REMONI: An Autonomous System Integrating Wearables and Multimodal Large Language Models for Enhanced Remote Health Monitoring

Thanh Cong Ho*, Farah Kharrat*, Abderrazek Abid*, Fakhri Karray*†
*Mohamed bin Zayed University of Artificial Intelligence
Masdar City, Abu Dhabi, UAE
Email: {Thanh.Ho, Farah.Kharrat, Abid.Abderrazek, Fakhri.Karray}@mbzuai.ac.ae
†Department of Electrical and Computer Engineering
University of Waterloo, Waterloo, ON, Canada N2L 3G1
Email: karray@uwaterloo.ca
‡College of Computer and Information Sciences
Prince Sultan University
Email: akoubaa@psu.edu.sa

***Abstract*—With the widespread adoption of wearable devices in our daily lives, the demand and appeal for remote patient monitoring have significantly increased. Most research in this field has concentrated on collecting sensor data, visualizing it, and analyzing it to detect anomalies in specific diseases such as diabetes, heart disease and depression. However, this domain has a notable gap in the aspect of human-machine interaction. This paper proposes REMONI, an autonomous REmote health MONItoring system that integrates multimodal large language models (MLLMs), the Internet of Things (IoT), and wearable devices. The system automatically and continuously collects vital signs, accelerometer data from a special wearable (such as a smartwatch), and visual data in patient video clips collected from cameras. This data is processed by an anomaly detection module, which includes a fall detection model and algorithms to identify and alert caregivers of the patient's emergency conditions. A distinctive feature of our proposed system is the natural language processing component, developed with MLLMs capable of detecting and recognizing a patient's activity and emotion while responding to healthcare worker's inquiries. Additionally, prompt engineering is employed to integrate all patient information seamlessly. As a result, doctors and nurses can access real-time vital signs and the patient's current state and mood by interacting with an intelligent agent through a user-friendly web application. Our experiments demonstrate that our system is implementable and scalable for real-life scenarios, potentially reducing the workload of medical professionals and healthcare costs. A full-fledged prototype illustrating the functionalities of the system has been developed and being tested to demonstrate the robustness of its various capabilities.**

***Index Terms*—Remote Health Monitoring, Wearable Technology, Multimodal Large Language Models, Healthcare.**

I. Introduction

The imbalance between the number of patients and medical professionals has always been a significant challenge for the healthcare system. While the population is rapidly increasing, the workforce of medical professionals is growing slowly due to the high requirements of this field. This leads to a high and stressful workload for doctors and nurses, which could affect the quality of care services.

Since the development of the Internet of Things (IoT), many studies have tried to apply it to the healthcare environment to support doctors and nurses with their daily tasks, such as systematizing medical records [1], automating medical appointment booking [2], and moving essential medical examinations to telehealth [3]. Recently, large language models (LLMs) have been experiencing significant growth. Numerous LLMs related to healthcare have been developed [4]–[7] to improve human-machine interaction, enabling them to respond to medical inquiries from patients or questions in medical textbooks. However, to date, there has been limited research on creating virtual assistants for medical professionals.

This paper proposes REMONI, an autonomous REmote health MONItoring system. Our system has two main components: an anomaly detection module and a smart natural language processing component capable of recognizing patient status (activity and emotion). These two components are built with a scalable and implementable IoT system to operate in practice. The main functions of the system are to promptly report the patient's abnormalities to the doctor, remotely monitor the patient's health, and support the doctor in retrieving patient data.

Our contributions are as follows:

- To the best of our knowledge, this is the first work for the concept of an all-in-one virtual assistant for medical professionals, which automatically gathers data from sensors, stores it, detects anomalies, and uses LLMs to converse with doctors.
- We propose an IoT system architecture that connects sensors, wearable devices, LLMs, and end-user web applications using edge devices, cloud storage, and cloud computing. The architecture is flexible and easy to scale

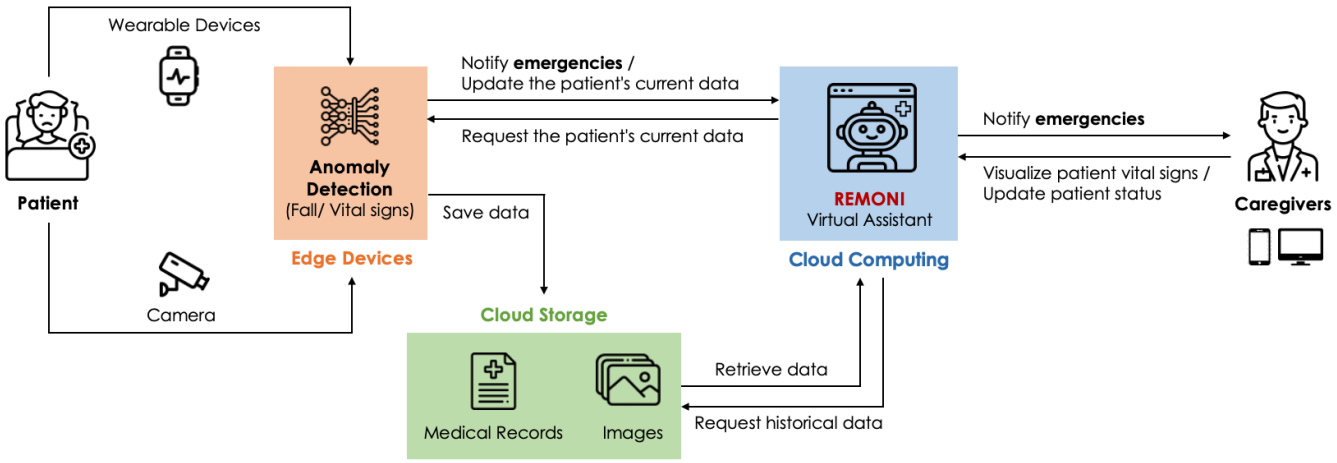

Fig. 1. Proposed Internet of Things System.

to a larger number of sensors and users, like in a hospital system.

- We demonstrate the medical virtual assistant's effectiveness and potential in improving the telehealth field.

II. Related Work

For a long time, researchers have been keen on developing virtual assistants for the healthcare sector to enhance human-machine interaction and improve care service quality.

During the early stages of Computer Vision development with Convolutional Neural Networks (CNNs), and when Natural Language Processing (NLP) was still in its infancy, Yoon et al. [8] created a projection-based augmented reality system to assist the elderly daily. They developed a camera capable of panning and tilting to monitor a wide area. Additionally, they utilized some deep learning frameworks for posture estimation, facial recognition, and object detection. With these capabilities, they constructed a system to project images onto surfaces (walls or tables) for the elderly to interact with. This system enabled the elderly to use their phones to see who was at the door and decide whether to grant entry. It also displayed daily reminders for taking medications or closing doors.

As NLP advanced, Nikita et al. [9] proposed a chatbot that employed machine learning techniques with metrics like TF-IDF, Stemming, n-grams, and cosine similarity to analyze users' healthcare-related queries. They aimed to create a chatbot to offer patients preliminary advice before consulting a doctor.

In the past two years, with the rapid development of LLMs, several medical chatbots have been created [4]–[7].These chatbots primarily focus on answering patients' medical questions. Yunxiang et al. [10] constructed a dataset of authentic patient-doctor conversations, gathering around 100k interactions from the online medical consultation website HealthCareMagic and an additional 10k conversations from another online medical consultation site, iCliniq.

While there is significant research in this field, it is evident that most efforts are centered on the patient side. A noticeable gap exists in developing virtual assistants specifically designed for medical professionals.

III. Proposed Framework

The framework proposed in this study comprises three modules: an Anomaly Detection module, a Natural Language Processing Engine, and an Internet of Things module for system deployment. Each of these is described in details below.

A. Anomaly Detection

Within our system, anomaly detection is a crucial process for recognizing irregular patterns that may indicate falls or critical changes in the patient's health status.

1) FallAllD Dataset for Fall Detection:

- **Dataset Description:** This study uses the FallAllD dataset [11], capturing 35 human falls and 44 daily activities, collected with three data loggers in an outdoor environment, where the subjects fell on grass instead of on soft mats. Each data-logger is equipped with the inertial module LSM9DS1, containing a 3-axial accelerometer configured with a sampling frequency of 238 Hz and a measurement range of ± 8 g. Fifteen participants wore these devices simultaneously on their necks, wrists, and waists to comprehensively capture their motion signals. The data collection protocol covers various fall types, including scenarios followed by recovery, and considers different postures and causes such as slips, trips, and syncope. This inclusivity is crucial for developing fall detection systems that are robust and effective across diverse real-world situations.
- **Dataset Preprocessing:** In preprocessing the FallAllD dataset, we focused on the accelerometer data obtained from the subjects' wrists. This choice is supported by the ability of accelerometer data to accurately capture motion patterns and dynamics critical to identifying falls, distinguishing them from activities of daily living (ADL) with high reliability. The literature further validates our approach as many researchers acknowledge the sufficiency of accelerometer data for effective fall detection [12]–[14]. Furthermore, we

rescaled the original FallAllD dataset's accelerometer data from $\pm 8\,\mathrm{g}$ to $\pm 1\,\mathrm{g}$ and downsampled it from $238\,\mathrm{Hz}$ to $32\,\mathrm{Hz}$. Fig. 2 compares the original FallAllD dataset and the preprocessed data for one random fall activity. It's evident that despite the reduced sampling rate and the range, the peaks associated with the fall are still clear in the preprocessed data, which indicates that the preprocessing retains the critical features necessary for fall detection.

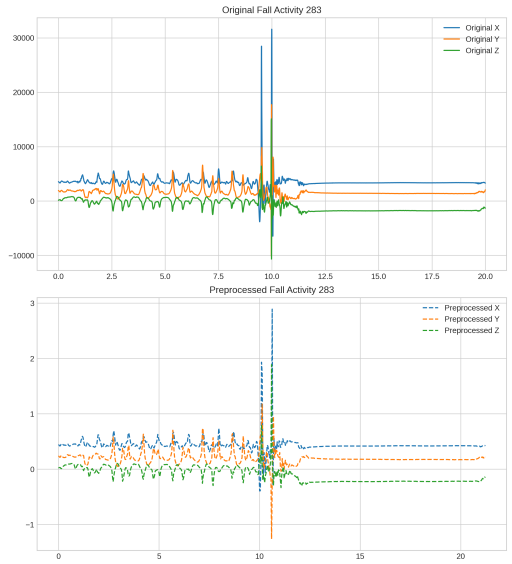

Fig. 2. Comparison of Original and Preprocessed Fall Activity from FallAllD Dataset.

2) Proposed Hybrid Deep Learning Model: We build upon the foundation of our previously proposed hybrid deep learning (HDL) model [15], which combines the strengths of convolutional neural networks (CNNs) for spatial feature extraction and long short-term memory (LSTM) networks for temporal sequence analysis, to address the time series challenge of human activity and fall detection. Retaining the model's core architecture, we have implemented minor refinements to enhance its performance further. The revised model, employing a sigmoid activation function in the final layer for optimal binary classification and compiled with an Adam optimizer and binary cross-entropy loss, offers enhanced performance and reliability in our real-world application.

3) Threshold-Based Assessment of Vital Sign Anomalies: In addition to the fall detection capability, our system integrates a threshold-based algorithm to detect anomalies in vital signs and automatically alert caregivers in case of health risks or emergencies. This Python model employs predefined healthy ranges for five key vital signs, namely body temperature (36.5–37.2°C), heart rate (60–100 beats per minute), respiration rate (12–20 breaths per minute at rest), blood pressure ($90/60$ to $120/80$ mmHg) [16], and oxygen saturation ($SpO_2 \geq 95\%$) [17]. REMONI continuously monitors these vital signs and upon detecting any values outside the healthy range, it promptly sends a notification alerting the medical personnel.

B. Natural Language Processing Engine

In the system, medical professionals communicate directly with REMONI, a virtual assistant, through a web application. REMONI is powered by an NLP engine, as shown in Fig. 3, which operates in three main stages: intention detection, data preparation, and final output production.

Initially, caregiver inquiries are processed by a General Large Language Model (LLM) to discern the user's intent. A system prompt, crafted to detail the intent detection task, is supplied to the General LLM alongside the doctor's question. This enables the General LLM to identify and output the user's intention in JSON format with the following keys: $patient\_id$, $list\_date$, $list\_time$, $vital\_sign$, is_plot, $is\_recognition$, and is_image.

In the second stage, the engine utilizes the $list\_date$ and $list\_time$ keys to determine whether to retrieve necessary data from cloud storage or the edge device. The is_plot and $is\_recognition$ keys indicate whether a plotting function or a MLLM is required to support the final response. The MLLM, equipped with its system prompt, is tasked with describing the patient's current status in an image, focusing on activity and emotion, which are crucial for depicting the patient's condition.

Finally, after collating all the information, the agent compiles it into an endpoint prompt that includes the patient's personal information, activity and emotion (as identified by the MLLM), vital signs (sourced from medical records in cloud storage or the edge device), and the user's question as instructions. A system prompt at the beginning of the endpoint prompt instructs the General LLM to use the data in the prompt to answer the user's questions accurately.

In summary, the NLP engine employs a General LLM for two distinct tasks (intent detection and final output production), each with a specific system prompt to guide the LLM's output according to the task. Additionally, the engine features a function library that includes a MLLM for recognizing the patient's activity and emotion and a plot function for visualizing the vital sign data.

C. Internet of Things System

The proposed IoT system comprises four main components: sensors, edge devices, cloud storage, and cloud computing. An overview of the architecture is illustrated in Fig. 1. Currently, our system is compatible with smartwatches and cameras as sensors.

At the core of our deployment lies the seamless integration of wearable technology. A custom wearable application serves as the conduit for data acquisition, harnessing the watch's sensors to collect accelerometer and physiological data. This application uses the Wi-Fi API to transmit the collected data streams to a designated Python module hosted on an edge device for further processing.

The edge device collects real-time data (accelerometer readings, vital signs, and visual information) from these sensors. It processes it through anomaly detection models (for falls and vital signs) to quickly identify emergencies. The device

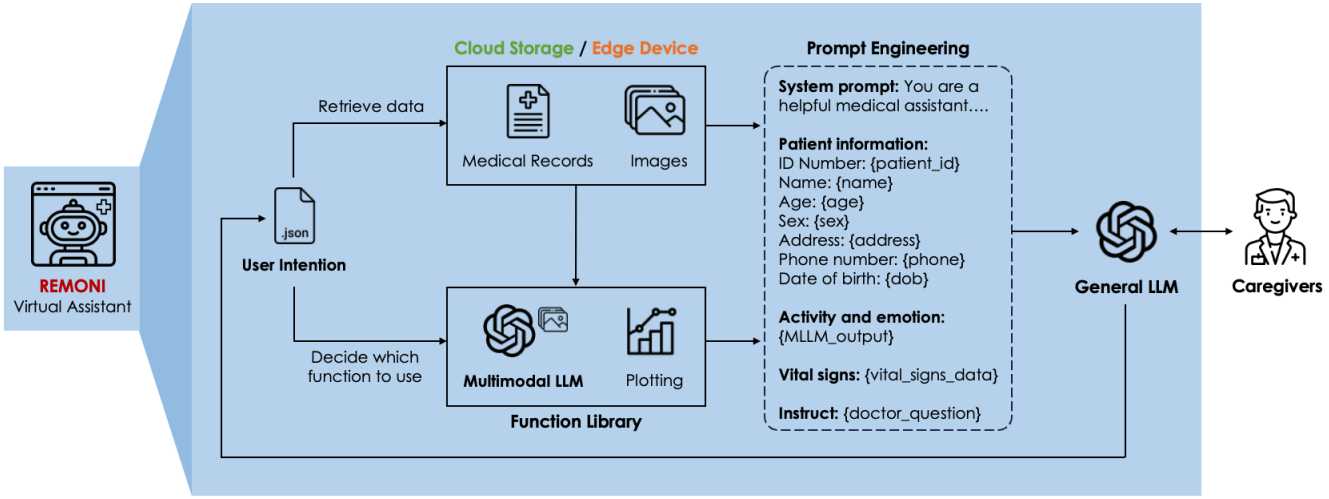

Fig. 3. Natural Language Processing Engine.

immediately sends an alert to the cloud computing platform if an emergency is detected. Otherwise, it periodically uploads vital signs and visual data to cloud storage for preservation and future use. Cloud storage archives historical data and responds to requests from the NLP engine for the data required to address user queries.

Regarding cloud computing, it serves as the host for the web application. Its primary role is to receive emergency alerts from the edge device and notify medical professionals. Additionally, it facilitates the NLP engine's communication with the edge device for current data and with cloud storage for historical data.

IV. Experiments

A. Deployment

The implementation of REMONI involves a carefully orchestrated setup of interconnected components, all operating concurrently on the edge device, each contributing to the system's robustness and effectiveness in remote health monitoring.

The project utilizes the ACER Nitro AN515 as the edge device, which is equipped with an i5 9th-gen CPU, a GTX 1660 TI GPU, 8 GB of RAM, and 1 TB SSD storage. For imaging, we selected the Logitech 720p HD Webcam for its image quality and ease of use.

For data acquisition, we deployed a Tizen wearable web app specifically designed for the Samsung Galaxy Watch 3 (4GB of RAM, 380 mAh battery, multiple physiological sensors), chosen for its advanced sensors, reliability, and seamless integration with our system. AWS S3 and AWS EC2 were selected for their stable performance for cloud storage and computing.

These intricately integrated components form a robust deployment framework that delivers scalable, reliable, and responsive remote health monitoring capabilities to caregivers and healthcare professionals.

B. Results and Discussion

To assess the system's effectiveness, we examine the performance of the fall detection model, which utilizes accelerometer data from the Samsung Galaxy Watch 3. Additionally, we evaluate the accuracy of activity and emotion recognition using MLLMs and analyze the implementation time required for various tasks within the system.

1) Fall detection model: Fig. 4 illustrates the training and validation accuracy over epochs. The model quickly reaches a high level of accuracy during training, which is consistently maintained throughout the validation phase. The convergence of training and validation accuracy suggests that the model generalizes well and exhibits minimal overfitting.

The confusion matrix in Fig. 5 displays the model's performance in classifying fall and non-fall events in the test data from the preprocessed FallAllD dataset. The high values on the matrix's diagonal (0.99 for Non-Fall, 0.98 for Fall) indicate a high true positive rate for both classes, demonstrating the model's accuracy. The low off-diagonal values (0.01 and 0.02) show a small fraction of misclassifications, confirming the model's reliability in fall detection.

Table I provides a comparative analysis of various models tested on the FallAllD dataset, such as Light Gradient-Boosting Machine (LightGBM), Support Vector Machine (SVM), Coarse-fine CNN combined with Gated Recurrent Unit (GRU), and Late AFVF (feature-level fusion applied to Actual Fusion within Virtual Fusion).The proposed HDL model demonstrates superior performance with a precision of 99%, an accuracy, recall, and F1-score all at 98%. This highlights the model's high performance using solely wrist-acquired accelerometer data, underscoring its potential for reliable and accurate fall detection in real-world applications.

2) Activity and Emotion Recognition: To assess the system's capability to recognize human activity and emotion, we selected the HACER dataset [20] for its high relevance to the project. The dataset comprises around 500 video clips featur-

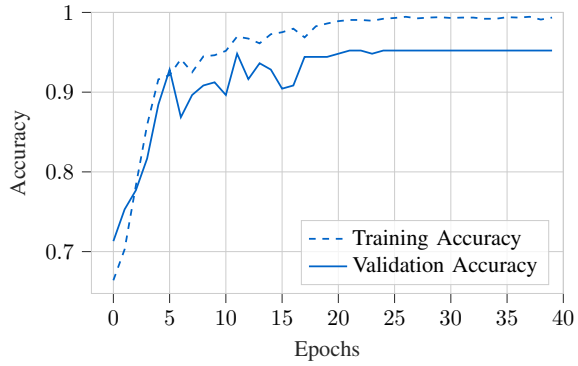

Fig. 4. CNN-LSTM Traning vs. Validation Accuracy on the Preprocessed FallAllD dataset.

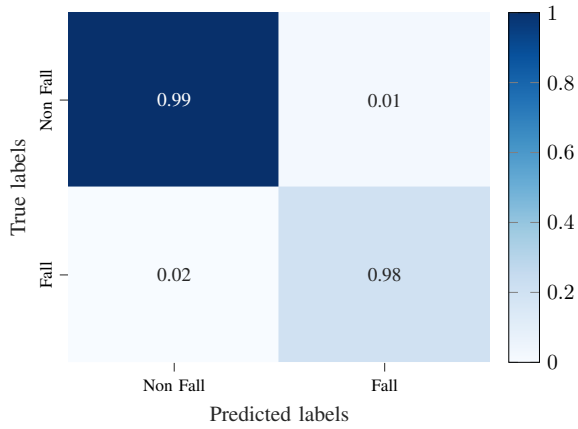

Fig. 5. CNN-LSTM Model: Confusion Matrix for Test Data on the Preprocessed FallAllD Dataset.

TABLE I
BENCHMARKING FOR THE FALLALLD DATASET

Model	Sensor/Location	A	P	R	F1
LightGBM [18]	Acc+Gyro/Wrist	95%	N/A	91%	91%
SVM [19]	Acc+Gyro/Wrist	93%	N/A	87%	93%
CNN-GRU [13]	Acc/Waist	98%	96%	93%	94%
Late AFVF [14]	Acc/Waist+Wrist	N/A	N/A	N/A	96%
Proposed model	Acc/Wrist	**98%**	**99%**	**98%**	**98%**

Acc: Accelerometer, Gyro: Gyroscope, A: Accuracy, P: Precision, R: Recall, F1: F1-score

ing ten participants, including men and women. Each video is labelled with both activity and emotion categories. The dataset features nine activity classes: *drinking*, *putting_on_glasses*, *putting_on_jacket*, *reading*, *sitting_down*, *standing_up*, *taking_off_glasses*, *taking_off_jacket*, and *writing*. Additionally, it includes five emotion classes: *angry*, *disgust*, *happy*, *neutral*, and *sad*.

We crafted a prompt similar to the system prompt used by the MLLM in our system. The difference is that we explicitly list the activity and emotion classes in the system prompt and ask the models to identify the correct one. If no person is visible in the image, the models should respond with 'unidentifiable' for both activity and emotion. Similarly, if the person's face is unclear, particularly in surveillance footage, the models should output 'unidentifiable' for the emotion while still attempting to identify the activity.

TABLE II
PERFORMANCE OF MLLMS IN ACTIVITY AND EMOTION RECOGNITION

Model	Accuracy	Precision	Recall	F1-score
LLaVA-Activity	10%	22%	10%	7%
Video-ChatGPT-Activity	35%	49%	35%	33%
GPT4-Vision-Activity	**51%**	**58%**	**51%**	**45%**
LLaVA-Emotion	15%	**39%**	15%	12%
Video-ChatGPT-Emotion	20%	**39%**	20%	19%
GPT4-Vision-Emotion	**41%**	35%	**41%**	**31%**

For evaluation, we utilized three multimodal large language models: LLaVA [21], Video-ChatGPT [22], and GPT4-Vision [23]. LLaVA and GPT4-Vision analyze a single frame, specifically the middle frame extracted from the video. In contrast, Video-ChatGPT processes approximately 100 frames.

As shown in Table II, LLaVA had the lowest performance, with 10% accuracy for activity and 15% for emotion. This may be due to its reliance solely on image input. Additionally, it frequently classifies activities and emotions as 'unidentifiable', suggesting a lack of focus on human faces or poses for activity and emotion detection.

While the training process of Video-ChatGPT enriched the dataset's context, enabling it to provide more detailed responses, it ranked second with an accuracy of 35% for activity recognition and 20% for emotion detection. The model's propensity for lengthy answers sometimes resulted in missing the correct classes, which did not align with our need for straightforward recognition of emotions and activities.

GPT4-Vision showed the best overall performance, with 51% accuracy for activity and 41% for emotion, despite only using the middle frame of the video. Its performance significantly surpasses that of LLaVA. However, a closer examination of the precision values for emotion reveals that GPT4-Vision tends to output 'happy' and 'neutral' as the emotion class, resulting in slightly lower precision. In contrast, despite their lower accuracy, the other two models produce outputs that span across various emotion classes.

Although GPT4-Vision scores the highest in most metrics, its performance is still relatively low. This may be attributed to our stringent criteria for determining accurate classifications. For instance, if a person is sitting and drinking water, the dataset classifies it as drinking, but if the models detect it as sitting, it is considered a false detection. We are in the process of developing a more equitable evaluation method. Additionally, the system is set to be implemented in a real-world scenario shortly and will undergo further fine-tuning with actual data from a clinic in Abu Dhabi. As a result, we anticipate an improvement in the performance of the MLLMs. For the demo phase, GPT4-Vision has been selected as the MLLM for the NLP engine.

3) Response Time Delays: The overall system latency was evaluated through three main approaches, each encompassing

various questions based on the data requested. Each scenario's time is measured ten times before calculating the average and variance.

Firstly, as shown in Table III, the response time to queries about real-time (instant) data, such as the patient's current state or vital signs, is assessed. . These questions necessitate communication between the cloud computing system and the edge device to retrieve the latest sensor data and provide answers. Moreover, the latency in responding to requests for historical information stored in the cloud database is also investigated.

TABLE III
RESPONSE TIME ANALYSIS OF THE REMONI SYSTEM FOR VARIOUS QUESTION TYPES

Question Type \Data type	Instant Data	Historical Data
Image Retrieval	$6.84s \pm 2.09s$	$3.37s \pm 0.15s$
Vital Sign Retrieval	$5.69s \pm 0.06s$	$6.62s \pm 8.73s$
Image Description	$9.56s \pm 6.11s$	$9.75s \pm 9.70s$
Plotting Vital Sign	N/A	$4.61s \pm 1.58s$

Overall, the system's response time to user queries is under 20 seconds. Questions requiring the NLP engine to use MLLM to determine the patient's current state (activity and emotion) take longer to answer. Additionally, retrieving instant image data from edge devices takes much longer than accessing historical image data from cloud storage.

Lastly, the system's responsiveness in sending emergency alerts from the edge device to the caregiver was examined, with response times averaging around 0.2 seconds.

V. CONCLUSION

Our proposed system, REMONI, is an autonomous system that enhances human-machine interaction in REmote health MONItoring. Utilizing MLLMs, IoT, and wearable devices, REMONI provides a comprehensive solution for the continuous and automatic collection of vital signs, accelerometer data, and visual data. It also facilitates anomaly detection and seamless communication between medical professionals and the system. The system is designed to be easily scalable to larger environments involving multiple medical professionals and many patients. Additionally, more wearable devices and anomaly detection algorithms can be quickly integrated into our proposed IoT system architecture. Our experiments demonstrate the system's feasibility for real-life applications. We believe in its potential to alleviate the workload of medical professionals and reduce healthcare costs. In future work, we plan to deploy the system in real-life settings and further enhance the performance of its components.

REFERENCES

[1] A. Subahi, "Edge-based iot medical record system: Requirements, recommendations and conceptual design," *IEEE Access*, vol. PP, pp. 1–1, 07 2019.

[2] Ruchit, P. Suryavanshi, S. Kaushik, and D. Dev, "An automated model for booking appointment in health care sector," in *2021 International Conference on Technological Advancements and Innovations (ICTAI)*, 2021, pp. 7–10.

[3] L. Romeo, R. Marani, T. D'Orazio, and G. Cicirelli, "Video based mobility monitoring of elderly people using deep learning models," *IEEE Access*, vol. 11, pp. 2804–2819, 2023.

[4] Q. Lu, D. Dou, and T. Nguyen, "ClinicalT5: A generative language model for clinical text," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5436–5443. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.398

[5] Z. Chen, A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, A. Sallinen, A. Sakhaeirad, V. Swamy, I. Krawczuk, D. Bayazit, A. Marmet, S. Montariol, M.-A. Hartley, M. Jaggi, and A. Bosselut, "Meditron-70b: Scaling medical pretraining for large language models," 2023.

[6] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B. A. y Arcas, N. Tomasev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. Barral, D. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, and V. Natarajan, "Towards expert-level medical question answering with large language models," 2023.

[7] T. Han, L. C. Adams, J.-M. Papaioannou, P. Grundmann, T. Oberhauser, A. Löser, D. Truhn, and K. K. Bressem, "Medalpaca–an open-source collection of medical conversational ai models and training data," *arXiv preprint arXiv:2304.08247*, 2023.

[8] Y. Park, H. Ro, N. Lee, and T.-D. Han, "Deep-care: Projection-based home care augmented reality system with deep learning for elderly," *Applied Sciences*, vol. 9, p. 3897, 09 2019.

[9] N. V. Shinde, A. Akhade, P. Bagad, H. Bhavsar, S. Wagh, and A. Kamble, "Healthcare chatbot system using artificial intelligence," in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2021, pp. 1–8.

[10] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge," *Cureus*, vol. 15, no. 6, 2023.

[11] M. Saleh, M. Abbas, and R. B. L. Jeannès, "Fallalld: An open dataset of human falls and activities of daily living for classical and deep learning applications," *IEEE Sensors Journal*, vol. 21, pp. 1849–1858, 2021.

[12] K.-C. Liu, K.-H. Hung, C.-Y. Hsieh, H.-Y. Huang, C.-T. Chan, and Y. Tsao, "Deep-learning-based signal enhancement of low-resolution accelerometer for fall detection systems," *IEEE TCDS*, vol. 14, pp. 1270–1281, 2022.

[13] C.-P. Liu, J.-H. Li, E.-P. Chu, C.-Y. Hsieh, K.-C. Liu, C.-T. Chan, and Y. Tsao, "Deep learning-based fall detection algorithm using ensemble model of coarse-fine cnn and gru networks," 2023.

[14] D.-A. Nguyen, C. Pham, and N.-A. Le-Khac, "Virtual fusion with contrastive learning for single sensor-based activity recognition," *arXiv preprint arXiv:2312.02185*, 2023.

[15] F. Kharrat, W. Gueaieb, F. Karray, and A. Elsaddik, "A hybrid deep learning model for human activity recognition and fall detection for the elderly," in *2023 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE, 2023, pp. 1–6.

[16] Health Encyclopedia, "Vital signs," (Accessed: Mar. 21, 2024). [Online]. Available: https://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentTypeID=85&ContentID=P00866

[17] National Library of Medicine, "Oxygen saturation," (Accessed: Mar. 21, 2024). [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK525974/

[18] J.-K. Kim, K. Lee, and S. G. Hong, "Detection of important features and comparison of datasets for fall detection based on wrist-wearable devices," *Expert Systems with Applications*, vol. 234, p. 121034, 2023.

[19] T. O. Nuñez, A. Ramirez, and L. B. Marón, "Analysis of waist and wrist positioning wearable machine learning models to detect falls," Jan. 2024.

[20] S. Ammar, T.-C. Ho, F. Karray, and W. Gueaieb, "Hacer: An integrated remote monitoring platform for the elderly," in *2023 International Conference on Intelligent Metaverse Technologies & Applications (iMETA)*, 09 2023, pp. 1–6.

[21] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[22] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *arXiv:2306.05424*, 2023.

[23] OpenAI, "Gpt-4 technical report," *ArXiv*, vol. abs/2303.08774, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774