architecture with associated machine learning algorithms.

Algorithm 1 – Elderly Health Anomaly Detection algorithm

Input: Real-time voice input, Video feed
Output: Detection and Reporting of Elderly Health Anomalies

1. Initialize microphone and camera sensors for real-time voice and video input.
2. Load pre-trained models for voice emotion recognition and posture classification.
3. Preprocess voice input, segmenting it into short frames.
4. Extract acoustic features from audio.
5. Utilize CNN – LSTM Deep Neural Network to detect emotional states.
6. Check for severity in detected emotional states.
7. Activate video module upon detection of severe anomalous emotional states.
8. Segment video frames using YOLO segmentation to isolate person from background.
9. Employ MobileNet-based enhanced transfer learning model to classify postures (Standing, Sitting, or Fallen).
10. Ascertain for high-severity emotions (e.g., a fall), and trigger alerts.

11. Update on individual's health condition in networked database.

4. IMPLEMENTATION AND EXPERIMENTATION

This section describes an implementation of the proposed system architecture of health anomaly detection.

4.1 Real time voice emotion recognition system

The proposed system utilizes a Convolutional Neural Network (CNN) followed by max-pooling layers to extract relevant features from the input voice data. Batch normalization and dropout layers are incorporated to enhance model stability and prevent overfitting. The CNN architecture is designed to capture complex patterns and variations in the audio signals, facilitating accurate emotion recognition. Subsequent layers of max-pooling further down sample the features, focusing on the most salient information while reducing computational complexity. Following the convolutional layers, Long Short-Term Memory (LSTM) units are employed to capture temporal dependencies and sequential patterns in the voice data. The LSTM layers, coupled with dropout regularization, enable the model to effectively learn and represent the dynamic nature of emotional expressions over time. The final dense layers perform classification, mapping the learned features to different emotional states. By leveraging a combination of CNN and LSTM layers, the proposed system achieves robust and accurate emotion recognition from voice inputs.

The dataset used in this module includes a diverse range of sources, each contributing distinct attributes and characteristics to the training process. The Surrey Audio-Visual Expressed Emotion (SAVEE) [10] dataset features recordings from four male actors, expressing a total of seven different emotions across 480 British English utterances. These sentences meticulously chosen from the standard TIMIT corpus, ensure phonetic balance for each emotion. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [11] presents a multimodal collection of emotional speech and song recordings. With contributions from 24 professional actors vocalizing lexically-matched statements in a neutral North American accent, RAVDESS encompasses a vast repository of 7,356 files covering seven distinct emotions. The Toronto Emotional Speech Set (TESS) [12] contributes stimuli for emotional speech research, featuring 200 target words spoken by two actresses across various emotional states. Further, the Indian Emotional Speech Corpora (IESC) [13] was used in training / testing of the system for emotion classification in speech. With 600 speech samples recorded from eight speakers, each uttering two sentences in five emotions, IESC provides a rich source of English-language data. Collectively, these datasets offer a comprehensive foundation for training and validating the real-time voice emotion recognition system, enabling robust and accurate emotion detection in elderly individuals.

In addition to the established datasets, the proposed work also incorporates Indian regional audio clips sourced from various social media platforms. These clips, representing a diverse range of linguistic and cultural backgrounds, offer valuable insights into emotional expressions. To ensure compatibility with the existing datasets, extensive normalization techniques have been employed. This normalization process involves standardizing the format, quality, and linguistic characteristics of the collected audio clips to align seamlessly with the established datasets, thereby facilitating integration and enhancing the diversity of emotional expressions represented in the training data. By incorporating these additional regional audio clips, our dataset becomes more comprehensive and reflective of the diverse emotional expressions prevalent among elderly individuals. This augmentation not only enriches the training process but also enhances the generalizability and effectiveness of the real-time voice emotion recognition system in accurately detecting and interpreting emotional states across different cultural and linguistic contexts.

The finalized input features for the current model encompass a comprehensive set of acoustic characteristics that are instrumental in capturing the nuances of emotional speech. These features include Zero Crossing Rate (ZCR), Root Mean Squared Error (RMSE), and Mel-frequency cepstral coefficients (MFCCs). Zero Crossing Rate (ZCR) measures the rate at which the signal changes its sign, providing insights into the frequency content and periodicity of the audio signal. With 108 instances of ZCR calculated across the dataset, this feature offers valuable information regarding the temporal characteristics of the speech signals.

Root Mean Squared Error (RMSE) serves as a measure of the amplitude variation in the audio signal, quantifying the energy distribution across the signal's time domain. Similar to ZCR, RMSE is computed 108 times across the dataset, capturing variations in signal intensity and dynamics. Mel-frequency cepstral coefficients (MFCCs) represent a powerful feature set widely used in speech processing tasks. Comprising 2160 coefficients computed across the dataset, MFCCs capture the spectral characteristics of the speech signal, providing insights into the frequency distribution and phonetic content.

The chosen parameters for feature extraction include a hop length of 512 and a frame length of 2048, ensuring efficient processing and capturing relevant temporal and spectral information. In total, the dataset comprises 2376 input features, combining ZCR, RMSE, and MFCCs, which collectively provide a rich representation of the acoustic properties of the emotional speech signals. These features serve as the foundation for training the real-time voice emotion recognition system, enabling accurate detection and interpretation of emotional states in elderly individuals.

The proposed system underwent training for a total of 38 epochs, with early stopping mechanisms employed to prevent overfitting and optimize model performance. Early stopping allows the training process to halt when the model's

performance on a validation dataset ceases to improve, thereby preventing unnecessary computation and ensuring efficient convergence to the optimal solution. This approach enables the system to strike a balance between maximizing training accuracy and generalizing well to unseen data. By monitoring performance metrics across epochs and halting training when further improvements are deemed unlikely, early stopping enhances the system's robustness and efficiency in emotion recognition tasks.

During development, the system captures audio input from a connected microphone and segments it into short frames for analysis. By leveraging lightweight model architecture and feature extraction techniques, the system efficiently processes these audio frames, extracting pertinent features and making emotion predictions with minimal latency. The integration with the Raspberry Pi's hardware ensures smooth operation, effectively meeting stringent latency requirements. The hardware experimental setup for the code development is shown in Figure 2.

Figure 2 – Experimental setup for system development

4.2 Video-based fall detection

The video module of the Elderly Wellness Companion system plays a critical role in ensuring timely intervention and support for elderly individuals, particularly in scenarios where anomalous emotions are detected through the voice module. Upon detection of such anomalies, the video module is triggered, activating the networked camera to capture video footage for further analysis. It's important to note that the camera is only enabled when necessary, ensuring privacy and minimizing intrusion into the user's space.

Once the video footage is captured, it is processed to extract relevant information about the person's posture and activity. The first step involves segmenting the video frames to identify the person's location using YOLO [14] based segmentation. This segmentation technique efficiently isolates the individual from the background, facilitating accurate analysis of their movements and posture.

Upon segmenting the individual in the video frames, the system integrates a MobileNet-based enhanced transfer learning model to analyze the person's posture. Renowned for its lightweight and efficient design, MobileNet ensures swift processing of video frames in real-time, maintaining high performance standards. Trained to categorize postures into three distinct states namely, standing, sitting, or fallen down as shown in Figure 3. The model assumes a critical role in monitoring the individual's condition and swiftly detecting potential emergencies, particularly instances of falls.

The accurate classification of posture serves as a fundamental indicator of the individual's well-being and safety. By adeptly discerning between various postures in real-time, the system enables proactive intervention in critical situations such as falls. This proactive approach is particularly valuable for elderly individuals who may be at increased risk of accidents. Moreover, the MobileNet-based enhanced transfer learning model's adaptability and generalization capabilities helps to deal with diverse environments and individuals. By leveraging pre-existing knowledge from extensive datasets, the model effectively recognizes subtle variations in posture and movement patterns, even under challenging circumstances. The incorporation of MobileNet-based posture detection enhances the system's responsiveness and effectiveness in multi-modal environment.

In pursuit of refining the model's efficacy in identifying falls, advanced concepts of attention and hourglass layers are seamlessly integrated. Attention mechanisms assume a pivotal role by dynamically assigning weights to different features within the input sequence. This dynamic allocation enables the model to prioritize pertinent cues indicative of fall events, thereby bolstering the accuracy of fall detection. By amplifying the importance of features associated with falling, such as sudden alterations in acceleration patterns or specific movement traits, the attention model effectively distinguishes falls from other activities or gestures, mitigating false alarms and facilitating prompt interventions. Furthermore, the hourglass model, originally devised for human pose estimation tasks, contributes significantly to accurately pinpointing the person's body joints or key points within the video frames. Leveraging an encoder-decoder architecture, the hourglass model adeptly captures multi-scale features while preserving spatial information, enabling precise estimation of the person's posture. Through systematic down sampling and subsequent up sampling of the input data, spatial context is meticulously maintained throughout the encoding process. This fidelity to spatial details is further reinforced by skip connections between encoder stages, ensuring the propagation of fine-grained information crucial for posture localization.

The symbiotic integration of attention and hourglass models empowers our system with exceptional precision in identifying falls and monitoring the individual's posture. By meticulously scrutinizing the temporal and spatial attributes of video frames, the model swiftly detects anomalous events and triggers appropriate responses, such as alerting caregivers or initiating emergency protocols. This proactive stance not only enhances the safety and well-being of elderly individuals but also instills confidence and assurance among users and their families. When visual data (Figure 4) is

passed to the proposed module, the Person localization is achieved with a confidence score of 0.75 and further classified as fall detected.

Figure 3 Realtime testing input image

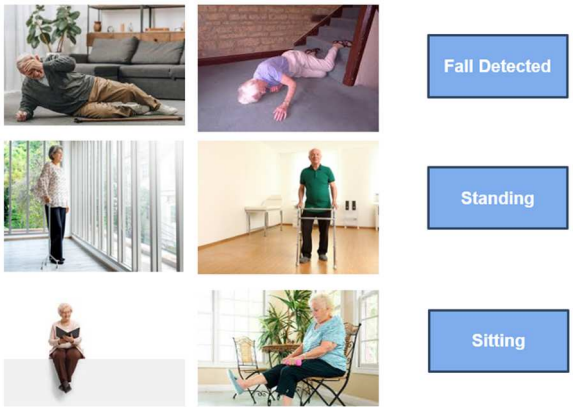

Figure 4 - Dataset sample images

The video module combines state-of-the-art techniques in computer vision and deep learning to monitor the activity and posture of elderly individuals in real time. By integrating MobileNet-based posture detection with attention and hourglass models, the system ensures accurate and efficient identification of anomalies, such as falls. This proactive approach to elderly care enhances the safety and well-being of individuals living independently, providing peace of mind to both users and their caregivers.

5. RESULTS

5.1 Performance of voice-based model

In the evaluation of the proposed system's voice model performance, several key metrics were meticulously analyzed to ascertain its efficacy in real-world applications. Among these metrics, the validation loss assumed significance as it quantified the error between the predicted labels and the actual labels for the validation dataset, offering insights into the model's generalization capabilities and predictive accuracy specific to voice analysis. The system performance measure on speech classification is listed in Table-1 and corresponding confusion matrix is depicted in Figure 5.

Table 1 – Speech classification

Emotion	Precision	Recall	F1-Score
Angry	0.98	0.96	0.97
Disgust	0.96	0.97	0.97
Fear	0.96	0.97	0.96
Happy	0.96	0.95	0.96
Neutral	0.97	0.98	0.97
Sad	0.96	0.97	0.97
Surprise	0.97	0.98	0.98

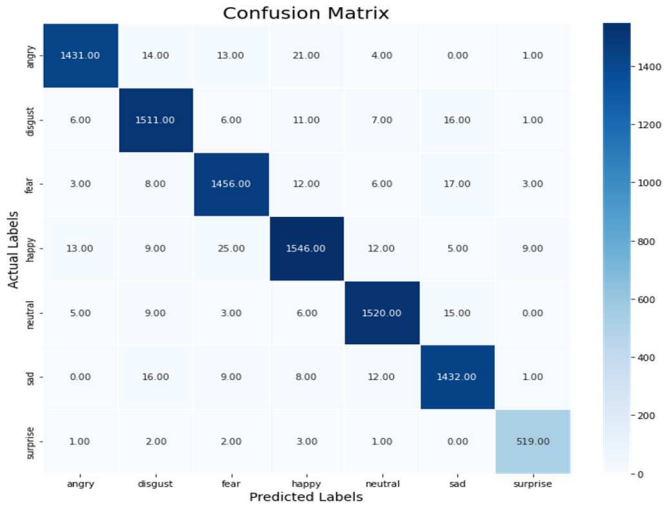

Figure 5 – Confusion matrix of voice model

Validation accuracy has emerged as a pivotal metric in assessing the voice model's proficiency. This metric measured the proportion of correctly classified instances within the validation dataset, serving as a reliable indicator of the overall robustness and reliability of the voice model. With an impressive validation accuracy of 96.34%, the proposed system's voice model demonstrated a high level of accuracy and proficiency in discerning and classifying different speech patterns and emotions. The relative plot between epochs and training and validation accuracies is shown in Figure 6.

Precision, a fundamental metric in classification tasks, provided valuable insights into the accuracy of the voice model's predictions. Both weighted average precision and macro average precision were evaluated to assess the model's performance across all classes in a multi-class classification scenario specific to voice analysis. The weighted average precision, which accounted for the number of samples in each class, offered a comprehensive assessment of the voice model's overall precision. Similarly, the macro average precision provided an unbiased measure of the model's performance across all classes, regardless of class imbalances. With both metrics yielding a consistent precision score of 96%, the proposed system's voice model demonstrated a commendable level of accuracy and reliability across diverse speech patterns and emotions.

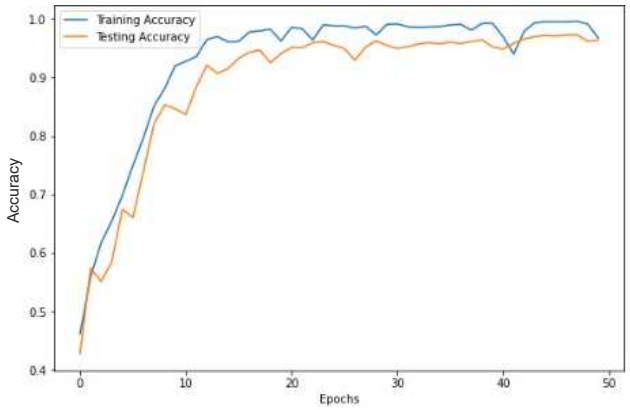

Figure 6 - Epoch vs. accuracy graph of voice model

In addition to these performance metrics, the latency of the voice model was evaluated to ensure its practicality in real-time applications. The average latency for processing voice inputs and generating predictions was measured at 1220 milliseconds across all classes, using Python's time package for precise computation. This latency demonstrates the system's ability to perform timely analysis and response, which is crucial for real-time monitoring and intervention.

5.2 Performance of video-based model

In the evaluation of the video-based model, several key metrics were analyzed to gain comprehensive insights into its performance and efficacy. Among these metrics, the validation loss held paramount importance as it quantified the degree of error between the model's predicted labels and the actual labels for the validation dataset across each epoch of training. A lower validation loss indicated a higher level of accuracy and precision in the model's predictions, signifying an improved performance trajectory. Validation accuracy is used as a key metric in assessing the model's effectiveness during the training phase. It measured the proportion of correctly classified instances within the validation dataset, serving as a reliable indicator of the overall robustness and reliability of the machine learning model. The model's achievement of a validation accuracy of 87.91% as shown in Figure 7, underscored its capability to make accurate predictions across diverse data samples.

Precision score computed on video model, provided valuable insights into the accuracy of positive predictions generated by the model. Weighted average precision and macro average precision were analyzed to gauge the model's performance across all classes in a typical multi-class classification scenario. With both metrics yielding a consistent precision score of 0.88, the model demonstrated a creditable level of accuracy and consistency across diverse classes, affirming its suitability for real-world applications in healthcare and beyond. The comparative analysis of Mean Average Precision with existing systems is tabulated in Table 2.

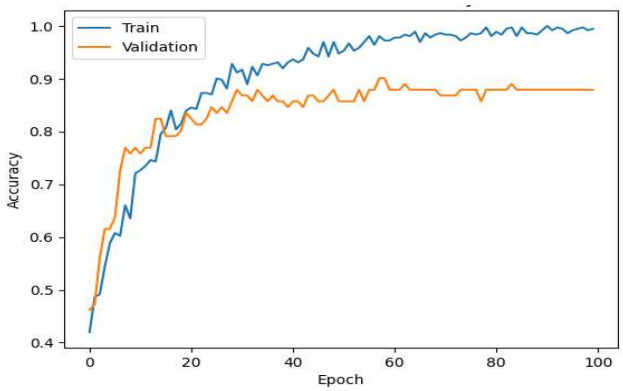

Figure 7 - Epoch vs. accuracy graph of video model

Table 2 – Comparison of mean average precision

Sl. No	Method	Mean Average precision
1	Tan [15]	65.38%
2	Jain [7]	70.17%
3	Hao [16]	82.57%
4	Li [17]	82.37%
5	Yinlong [6]	85.89%
6	Ours	88.00%

The implementation of the video module in proposed system is targeted to detect and accurately interpret subtle visual cues (such as facial expressions, body language, and posture) at an early stage, before any further damage occurs or the health condition deteriorates. The impressive combination of high accuracy and low latency underscores the proposed system's practicality and effectiveness in real-world environment, particularly in enhancing elderly care through prompt and reliable voice / video based anomaly detection and support. The Raspberry Pi-based system implementation facilitates versatile cost-effective healthcare solutions, which are easy to install and maintain in a home environment.

6. CONCLUSION

This paper presented an innovative approach to addressing the multifaceted challenges of elderly care through the design and implementation of an elderly wellness companion system. The voice module demonstrated exceptional performance in analyzing speech patterns and detecting emotional cues, achieving an impressive validation accuracy of 96.34%. Through meticulous evaluation of key metrics such as validation loss and precision, the voice model exhibited robustness and reliability in discerning and classifying different speech patterns and emotions. The voice module acts as a preliminary check, which triggers the video module only when severity in state is detected, ensuring camera privacy of the elderly. Furthermore, the video model complements the voice-based anomaly detection system by providing additional insights into the physical state of elderly individuals. Through the seamless integration of voice and video-based anomaly detection technologies, the proposed system offers comprehensive and holistic care for elderly individuals, promoting independence, safety, and well-being.

The proposed system while conforming to the Recommendation ITU-T Y.4220, can further contribute to the related technology standardization through the ITU-T Focus Group on Artificial Intelligence for Health.

REFERENCES

[1] United Nations Department of Economic and Social Affairs, Population Division, World Population Ageing 2020 Highlights: Living arrangements of older persons (ST/ESA/SER.A/451), 2020.

[2] World Health Organization, "Ageing and Health," https://www.who.int/news-room/fact-sheets/detail/ageing-and-health (accessed Apr. 26, 2024).

[3] Y. Jiang, "Construction of an Intelligent System for Elderly's Health and Elderly Care from the Perspective of the Integration of Smart Sensors and Physical Medicine," 4th International Conference on Smart Systems and Inventive Technology, Tirunelveli, India, 2022, pp. 541-544.

[4] B. Vasudevan, M. Vignesh, and M. Uppuliappan, "Artificial Intelligence (AI) Based Interactive Smart Robot for Elderly People Health Monitoring System", pp1-7, 2023.

[5] CarePredict "All-in-One Solution for Senior Living" https://www.carepredict.com/senior-living/ (accessed May 18, 2024).

[6] ITU-T Focus Group on Artificial Intelligence for Health, "ITU-T FG-AI4H Deliverable - TG Falls: Topic Group on Falls Among the Elderly". https://www.itu.int/en/ITU-T/focusgroups/ai4h/Pages/falls.aspx

[7] Recommendation ITU-T Y.4220 (03/2023), "Requirements and capability framework of abnormal event detection system for smart home".

[8] Y. Zhang, W. Liang, X. Yuan, S. Zhang, G. Yang and Z. Zeng, "Deep Learning-Based Abnormal Behavior Detection for Elderly Healthcare Using Consumer Network Cameras," in IEEE Transactions on Consumer Electronics, vol. 70, no. 1, pp. 2414-2422, Feb. 2024.

[9] R. Jain and V. B. Semwal, "A novel feature extraction method for preimpact fall detection system using deep learning and wearable sensors," IEEE Sensors Journal, vol. 22, no. 23, pp. 22 943–22 951, 2022.

[10] "Surrey Audio-Visual Expressed Emotion (SAVEE) database", 2011. http://kahlan.eps.surrey.ac.uk/savee/ (accessed Apr. 26, 2024).

[11] "RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)". https://paperswithcode.com/dataset/ravdess (accessed Apr. 26, 2024).

[12] "Toronto emotional speech set (TESS)" University of Toronto TSpace. https://tspace.library.utoronto.ca/handle/1807/24487 (accessed Apr. 26, 2024).

[13] "Indian Emotional Speech Corpora (IESC)" https://www.kaggle.com/datasets/ybsingh/indian-emotional-speech-corpora-iesc (accessed May 18, 2024).

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," Procs. of the IEEE Conferece on Computer Vision and Pattern Recognition, pp. 779-788, 2016.

[15] L. Tan and C. He, "Abnormal event detection based on appearance repair and motion consistency," in Third International Conference on Computer Science and Communication Technology (ICCSCT 2022), vol. 12506. SPIE, 2022, pp. 967–972.

[16] Y. Hao, Z. Tang, B. Alzahrani, R. Alotaibi, R. Alharthi, M. Zhao, and A. Mahmood, "An end-to-end human abnormal behavior recognition framework for crowds with mentally disordered individuals," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 8, pp. 3618–3625, 2021.

[17] J. Li, Q. Huang, Y. Du, X. Zhen, S. Chen, and L. Shao, "Variational abnormal behavior detection with motion consistency," IEEE Transactions on Image Processing, vol. 31, pp. 275–286, 2021.