International Journal on Advances in Life Sciences

2025 vol. 17 nr. 3&4

The *International Journal on Advances in Life Sciences* is published by IARIA.
ISSN: 1942-2660
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Life Sciences
Volume 17, Number 3&4, 2025

International Journal on Advances in Life Sciences
Volume 17, Numbers 3&4, 2025

CONTENTS

Pandemic-Related Social Media Content

Sujata Patil, KLE Technological University, India
Les Sztandera, Thomas Jefferson University, USA
Hemalatha Kivudujogappa Lingappa, Sri Krishna Institute of Technology, India

Realizing Seamless Connection Between Real and Virtual Spaces via Adaptive Virtual Reality: Objective Description of Cognitive Discrepancies Between These Spaces

Katsuko T. Nakahira
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: katsuko@vos.nagaokaut.ac.jp

Taichi Nakagawa
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: s223308@stn.nagaokaut.ac.jp

Thalpe Liyanage Amila Nirosh Chandrasiri
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: s245065@stn.nagaokaut.ac.jp

Nobuyuki Ogawa
National Institute of Technology(KOSEN), Gifu College
Motosu, Gifu, Japan
Email: ogawa@gifu-nct.ac.jp

Kuniaki Yajima
National Institute of Technology(KOSEN), Sendai College
Sendai, Miyagi, Japan
Email: yajima@sendai-nct.ac.jp

Muneo Kitajima
Nagaoka University of Technology
Nagaoka, Niigata, Japan
Email: mkitajima@kjs.nagaokaut.ac.jp

***Abstract*— Although the implementation of "Adaptive Virtual Reality" is becoming feasible, understanding the main effects of its realization on users based on cognitive models is essential. Here, we first described a model of the flow of information obtained by actual human perception through avatars in virtual reality (VR) and the resulting human reactions, and confirm the validity of the user models proposed thus far. We also considered the degree of immersion predicted due to the integration of multi-modal information. The cognitive processes of VR experiences are largely categorized into "perception and recognition of information (attention, memory, and decision making)" and "perception-based physical actions and interactions with VR objects". Based from this, we describe a cognitive model of VR experiences. In addition, as examples of the discrepancies in sensory perception experienced in real/VR spaces, we briefly describe the phenomena that occur in communication. We describe the cognitive models for these phenomena and qualitatively consider the degree to which sensory information obtained from the real/VR space affects the degree of chunks activation. The intensity of human sense is expressed as a logarithm according to Weber-Fechner's Law, suggesting that human senses can distinguish differences even with weak sensory information. We argue that the "slightly different from the real world" sense felt in VR content is caused by such slight differences in sensory information. Overall, we advance the cognitive understanding of the immersive experience particularly in the VR space, and qualitatively describe the possibility of designing highly immersive VR content which are adapted to each individual.**

Keywords*— *sensory perception; cognitive model; virtual reality; experience.

I. Introduction

This paper is based on the previous work originally presented in AIVR2024 [1]. The following changes were made: (1) a restructuring of the model diagram, (2) the addition of preliminary experimental results, and (3) the addition of corresponding discussions.

A concept called "Adaptive Virtual Reality (Adaptive VR)" has been discussed in recent years. Baker and Fairclough [2] described it as follows: Adaptive VR monitors human behavior, psychophysiology, and neurophysiology to create a real-time model of the user. This quantification is used to infer the emotional state of individual users and induce adaptive changes within the virtual environment during runtime. Therefore, the authors argued that the efficacy of the emotional experience can be increased by modeling individual differences in the way users interact within a particular virtual environment as a system.

Several methods exist for inferring emotional states. The most classical approach, following Russell's circular model [3][4], involves measuring a person's valence and arousal states to predict their emotions. These are often predicted through measurements such as pupil response or heart rate. Specifically, regarding visual behavior, Bao et al. [5] proposed a method to recognize learners' emotional states during distance learning, suggesting emotion recognition techniques for relatively simple emotions such as interest, happiness, confusion, and boredom. Furthermore, Sun et al. [6] suggested that arousal related to cognitive effort interacted informationally with luminance, and that the strongest pupil response due to arousal occurred at luminances below 37 cd/m^2.

Given this background, the implementation of Adaptive VR is becoming feasible. However, the main effects of its implementation on users should be understood based on a cognitive model. Studies have mainly focused on bottom-up content design with an awareness of Adaptive VR. However, it is difficult for empirical developments to provide effects that create a new phenomena. Hence, not only a bottom-up but also a top-down approach is necessary.

On the other hand, with the growth in Virtual Reality (VR) goggles and the low cost of equipment for shooting omnidirectional video, VR content has attracted substantial attention. In addition to games, a wide range of VR contents have been developed, including omnidirectional video playback, education, sightseeing, property previews, and shopping. VR systems that enable these contents to be viewed are also growing rapidly. For example, the following innovations have emerged in content design. VR systems using Head-Mounted Displays (HMDs) sold to general consumers cover the user's field of vision; thus, the user cannot see their own body. Therefore, VR systems using HMDs typically display a virtual body drawn from the user's first-person perspective. A mechanism for realizing the user's first-person perspective is the implementation of avatars. The effects of avatars have been described by researchers. Steed et al. [7] suggested that the use of avatars that follow the user's movements can reduce the cognitive load of certain tasks in the VR space. People around the world have been using VR social networking services, such as VRChat, where users enjoy interacting with other users using avatars that they have selected and edited to their liking. This shows that avatars are a means of self-expression in VR communication.

There are many research approaches to VR contents and systems, including research from the perspective of Human Computer Interaction (HCI), research on the relation between VR and working memory (WM), research on the differences in sensory perception between the real world and VR, and research on Adaptive VR that incorporates individual adaptability into VR contents.

Among the studies from the perspective of HCI, Mousavi et al. [8] integrate Emotion Recognition (ER) and VR to provide an immersive and flexible environment in VR. This integration can advance HCI by allowing the Virtual Environment (VE) to adapt to the user's emotional state.

According to Batra et al. [9], the following requirements for VR are listed: First, the primary component called "visualization" enables human-machine interaction to approximate real life; Additionally, VR requires removing the barrier between the real world and the virtual world. Through these means, a series of simulation technologies must generate artificial tactile, auditory, olfactory, and sensory experiences grounded in reality. For this simulation, it is crucial to capture human cognitive characteristics multi-modally.

As a stepping stone to this goal, we do the following in this study. We describe a model of the flow of information obtained by actual human perception through avatars in VR and the resulting human reactions, and confirm the validity of the user models proposed so far. The degree of immersion predicted because of the integration of multi-sensory information is also discussed. Understanding the role of multi-sensory information can enable us to design VR contents for individual users and how we can control sensory perception.

The remainder of this article is organized as follows. We describe the sense-perception cognitive model on VR in Section II. Section III describes experiments conducted to investigate the behavioral characteristics of participants' information acquisition and attention direction within VR spaces, as well as the relationship between the degree of recall and the variety of perceptual information quantity and quality within the VR space. Section IV argues the relationship between the variety of perceptual information quality combinations and cognitive load.

II. Description of the Cognitive Model for Sensory in VR

In general, physical information in the VR space is represented as follows. Objects in the VR space (VR objects) are represented by computer graphics, and their behavior is based on a program previously written to interact with the environment and other objects. The sound in the VR space is provided by artificially preparing audio data that is predicted in advance to be uttered in the space, and is played continuously in a background music-like manner, or by using a sound engine controlled by the user. Specifically, in the latter case, it can be attached to a VR object and played when certain conditions are met. Comprised of these elements, all human activities and virtual experiences in the VR space are performed by using the avatar as one's own body. The avatar's movement is performed by tracking the user's real-world body movements. Tracking methods include three-point tracking, which consists of an HMD and two hand controllers, and full-body tracking, which uses motion capture and a tracking suit.

Consequently, the human experience in the VR space differs slightly from perception and cognition in the real world, and can be said to be the result of the interaction between avatar and VR objects, as well as the perception of the accompanying environment such as sound linked to these objects. Considering this, the model of human perception, cognition, and behavior in the VR space should be described with an awareness of the various interactions in the VR space with those in the real world.

Based on the above, the integration process of information perceived in both the real world and VR is shown in Figure 1. The concept is as follows.

A. Transformation of Perceptual Information Provided by Objects

The perceptual information of an object existing in real space is expressed through the five senses—visual, auditory, touch (somatosensory), smell, and taste—in a form adapted to our sensory organs. Perceptual information directly received by humans from the real world is received without attenuation beyond the capabilities of the individual's sensory organs. However, in VR, analog-to-digital conversion is applied to the perceptual information possessed by real-world objects. Consequently, this information exists within the VR space in a form where some information is missing. This means that in Figure 1, the chunk c_j(transferred) provided by the object transferred to the virtual world—resulting from the analog/digital conversion of the analog perceptual information in chunk c_j(real) provided by the real-world object—exists in a form where information is missing.

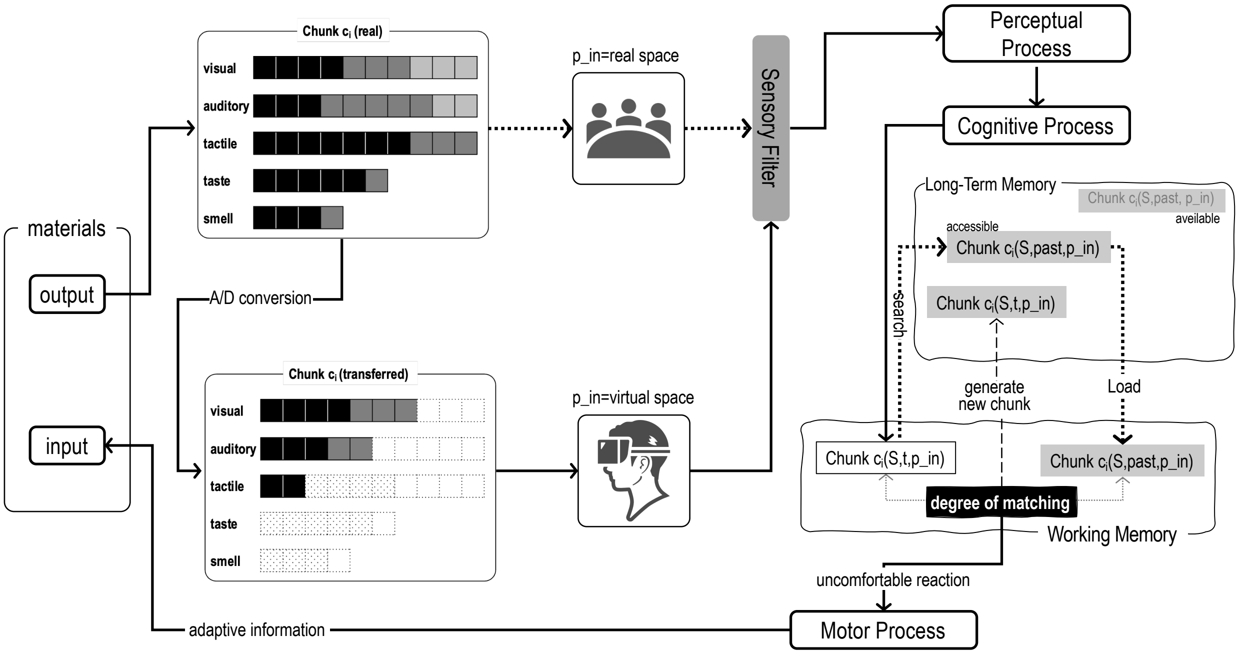


Figure 1. The information flow received by the senses in different environments.

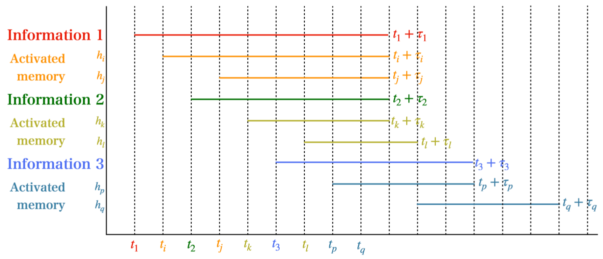


Figure 2. Staying timeline of sensory information stored in working memory and information invoked from long-term memory.

B. *Perception of Information Provided by Objects in Virtual World*

In general VR experiences using current HMDs, visual, auditory, and somatosensory information are used as perceptual information. The VR experience begins when the user puts on the HMD and views the images displayed on the lenses; by moving their head while wearing the HMD, the user can perceive the virtual space in the same way as they perceive the real world. Auditory information is output from the HMD's built-in or external speakers, and audio is played in response to the behavior of VR objects. The somatosensory information is used to make operations in the VR space clearer by vibrating the controllers in both hands to generate tactile feedback when operating the User Interface (UI) in the VR space or selecting VR objects.

C. *Cognition of Information in Virtual World*

The perception of information in virtual world is fundamentally no different from that in the real world. However, even when perceiving the same object, the provided information from the object is incomplete compared to that in the real world. The process differs in that it involves cross-referencing with similar memories. Thus, we describe the sequence of events as below.

1) *Attention:* Perceptual information moves to the sensory register, and then only the information to which the user's attention is directed passes through the selective filter and into the WM. Here, each sensory information does not completely enter the WM at the same time, but one piece of information passes through per processing.

2) *Memory:* If the sensory information obtained in the VR space is similar to that obtained in the real world, the user perceives the VR space as if it were a real space. In addition, based on the information in the Long-Term Memory (LTM), the user anticipates and expects the response of objects in the VR space to his or her actions, and engagement is generated.

3) *Decision:* Based on the perceptual information, the next action is determined. Here, when actions on a VR object are performed via a controller, the actions in the real world are converted into the corresponding controller operations.

D. *Body Movement Based on Perception*

The operator (actual body) moves, and the avatar in the VR space moves in response to the movement. There are two methods for incorporating human motion into VR:

- Image sensing by the camera attached to the HMD: Basic UI operations (clicking and screen scrolling) and grasping VR objects (realized by holding something with a hand gesture) are possible. The high degree of synchronization between the actual hand and the avatar's hand motion is an advantage of this method. Conversely,

TABLE I. The difference between past or current input information and situations, and the *degree of matching*

difference of situation	difference of input information		
	almost never	little	greatly
almost never	−	+	++
little	+	++	++
greatly	++	++	+++

precise manipulation, movements large enough to cause both hands to move out of the camera's field of view, and very fast hand movements are weaknesses.

- Yaw, pitch, roll + relative position by controller:
  The accurate tracking of position, posture, and motion information by sensors is possible, and the sense of actual body motion is directly reflected during the operation, resulting in a high sense of immersion. However, if the reflection of body motion by the HMD is not synchronized with the actual body motion, it may cause a sense of discomfort and reduce the immersiveness of the VR experience.

1) Interaction with VR Objects: VR objects not only appear to be three-dimensional, but can also be actually manipulated. Examples include playing a musical instrument or a push-button switch. Here, the immersiveness of the VR experience can be enhanced by providing not only a visual 3D effect, but also contextual information that one's actions affect the VR object.

E. Integration of Information Obtained in the Virtual World and Past Experiences

Based on Figures 1 and Table I, we consider the perception of a phenomenon in the real (R) or virtual (V) space as follows. The chunk C_j stored in the LTM is constructed from the information group $I_i^{env}(t)$ obtained from sensory organ i $(1 \leq i \leq 5)$ in the past. Here, i refers to the five sensory organs possessed by a person. Each $I_i^{env}(t)$ passes through the attention filter $F_i^{env}(t)$ via the sensory register. And at time t, only the information obtained from a specific sensory organ passes through. C_j contains the information obtained from each sensory organ as a set $I(t)$ and is denoted as $C_j(I(t))$. Here, $I(t)$ is represented as follows:

$$I(t) = \{\boldsymbol{I} \ | F_i^{env}(t) I_i^{env}(t),\ 1 \leq i \leq 5\}.$$

The information that has passed through the attention filter is stored in the WM for a specific time τ_k, and a set of information $I(t)$ is sent to the LTM at the same time or with a time lag. In the LTM, $C_j(I(t))$ is matched with $C_j(I(t))$ based on the information in $I(t)$, and the closest or matching $C_j(I(t))$ is used as knowledge. The used knowledge is overwritten in the LTM through the WM in the form that the information in $I(t)$ is enhanced. Here, we target two sensory organs – visual and auditory. We consider how the information flows through these three types of sensory organs in turn.

Suppose that at a certain time, a specific amount of information $I_i^{env}(t)$ $(1 \leq i \leq 5)$ is received from the external environment. $I_i^{env}(t)$ correspond to Information N in Figure 2. Information N simultaneously activates several chunks. Although the degree of chunk activation varies, $F_i^{env}(t) I_i^{env}(t)$ is integrated into a single piece of information and sent to the LTM. This difference, the integrated information $I^{syn}(t)$, can be expressed using the integration operator G as follows. However, since G depends on the individual, it does not take a unique form.

$$I^{syn}(t) = G(i, j, F_i^{env}(t) I_i^{env}(t), C_j(I(t)))$$

For the sake of simplicity, we simply add the amount of information and the degree of chunk activation as follows.

$$G^{env}(t) = \sum_{j}^{m} \sum_{i}^{n} F_i^{env}(t) I_i^{env}(t) C_j(I(t)) \qquad (1)$$

III. Information Acquisition and Attention Direction in Metaverse Space

If the cognitive framework described in Section II is correct, differences in sensory perception should be observed between the real and virtual worlds. The following are examples of what these differences in sensory perception might cause:

- Differences in memory quantity/quality:
  Information easily memorized in the real world may be difficult to contextualize in the virtual world, or vice versa.
- Differences in reaction:
  In the real world, even minor changes can trigger significant reactions. Conversely, in the virtual world, reactions may be difficult to elicit without substantial changes. Or the opposite may occur.

We consider the differences in sensory perception is caused by depending on the *degree of matching* within working memory, described in Figure 1, namely the value of $I^{syn}(t)$. We also consider $I^{syn}(t)$ as the cognitive load incurred during the integration of perceived similar information of past and current, the greater the divergence between the two pieces of information, the higher the load required to generate $I^{syn}(t)$. To realize this divergence, this research sets the number of objects in the virtual world as the excess or deficiency of integration targets for visual information, and the audio quality of explanatory narration for specific objects as the difficulty level of integration targets for auditory information. We confirm the possibility of measuring the load on information integration through the combination of these two factors. Based on the above, we propose the following hypotheses regarding the quality of visual and auditory information:

- Hypothesis 1: Different combinations of quality result in different cognitive load for information integration, and an optimal combination exists that provides the least load.
- Hypothesis 2: Consequently, differences are observed in the memory of information perceived during activities in a virtual world.
- Hypothesis 3: When the combination provides optimal load, the perceived cognitive load is closer to that experienced in the real world, leading to a sense of immersion.

In this study, we design experiments to gain insights into Hypotheses 1 and 2 and verify their validity.

To verify this, one method involves recreating real-world objects within a virtual world with appropriate explanations, then conducting visual behavior analysis and memory depth analysis using variables such as fixation time on the object and depth of memory for the explanation. The experimental method for this is described below.

A. The Configuration of the Target Virtual World and Experimental Conditions

To conduct experiments testing hypotheses, it is necessary to construct a virtual world and set the quality and quantity of objects. In this study, to perform trend analysis for the hypothesis, we designed the space as follows as a preliminary experiment.

The virtual world is structured with sightseeing in mind. Consequently, activities within the virtual world are as follows:

- Free exploration (primarily focused on acquiring visual information)
- Discovery of distinctive objects (discovery through visual information). We focuse on architectural structures.
- Receiving supplementary knowledge through explanatory narration on architectural styles and structures (learning involving auditory information).

The primary object is content featuring the construction of shrines—people often seen but rarely understood in detail in real world. A screenshot of the VR space used in the experiment and the intensity condition waveform of the sound source are shown in Figure 3.

The virtual world space was constructed using Unity. Within the VR space, a shrine model and an explanatory audio track about the shrine's construction were placed. For the shrine's 3D model, a commercially available standard architectural style was used. However, since the focus was specifically on learning architectural styles, decorative items that should be placed inside were excluded. The commentary audio is designed to play when the participant pushes the speaker icon. The content consists of standard explanatory text combined and read aloud by a automated voice, with only amplitude adjustments made. However, for the lower quality setting, a lowpass filter is applied to achieve telephone-like audio quality. Three explanatory audio clips were prepared for each shrine, with their auditory quality set to three varieties. For the building exteriors, objects related to shrines—such as trees, *torii* gates, and *chozuya* purification fountains—were placed to enhance visual and contextual information, making it closer to the real thing.

In this virtual world, each participant freely moves through the space, exploring both inside and outside the shrine. They memorize the space sometimes using only visual information, sometimes only auditory information, and sometimes by integrating both types of information. Therefore, a recall test for the memorized content can serve as an indicator of how information acquired in various ways is expressed.

For visual and auditory information, their quality was set according to the conditions shown in Table II. For visual information variations, three patterns were prepared:

- A simple plane with only a shrine (condition L),
- A simple plane with a shrine and related objects (trees, *torii* gate, *chozuya*) which represents condition N, and
- A shrine located within a forest, accompanied by a *torii* gate and *chozuya* which represents condition R.

For auditory information variations, we prepared three patterns:

- poor audio quality for the explanatory narration (with lowpass filter for normal sound) which represents condition L,
- standard commentary audio (default automated voice and no customize) which represents condition N, and
- consistently high volume (amplifying power) which represents condition R.

B. Experimental Procedure

The experiment was conducted as follows. The overall flow is shown in Figure 4. The HMD used for the experiment was the Meta Quest 2, and the Tobii Pro Glasses 3 were used for eye tracking. Data acquired with the Tobii Pro Glasses 3 was processed in Tobii Pro Labo to identify saccades, fixations, and obtain gaze point coordinates. Furthermore, to mitigate VR sickness, teleportation was adopted as the method of movement within the VR space. Before the experiment began, subjects were asked to answer questions regarding:

- Previous VR experience,
- Prior knowledge of architecture, and
- Prior knowledge of shrine construction.

Additionally, subjects were asked to answer several questions before the experiment began, including their knowledge of shrines, learning experiences, and whether they had recently visited a shrine. Sessions were conducted for each visual condition, and at the end of viewing each session,

- Did you feel as if you were actually present there?
- Did you feel the VR world was so realistic that you forgot the outside world?
- Did you feel like you were watching a video, or did you feel like you were actually in the space?

They were asked to rate their responses on an eight-point scale.

After the questionnaire, participants were given a tutorial and then experienced VR content with sensory information appropriately altered. To measure participants' gaze information, they wore an HMD over an eye tracker while experiencing VR content. Participants experienced one session consisting of an audio playback task with three different auditory conditions under the same visual condition, completing a total of three sessions for different visual conditions. At the end of each session, they answered questions about the content. After completing the three sessions, a recall test was conducted.

In the recall test, a free-response section was included where subjects were asked to write about what they remembered

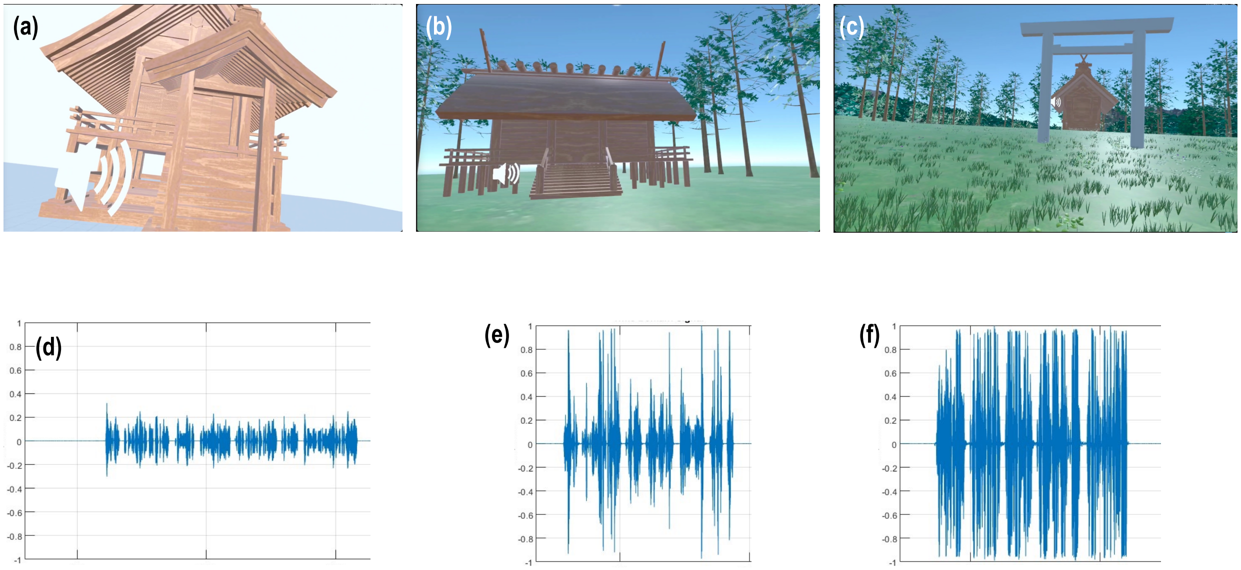


Figure 3. VR content presented to participants. Visual conditions (V) are: (a) less (weak) stimulus (L), (b) normal stimulus (N), (c) rich (strong) stimulus (R). Audio conditions (A) are: (d) low quality (L), (e) normal quality (baseline), (N)), and (f) rich quality (R).

TABLE II. VR content presentation stimulus change patterns.

	condition1 : weak stimuli L	condition2 : normal stimuli N	condition3 : strong stimuli R
visual V	Only the shrine is placed on the flat surface.	Arrange a shrine, 47 trees, a *torii* gate, and ground texture on a flat surface.	Arrange a shrine, 75 trees, three-dimensional terrain, grass, *torii* gates, a *chozuya*, and background on a flat surface.
auditory A	A muffled sound quality like over the phone, which apply a low-pass filter to achive the situation.	Unadjusted audio.	The volume is excessively loud, which change amplitude to achive the situation.

about the content and what they felt during the experience. The questions were:

- Please freely write down what you remember about the content.
- Please freely describe what you felt while experiencing the VR content.

The free-response session had no time limit, and subjects were permitted to write down as much information as they could recall. A 3-minute break followed the free-response session, during which subjects spent time with the HMD removed.

C. Experimental Conditions

The experimental conditions for visual and auditory information are as shown in Table II, with three variations prepared for each. For variations in visual information, three patterns were prepared: a simple plane with only a shrine, a simple plane with a shrine and shrine-related objects (trees, *torii* gate, *chozuya*), and a shrine located within a forest, accompanied by a *torii* gate and *chozuya*. For variations in auditory information, three patterns were prepared: a case with poor audio quality for the explanatory narration, a case with normal audio quality for the explanatory narration, and a case with high volume for the explanatory narration. Examples of scenes presenting each condition are shown in Figure 3.

TABLE III. Average fixation time for each subject across combinations of visual variety and auditory variety.

			visual (V)		
			R	N	L
auditory(A)	R	S_a	213.5	183.1	225.5
		S_b	159.1	124.7	172.3
		S_c	144.6	154.8	
	N	S_a	**210.0**	*198.5*	*190.2*
		S_b	**198.5**		*164.6*
		S_c	**190.2**	*138.1*	*101.1*
	L	S_a	**190.2**	*208.8*	**193.5**
		S_b	**208.8**	*173.0*	**177.6**
		S_c	**193.5**	*115.6*	**144.9**

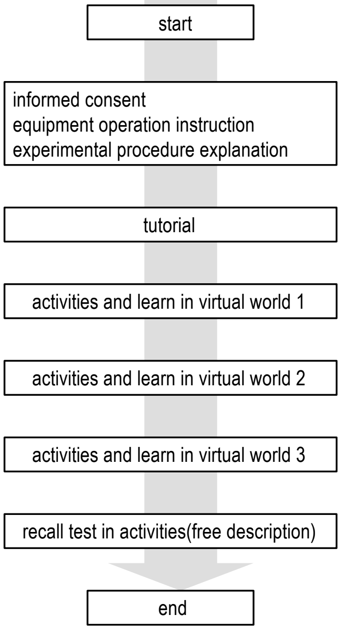

start
informed consent
equipment operation instruction
experimental procedure explanation
tutorial
activities and learn in virtual world 1
activities and learn in virtual world 2
activities and learn in virtual world 3
recall test in activities(free description)
end

Figure 4. Experimental design.

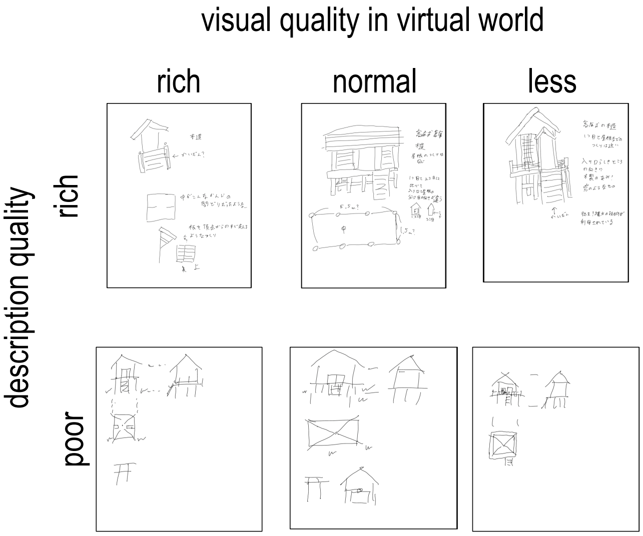


Figure 5. The result of free description in recall test.

D. Results (1) : Effect of Fixation Time and Visual/Auditory Quality

We focused on three participants $S_a,\ S_b,\ S_c$ whose recall test results were particularly distinctive as a case study to confirm the validity of this experiment. Among them, S_a and S_b both had almost no interest in architecture, while S_c was a graduate of an architecture department. The characteristics of their respective descriptions were as follows.

- S_a described information obtained within the virtual world simply but faithfully, as seen in the rich class of Figure 5,
- S_b described information obtained within the virtual world solely through deformed drawings (poor class of Figure 5), and
- S_c described information obtained within the virtual world using both simple text and detailed drawings.

Table III shows the average fixation time for each participant across visual and auditory condition combinations. The two blank entries indicate missing data due to malfunction of the auditory information presentation program. Although the values varied considerably across subjects, several trends were observed. When examining the (visual, auditory) conditions, fixation times were generally longer for $(R,\ N)$, $(R,\ L)$, and $(L,\ L)$. Furthermore, for $(N,\ R)$, fixation times tended to be shorter overall.

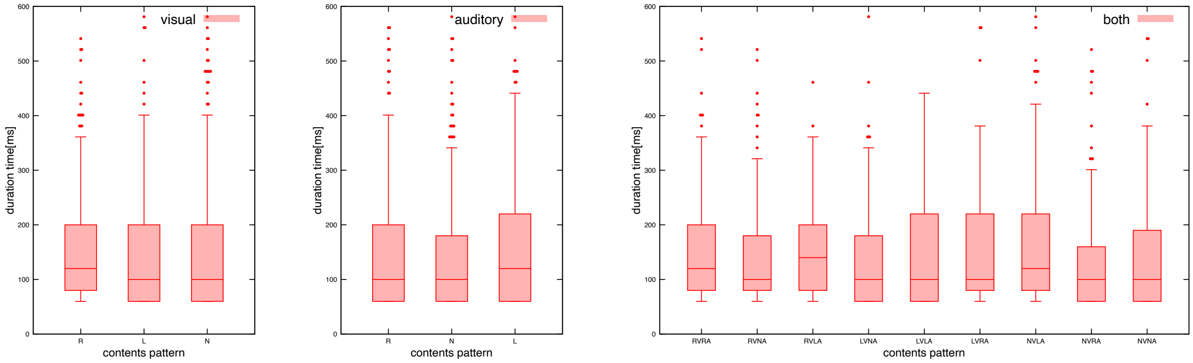
Figure 6 shows a boxplot of fixation time for combinations of manipulated perceptual information. In the figure, V denotes visual, A denotes auditory, and $L,\ N,\ R$ denote the quality of each perceptual information as low/normal/high. Figure 6(a) is a boxplot of fixation time for shots grouped by visual quality on the left and auditory quality on the right. Figure 6(b) is a boxplot of fixation time for shots segmented by visual quality-auditory quality combinations.

Figure 6 shows as follows: First, the distribution of fixation times for visual V was slightly longer for the R condition, with a median of approximately 120 ms; The distribution of fixation times for auditory A was slightly longer for the L condition, with a median of approximately 120 ms. These results suggest that fixation times are longer when the perceptual information condition is $RVLA$. Indeed, examining Figure 6(b) and confirming the median for $RVLA$ in Figure 6(a), it was approximately 150 ms, a value clearly larger than the medians for the V or A groups alone. Furthermore, both $RVRA$ and $NVLA$ were around 120 ms, larger than the median for the V or A single-stimulus groups. Considering the above, it is reasonable to conclude that for the perceptual conditions $RVLA$, $RVRA$, and $NVLA$ in the shot, information integration takes longer compared to a single perceptual condition.

To verify this, we conducted an analysis of variance (ANOVA) on the distribution of fixation times for the independent variables visual variety and auditory variety. First, a two-way ANOVA on the fixation times for three participants revealed a weak tendency toward an interaction effect ($F(2, 1201) = 2.258,\ p = 0.06$). A weak tendency was also observed for the main effect of auditory variety ($F(2, 1201) = 2.414,\ p = 0.09$). Next, focusing specifically on participant S_c who drew with particular precision in the recall test, a two-way ANOVA was performed using only S_c's data. No interaction was confirmed ($F(2, 402) = 1.65,\ p = 0.177$), a significant main effect was observed for auditory variety ($F(2, 402) = 4.081,\ p = 0.018$). Furthermore, multiple comparisons revealed a significant tendency for (visual variety, auditory variety) $= (N, R)$ and (L, N) $(p = 0.087)$.

E. Result (3): Distribution of Shot Duration for Visual/Auditory Variety

Next, we discuss the pupil diameter changes during auditory information listening for each of $S_a,\ S_b$, and S_c. Table IV



(a) For visual and auditory stimuli respectively, the left side shows visual stimuli grouped together, and the right side shows auditory stimuli grouped together.

(b) Combination of visual/auditory.

Figure 6. Boxplot of fixation time for combinations of manipulated perceptual information. In the figure, V denotes visual, A denotes auditory, and L, N, R denote the quality of each perceptual information as low/normal/high.

TABLE IV. Auditory Experimental Conditions Results and Measurements in Two-Second Time Period. The under number of each visual condition represents lightness for each condition which calculated by Matlab.

		Visual Condition		
		$L(166.98)$	$N(129.14)$	$R(122.99)$
Auditory Condition	L	0.196	0.341	0.172
	N	0.163	0.261	0.125
	R	0.172	0.291	0.148

shows the APD during auditory information listening for each visual variety.

The pupil diameter change was calculated as follows. First, it is necessary to determine the baseline r_b for the pupil diameter acquired simultaneously during gaze measurement, which is obtained for each gaze (∼20 [ms]). r_b was set as the average pupil diameter from 500 [ms] before the time t_{sa} of entering the actual space after the tutorial in the experiment until t_{sa}. Using the pupil diameter $r_p(t)$ measured at time t, $\bar{r_p}$ is calculated as follows.

$$\bar{r_p} = \frac{1}{n_{r_p}} \sum_{i}^{n_{r_p}} (r_p(t_i) - r_b),$$

where n_{r_p} represents the number of data observed $r(t)$ from t_{sa} to $t_{sa} + \Delta t$. We set Δt to 2000 [ms], which is considered sufficient for the response to the presented stimuli to settle.

In Table IV, when visual condition is N, $\bar{r_p}$ shows a larger value compared to the others. Particularly in $(N,\ L)$, $\bar{r_p}$ shows a large value. For $(N,\ L)$, the value is nearly twice of that of the other $\bar{r_p}$'s. Furthermore, comparing $\bar{r_p}$ in auditory condition, all variety of visual condition have large $\bar{r_p}$ when auditory condition L.

F. Result (2): Distribution of Shot Duration for Visual/Auditory Variety

Next, we perform a preliminary analysis of fixation behavior among participants. For participants' fixation behavior, we classified visual actions according to the conceptual diagram shown in Figure 7. Participants' eye movement behavior is broadly categorized into saccades and fixations. Regarding fixations, participants repeatedly make very short fixations to acquire information from the target object. In this process, the distribution of fixations that can occur can be categorized into the following three types:

- Remaining stationary on a specific point of a specific object for an extended period ((a) in Figure 7)
- Remaining stationary on the same object while shifting gaze to several parts of it ((b) in Figure 7)
- Repeating very short stationary periods and saccades, each time stationary on a different object ((c) in Figure 7)

We defined the three visual action classifications shown in Figure 7 as "1 shot", and examined the distribution of the time t_{shot} required for each shot to reveal the distribution of visual actions among participants. Table V shows the statistics for this. A clear difference in trend is that the visual behavior of S_c differs significantly from that of S_a and S_b. For S_c, the median t_{shot} was $(R,\ N,\ L) = (581,\ 621,\ 641)$ [ms], approximately half the time compared to S_a and S_b. Additionally, S_c exhibited very small values for other metrics such as $Q_1, Q_2,\ Q_3$, and IQR compared to the other groups.

IV. The Relationship Between the Variety of Perceptual Information Quality Combinations and Cognitive Load

Based on the above results, we consider the relationship between the combination of variety of perceptual information quality and the cognitive load experienced by participants.

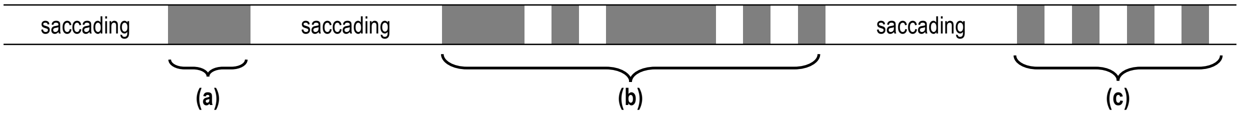


Figure 7. Setting fixation variety. (a) represents a state of continuously fixating on the same object, (b) represents a state of continuously fixating on different locations within the same object with saccades in between, and (c) represents a state of rapidly shifting gaze(short fixation) between different objects with saccades in between.

TABLE V. Visual variety-individual shot distribution statistics for each participant.

	S_a			S_b			S_c		
	R	N	L	R	N	L	R	N	L
Q_1	426.0	481.0	721.0	530.8	761.0	641.5	290.5	281.0	270.5
Q_2	891.5	761.0	1323.0	1282.5	1683.0	1302.0	**581.0**	**621.0**	**641.0**
Q_3	2083.3	1884.0	2124.0	2519.0	2424.0	4077.5	1202.0	1503.0	1542.5
average	1505.3	1443.4	1810.7	2335.6	2215.8	2336.2	1133.6	1087.9	1264.0
stdev	1578.0	1698.9	1631.6	2929.2	2208.3	2246.6	1863.3	1132.4	1472.5
min	60.0	161.0	40.0	100.0	201.0	160.0	80.0	80.0	20.0
max	8556.0	9056.0	8055.0	13425.0	10800.0	9598.0	15729.0	5631.0	7434.0

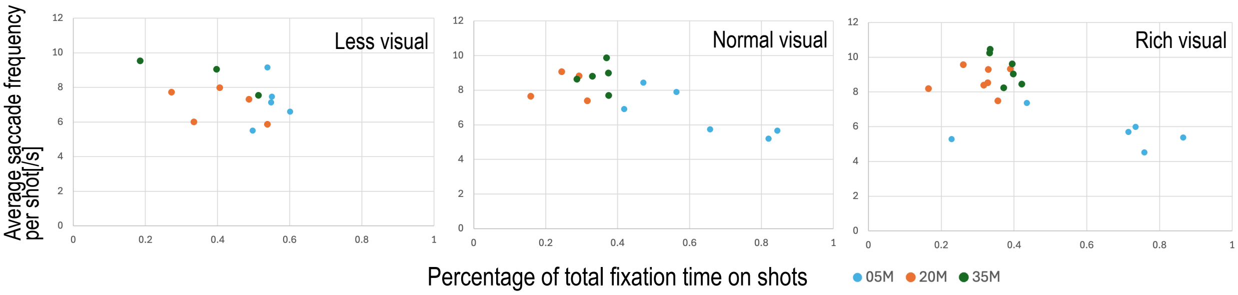


Figure 8. Relationship between the ratio of total fixation-saccadic time per memorized shot and the number of saccades per unit time.

TABLE VI. Summary of experimental results.

			visual condition		
			$L(++)$	$N(-)$	$R(-)$
Auditory Condition	L	Fixation time	N/A	N/A	N/A
		ANOVA	N/A	N/A	N/A
		$\bar{r_p}$	$+$	$++$	$+$
	N	Fixation time	$++$	N/A	$--$
		ANOVA	$*$	N/A	N/A
		$\bar{r_p}$	$+$	$+$	$--$
	R	Fixation time	$++$	$--$	$++$
		ANOVA	N/A	$*$	N/A
		$\bar{r_p}$	$+$	$+$	$-$

A. *Relationship between Average Fixation Time, Information Integration, and Cognitive Load*

Characteristics of fixation time distributions across different perceptual information qualities revealed a fundamental interaction effect dominated by auditory information. Furthermore, fixation time statistics showed that combining visual and auditory qualities tend to result in longer fixation durations compared to single information quality varieties, relative to the average fixation time for either visual or auditory quality alone. This phenomenon is explained based on Figure 1 as follows. In multimodal information processing, as depicted in Figure 1, integrating $I^{sym}(t)$ at the *degree of matching* requires searching for objects within long-term memory that contain multiple perceptual information types. In the example from Section III, focusing on the quality of the input information (visual and auditory), Table VI shows various results for the combination of variety.

The main effect was observed for auditory information, so we will examine each auditory information category.

First, for the auditory information condition L, there was no significant trend in fixation time, while $\bar{r_p}$ showed a tendency toward dilation. This suggests complex information processing is occurring due to the auditory information. Specifically, $(N,\ L)$ exhibited significant pupil dilation. If this auditory condition is appropriate for the listener, it indicates cognitive load arising from language processing.

Next, for auditory information with condition N, the tendency differs depending on the visual condition. For $(L,\ N)$, pupil dilation occurred, while for $(R,\ N)$, pupil constriction occurred. Additionally, for $(L,\ N)$, fixation time was longer, while for $(R,\ N)$, fixation time was shorter. From them, it suggests that pupil response and fixation are linked. Therefore, depending on the quality of visual information, the following behavioral differences are expected:

- When quality is low, auditory information is obtained in a state where nothing else is visible. To confirm which part the explanation refers to, longer fixation times occur.

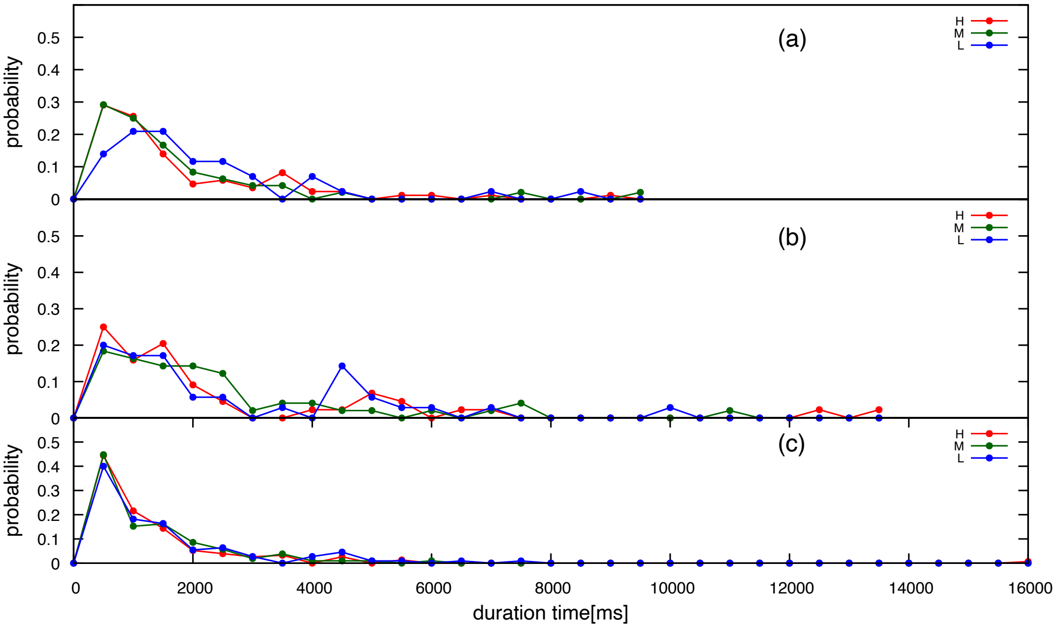


Figure 9. Distribution of fixation activity time per subject. (a) and (b) represent subjects with no interest in architecture $(S_a,\ S_b)$, (c) represents subjects with interest in architecture (S_c).

- Under the conditions of this study, excessively high quality is approximately equivalent to having too much visual information (objects). That is, obtaining auditory information amidst numerous visual targets leads to unstable visual point, resulting in shorter fixation times. This cognitive overload may explain the smaller $\bar{r_p}$ values.

Therefore, under the $(N,\ N)$ condition, where a moderate cognitive load is applied, the results suggest that the value of $\bar{r_p}$ may show a slight upward trend.

Next, the condition where auditory information was R also had a very large effect on fixation time. Condition pair $(N,\ R)$ had very short fixation times, while the others had long ones. Furthermore, the pupil response in condition pairs $(L,\ R)$ and $(N,\ R)$ was somewhat large. This suggests that when visual information is excessively scarce or abundant, longer fixations may be maintained to integrate it with auditory information. In our case, the origin of complexity is considered to stem from the integration of auditory and visual information. On the other hand, for condition pair $(R,\ R)$, where there is much to integrate, it is suggested that subjects may abandon memory of what they saw and heard in the VR space, primarily as a result of receiving excessive stimuli from visual/auditory information.

B. *Individual Differences in Eye-Movement Behavior*

Due to individual differences in human behavior, we focus on eye-movement behavior within the virtual world to perform individual-level analysis. Figure 8 shows the total fixation time - total saccadic time ratio and the average saccade frequency per second for shots recalled in the recall test. Although individual differences exist, a general trend shows a slight increase in saccade frequency as visual information moves from left to right. From the figure, the fixation-to-saccade ratio per shot was generally around 0.4, and the saccade frequency averaged approximately 8 to 10 saccades per second. We consider this trend to show no significant variation.

Figure 9 shows the shot time distribution for each participant. The number of shots for S_a was $(L,\ N,\ R) = (87, 50, 44)$, for S_b it was $(45,\ 50,\ 36)$, and for S_c it was $(153,\ 106,\ 111)$. Looking at the distributions for S_a and S_b, while there are differences in peak locations and visual conditions, they show distinct distributions for each condition. Generally, the peak is around 500ms, but the shot distribution extends relatively far up to about 1500ms. Additionally, there is a second peak around 4000ms. This trend is particularly pronounced when the visual condition is L.

In contrast, S_c consistently changed shot scenes at similar time intervals regardless of visual condition, showing no variation based on visual condition. On the other hand, during the recall test, S_c provided quite detailed descriptions regarding drawing but used few verbal expressions.

This suggests that the way information is acquired in the virtual world is significantly influenced by the participant's timing for selecting specific scenes—that is, the shots. Participants like S_c, who possess an interest in architecture and are skilled at information acquisition, extract shots at consistent intervals regardless of visual appearance. They acquire a large amount of information in relatively short, fragmented intervals, enabling the extraction of detailed features. In contrast, participants like S_a and S_b, who lack interest in architecture and are beginners in information acquisition, likely attempt to gather as much information as possible in a single shot,

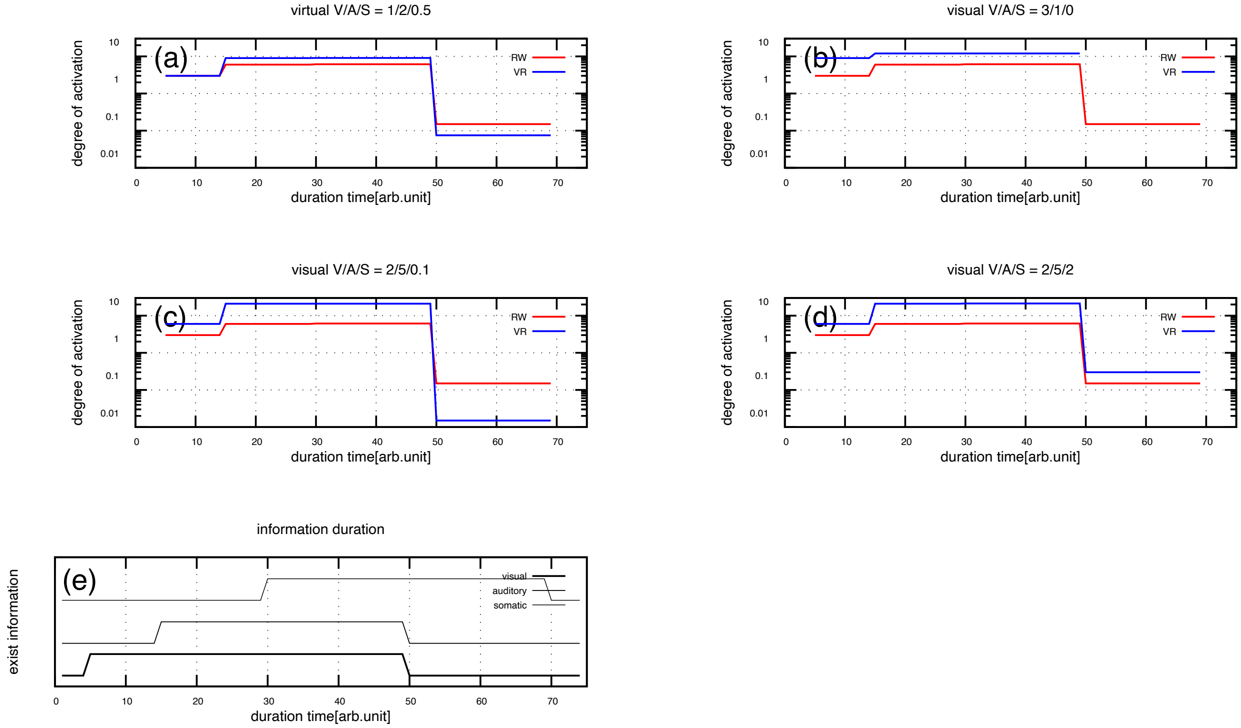


Figure 10. The trends of estimated $I^{syn}(t)$ which are changed three perceptual information(visual, auditory, somatic) amplified in Virtual Reality space.

resulting in a more variable shot time distribution.

Specifically, focusing on S_c and analyzing the details, when the objects of interest were A, B, C, a tendency was frequently observed where the fixation point would shift from the object of interest to another object—such as $A - B$, $A - C$, or $B - C$—before returning to the original location. This suggests a connection to the findings of Kurihara et al. [10], who demonstrated that temporarily shifting the fixation point away and then returning it to the same location enhances memory consolidation.

C. Perception in the Real/Virtual World

So far, we discussed that the impact of combining multimodal perceptual information in virtual spaces on cognitive load. Finally, we will mention what can be expected regarding the relationship between perception and cognition in real and virtual spaces. Figure 10 shows the trend of $I^{syn}(t)$ when the degrees to which visual, auditory, and somatic information are emphasized in VR are varied. The solid red line in the figure shows $I^{syn}(t)$ when visual, auditory, and somatic information are received in the real world. Here, we set $j = 1, 2$. Both visual and auditory information equal 1 for one, and 2 for the other. The somatic information is set to 0.5 on one side and 0.3 on the other. The solid blue lines indicate the degree to which the same information is distorted in VR.

Figure 10 (e) shows the duration of information obtained from each sense. In contrast, Figures 10 (a)∼(d) show the degree of integrated information activation calculated by Equation (1). Figure 10 (a) shows the case where auditory is multiplied by a factor of 2 and somatic by a factor of 0.5. For $t < 50$, the VR space is slightly more chunk activated, but the characteristics are almost same. However, at $t \geq 50$, when only somatic information is perceived, the chunk activation in the VR space is lower. In Figure 10 (b), the visual information is markedly increased, while the somatic information is not reproduced in the VR space. For $t < 50$, the activation of chunk in the VR space is markedly increased, but at $t \geq 50$, the somatic information is lost; Hence, there is no chunk activation in the VR space. In Figure 10 (c), the somatic information is lowered to 0.1 and the information is emphasized in the form of visual<auditory. In particular, at $t \geq 50$, the somatic information is still present, but its effect is much smaller. Figure 10 (d) is the case where the somatic information is also doubled. Compared with Figures 10 (b) and (c), chunk activation remains high at $t \geq 50$.

The intensity of human sensation is expressed as a logarithm according to Weber-Fechner's Law. Therefore, as shown in Figure 10, even if the difference in sensory information is very slight, it suggests that the human senses can distinguish this difference. The sense of "slightly different from the real world" felt in VR content is thought to be caused by such slight differences in sensory information. The sensory information obtained in real space is not necessarily large, as shown in the example in Section III. However, it is easy to understand that these small differences lead to a sense of discomfort, which in turn indicates a decrease in immersive perception.

In the present case, we only dealt with a very simple integration of information. To advance our understanding of human sensory perception and use knowledge in VR spaces, scholars should develop a new approach that uses operators in Equation (1), such as Adaptive Control of Thought—Rational (ACT-R) [11] and Model Human Processor with Realtime Constraints (MHP/RT) [11] which incorporate Two Minds, to integrate information in a cognitive architecture [12][13][14].

V. Conclusion and Future Work

To realize adaptive VR, we need to design deeper immersion resulting from human interaction with real/VR spaces. As a first step, this study describes a sensory-cognitive model for VR spaces. Based on the described model, we analyzed how information is acquired in a virtual world, focusing on visual and auditory information, and how behavior changes when conditions are altered, using results of recall test. The results suggested that some malfunction occurs under conditions other than $(N,\ N)$. Furthermore, it suggested that the degree of proficiency in acquiring information from space may influence eye-movement behavior and, consequently, the state of memory. Connecting the two issues, multimodal information and chunk activatin, we undertake the research qualitatively and explain the phenomenon that can occur when one or more types of information (visual, auditory, or somatic) is overemphasized or surpressed in a VR space. Expressing human sensory intensity as a logarithm according to Weber-Fechner's Law, we suggest that human senses can distinguish differences in sensory information, even if the differences are very slight. Considering these points, we are able to deepen our understanding of how the VR space realizes the immersive effect with impressive each other. Moreover, we are able to design "adaptive" immersive contents. In the future, it is necessary to investigate in experiments whether the degree of immersion felt by users changes when they experience VR content by changing the degree of emphasis of each sensory information. The metrics used to judge the degree of similarity between the real and virtual worlds can be defined as the overlap between the information held in the WM and the information in the LTM that has been activated up to that point in time. As the activation of information in the LTM is considered to be reflected in biological information, future experiments could be conducted using eye gaze and skin resistance measurements and subjective evaluation by means of questionnaires. Hysteresis can be considered based on the impact of inputs from the environment on the memory of the time series.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 20H04290, 22K12284, 23K11334, and National University Management Reform Promotion Project. The authors would like to thank Editage (www.editage.com) for the English language editing.

References

[1] T. Nakagawa, M. Kitajima, and K. T. Nakahira, "Model-Based Analysis of the Differences in Sensory Perception between Real and Virtual Space : Toward "Adaptive Virtual Reality"," in AIVR 2024 : The First International Conference on Artificial Intelligence and Immersive Virtual Reality, 2024, pp. 39–44.

[2] C. Baker and S. H. Fairclough, "Chapter 9 - adaptive virtual reality," in Current Research in Neuroadaptive Technology, S. H. Fairclough and T. O. Zander, Eds. Academic Press, 2022, pp. 159–176. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128214138000142

[3] J. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, vol. 39, 12 1980, pp. 1161–1178.

[4] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." Journal of personality and social psychology, vol. 76 5, 1999, pp. 805–19. [Online]. Available: https://api.semanticscholar.org/CorpusID:14362153

[5] J. Bao, X. Tao, and Y. Zhou, "An emotion recognition method based on eye movement and audiovisual features in mooc learning environment," IEEE Transactions on Computational Social Systems, vol. 11, no. 1, 2024, pp. 171–183.

[6] N. Sun and Y. Jiang, "Eye movements and user emotional experience: a study in interface design," Frontiers in Psychology, vol. Volume 16 - 2025, 2025. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2025.1455177

[7] A. Steed, Y. Pan, F. Zisch, and W. Steptoe, "The impact of a self-avatar on cognitive load in immersive virtual reality," in 2016 IEEE Virtual Reality (VR), 2016, pp. 67–76.

[8] S. M. H. Mousavi et al., "Emotion recognition in adaptive virtual reality settings: Challenges and opportunities," CEUR Workshop Proceedings, vol. 3517, jan 2023, pp. 1–20. [Online]. Available: https://sites.google.com/view/wamwb/

[9] T. Batra and P. Chunarkar-Patil, "Virtual reality in bioinformatics," Open Access Journal of Science, vol. 3, no. 2, 2019, pp. 63–70.

[10] Y. Kurihara, M. Shino, K. Nakahira, and M. Kitajima, "Visual behavior based on information foraging theory toward designing of auditory information," in Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 1: HUCAPP, INSTICC. SciTePress, 2024, pp. 530–537.

[11] F. E. Ritter, F. Tehranchi, and J. D. Oury, "ACT-R: A cognitive architecture for modeling cognition," WIREs Cognitive Science, vol. 10, no. 3, 2019, p. e1488. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1488

[12] M. Kitajima, Memory and Action Selection in Human-Machine Interaction. Wiley-ISTE, 2016.

[13] M. Kitajima and M. Toyota, "Simulating navigation behaviour based on the architecture model Model Human Processor with Real-Time Constraints (MHP/RT)," Behaviour & Information Technology, vol. 31, no. 1, 2012, pp. 41–58.

[14] M. Kitajima and M. Toyota, "Decision-making and action selection in Two Minds: An analysis based on Model Human Processor with Realtime Constraints (MHP/RT)," Biologically Inspired Cognitive Architectures, vol. 5, 2013, pp. 82–93.

Characterising Emotion Shifts Using Markov Processes

1st Clement Leung
School of Science and Engineering
Chinese University of Hong Kong
Shenzhen, China
clementleung@cuhk.edu.cn

2nd Zhifei Xu
School of Science and Engineering
Chinese University of Hong Kong
Shenzhen, China
zhifeixu1@link.cuhk.edu.cn

***Abstract*—In many operational contexts, particularly those that are safety-critical, it is imperative that human participants maintain appropriate emotional conditions. Consequently, the accurate recognition of these states is a central challenge in modern research. While mainstream methods have utilized Pretrained Language Models (PLMs) for emotional understanding, the emergence of Large Language Models (LLMs) like ChatGPT offers new possibilities. This study investigates the underexplored zero-shot capabilities of ChatGPT-4 for image-based emotion analysis. We focus on its performance in classifying emotional valence (positive vs. negative) and predicting its temporal evolution. Our findings demonstrate that ChatGPT-4 can effectively forecast changes in emotional states, surpassing expectations. Nonetheless, we note deficiencies in its ability to accurately discern specific negative emotions, highlighting a need for further refinement. The study further introduces a hierarchical stochastic model to formalize these emotional shifts, providing a theoretical bridge between empirical LLM outputs and psychological stability parameters.**

Keywords-image emotion prediction; large language model; ChatGPT4; zero-shot; markov chain; emotion stability parameter.

I. Introduction

Accurately interpreting human emotion is fundamental to communication, enabling connection while revealing underlying mental states and intentions. For this reason, research has increasingly focused on integrating emotional insight into AI, from early human-computer dialogue systems [1][2] to the advanced Large Language Models (LLMs) of today. The arrival of models like ChatGPT [3] and Instruct-GPT [4] has sparked immense interest in LLM-based emotion recognition, particularly for providing emotional support in personal, clinical, and customer service settings. This study evaluates how effectively the latest iteration, ChatGPT-4 [5], can infer emotions from facial expressions alone.

The need for reliable emotion recognition is not merely academic; it is critical for safety, mental health, and user experience [6, 7]. Social stressors such as occupational strain, perceived injustice, and relationship loss can precipitate significant harm [8, 9]. Tragic incidents, including suicidal ideation linked to work demands [8], school shootings, road rage, and even a depressed pilot's attempt to shut down engines mid-flight [9], underscore the urgent need for better technological aids. Advanced emotion recognition and prediction systems could offer critical support for safety and mental health interventions [10].

While neural networks have long enabled emotionally responsive generation [11], the nuanced linguistic competence of modern LLMs like ChatGPT-4 has transformed conversational AI. Yet, the extent to which these systems can track or express emotion, especially through non-textual data, remains underexplored. This research assesses the strengths and limitations of ChatGPT-4 in multimodal emotion recognition and prediction [12, 13, 14]. By leveraging its capabilities, we can also reduce the human-rater bias often present in psychological studies, thereby promoting fairness and ethically tailored interventions.

A. Related Work: From Static to Generative Approaches

Historically, emotion recognition has relied on static classification models, such as Convolutional Neural Networks (CNNs) trained on fixed datasets like FER-2013 or AffectNet [15]. These "discriminative" models are excellent at categorizing a single frame but often fail to capture the temporal fluidity of human emotion. They view emotion as a snapshot rather than a process.

In contrast, Generative AI and LLMs offer a "generative" approach. They can synthesize context, history, and multimodal cues (text + image) to infer not just the current state, but the likely future state. However, the stochastic nature of LLMs introduces variability. This necessitates a robust mathematical framework to model that variability. Our work bridges this gap by applying stochastic process theory—specifically Markovian dynamics—to the output of generative models, providing a rigorous structure to the fluid predictions of an LLM.

Our work is grounded in established theories of emotion. These include categorical models, such as Ekman's six universal emotions (joy, sadness, fear, anger, surprise, disgust) [16] and Plutchik's wheel of eight (joy, trust, fear, surprise, sadness, disgust, anger, anticipation) [17], which posit a fixed set of basic emotions. In contrast, dimensional models view emotions along continuous axes of valence (positive/negative), arousal (intensity), and dominance [18, 19].

B. Contribution and Relation to Prior Work

This manuscript represents a substantial extension of our preliminary study presented at the BRAININFO 2025 conference [1]. While our initial work established the baseline feasibility of using ChatGPT-4 for zero-shot emotion prediction under hypothetical situations, the current study significantly expands the theoretical framework, experimental scope, and comparative analysis. The specific contributions that distinguish this article from the conference version are as follows:

1) **Hierarchical Stochastic Modeling:** We upgrade the mathematical framework from a standard Markov chain to a hierarchical model. This includes the introduction of a binary valence layer based on a Poisson process (Section II-B), which mathematically links global emotional volatility to categorical transitions.
2) **Multimodal Dataset Expansion:** Whereas [1] relied exclusively on static facial expression datasets, this study incorporates the Multimodal EmotionLines Dataset (MELD). This allows us to evaluate the model's performance on complex scenarios involving dialogue and sentiment-tagged utterances.
3) **New Experimental Tasks:** We introduce a new prediction task involving emotion-conditioned sentences. Unlike the situational prompts used in [1], this task tests the model's ability to predict emotional evolution based on specific verbal cues (e.g., an angry utterance vs. a surprised utterance).
4) **Comparative Analysis:** We provide a comprehensive comparison between ChatGPT-4 and the Doubao (TikTok) Large Language Model, highlighting critical divergences in how these models interpret negative emotional states and zero-shot multimodal prompts.

Section II introduces the hierarchical stochastic model used to formalise emotion shifts. Section III describes the datasets, prompting protocol, and quantitative evaluation results. Section IV discusses limitations, ethical considerations, and future directions.

C. Problem Setting and Research Questions

We study *zero-shot* emotion inference where the model receives (i) a facial image and (ii) an optional textual continuation (a scenario description or an emotion-conditioned utterance), and must output both a current emotion label and a plausible next emotion label. This differs from standard facial-expression classification in two ways. First, the output is inherently *temporal* (a transition rather than a single label). Second, the "ground truth" for a hypothetical future emotion is not directly observable; therefore, our evaluation separates (a) *recognition correctness* (agreement with dataset labels for the current frame) from (b) *transition consistency* under controlled polarity cues (positive vs. negative situations) and under utterances drawn from MELD-style emotion categories.

Accordingly, we structure the study around three research questions:

- **RQ1 (Recognition):** When prompted with facial images only, how reliably can ChatGPT-4 infer the dataset emotion label, and how does performance differ between positive vs. negative categories?
- **RQ2 (Shift prediction):** Given an initial facial emotion, does the model predict transitions that are *consistent* with the polarity of the subsequent situation/utterance (e.g., reward-like vs. breakup-like contexts), and where does it fail?
- **RQ3 (Mechanism):** Can a compact stochastic process model (Poisson + Markov + persistence) explain the empirical pattern that valence is often correct while fine-grained negative categories are frequently confused?

These questions motivate our hierarchical model in Section II and the prompting/evaluation protocol in Section III.

II. Mathematical Model

This section formalizes the stochastic model that we use to describe the temporal evolution of emotions and to interpret the empirical behaviour of ChatGPT-4 and Doubao in Section III. The construction proceeds in three layers: (i) a binary valence layer based on a Poisson process, (ii) a categorical layer using an eight-state Markov chain, and (iii) a stability layer with emotion-specific persistence parameters.

A. Notation

Table I summarises the main notation used in this section.

B. Binary valence model (Poisson switching)

At the coarsest level, we distinguish positive from negative valence. Let

$$S(t) \in \{+1, -1\} \tag{1}$$

denote the valence state at continuous time t, with $S(0) = +1$ indicating an initially positive state.

Valence switches are driven by a homogeneous Poisson process $N(t)$ with rate $\lambda > 0$. Each arrival of the process flips the sign of $S(t)$. If the number of arrivals in $(0, t]$ is even, the valence remains positive; if it is odd, the valence is negative.

Let $p_k = \Pr\{N(t) = k\}$ be the Poisson probabilities with parameter λt. The probability that valence is still positive at time t, given that it started positive, is

$$\Pr\{S(t) = 1 \mid S(0) = 1\} = p_0+p_2+p_4+\cdots = e^{-\lambda t}\cosh(\lambda t). \tag{2}$$

Similarly, the probability that the state has flipped to negative is

$$\Pr\{S(t) = -1 \mid S(0) = 1\} = e^{-\lambda t}\sinh(\lambda t). \tag{3}$$

The parameter λ therefore acts as a global emotional volatility parameter: small λ implies long-lasting valence (rare switches), whereas large λ produces rapid alternation between positive and negative states.

1) Discrete-step interpretation and an explicit stay/flip form: In many applications the model is queried at discrete steps (e.g., turns in a dialogue or time bins of a fixed duration Δt). Under Poisson-driven sign flips, the probability of *staying* in the same valence over one step is

$$\Pr\{S(t+\Delta t) = S(t)\} = e^{-\lambda\Delta t}\cosh(\lambda\Delta t) = \frac{1+e^{-2\lambda\Delta t}}{2}, \tag{4}$$

and the probability of a *flip* is

$$\Pr\{S(t+\Delta t) \neq S(t)\} = e^{-\lambda\Delta t}\sinh(\lambda\Delta t) = \frac{1-e^{-2\lambda\Delta t}}{2}. \tag{5}$$

TABLE I
MAIN NOTATION USED IN THE MODEL.

Symbol	Description
$S(t)$	Valence state at continuous time t ($+1$ = positive, -1 = negative)
$N(t)$	Poisson process counting valence switches up to time t
λ	Global valence switching rate (Poisson intensity)
E_t	Categorical emotion at discrete step t
E	Emotion set $\{$Joy, Trust, Surprise, Anticipation, Sadness, Disgust, Anger, Fear$\}$
E_+	Positive emotions $\{$Joy, Trust, Surprise, Anticipation$\}$
E_-	Negative emotions $\{$Sadness, Disgust, Anger, Fear$\}$
$p_i(t)$	Probability $\Pr\{E_t = E_i\}$ of emotion E_i at step t
$\mathbf{p}(t)$	Column vector $[p_1(t), \dots, p_8(t)]^\top$
$\tilde{p}_i(t)$	Stability-adjusted probability of emotion E_i at step t
λ_i	Stability parameter for emotion E_i (smaller = more persistent)
P_{ij}	One-step transition probability from E_i to E_j
P	8×8 row-stochastic state transition matrix

These closed forms clarify how λ controls volatility: for small $\lambda\Delta t$, flips are rare; as $\lambda\Delta t$ grows, the process approaches a near-random alternation with stay probability $\approx 1/2$.

Moreover, if an empirical estimate $\widehat{p}_{\text{stay}}$ of the valence stay probability over Δt is available, one may invert (4) to obtain

$$\widehat{\lambda} = -\frac{1}{2\Delta t}\ln\left(2\widehat{p}_{\text{stay}} - 1\right), \quad \text{valid when } \widehat{p}_{\text{stay}} > \tfrac{1}{2}. \tag{6}$$

This provides a principled link between observed stability (from repeated LLM trajectories) and the volatility parameter.

C. Categorical extension: eight-emotion Markov chain

To represent which emotion is being expressed, we refine the valence layer into eight categorical states,

$$E = \{\text{Joy}, \text{Trust}, \text{Surprise}, \text{Anticipation}, \tag{7}$$

$$\text{Sadness}, \text{Disgust}, \text{Anger}, \text{Fear}\}. \tag{8}$$

We partition these into positive and negative subsets,

$$E_+ = \{\text{Joy}, \text{Trust}, \text{Surprise}, \text{Anticipation}\} \tag{9}$$

$$E_- = \{\text{Sadness}, \text{Disgust}, \text{Anger}, \text{Fear}\} \tag{10}$$

and define a simple valence map $g : E \to \{+1, -1\}$ with $g(E_i) = +1$ for $E_i \in E_+$ and $g(E_i) = -1$ for $E_i \in E_-$.

Time is now indexed in discrete steps $t \in \{0, 1, 2, \dots\}$ (e.g., conversational turns or fixed-size time bins). Let E_t denote the emotion at step t, and define

$$p_i(t) = \Pr\{E_t = E_i\}, \qquad \mathbf{p}(t) = [p_1(t), \dots, p_8(t)]^\top, \tag{11}$$

with $\sum_{i=1}^{8} p_i(t) = 1$.

The categorical dynamics follow an eight-state Markov chain with transition matrix P:

$$P_{ij} = \Pr\{E_{t+1} = E_j \mid E_t = E_i\}, \qquad \sum_{j=1}^{8} P_{ij} = 1 \quad \forall i. \tag{12}$$

Using the column-vector convention, the one-step update is

$$\mathbf{p}(t) = P^\top \mathbf{p}(t-1). \tag{13}$$

1) Theoretical Implications: This hierarchical structure implies that emotional stability is not uniform. The Poisson layer dictates the "mood" (valence), while the Markov layer dictates the specific "affect" (emotion). This aligns with psychological appraisal theories where a general valence check often precedes specific emotional labeling. In our experiments with ChatGPT-4, we observe that the model often gets the valence correct (Poisson layer) even when it confuses the specific category (Markov layer), supporting the validity of this hierarchical separation.

D. Stability and persistence parameters

To keep the model simple and interpretable, we group emotions by polarity and assign

$$\lambda_i = \begin{cases} 0.2, & E_i \in E_+ \quad \text{(more persistent positive emotions)}, \\ 0.5, & E_i \in E_- \quad \text{(more volatile negative emotions)}. \end{cases} \tag{14}$$

Given a current distribution $\mathbf{p}(t) = [p_1(t), \dots, p_8(t)]^\top$, the probability that emotion E_i stays the same at time t is modelled analogously to (2):

$$P_{\text{stay},i}(t) = p_i(t)\, e^{-\lambda_i t} \cosh(\lambda_i t). \tag{15}$$

The complementary probability mass $p_i(t) - P_{\text{stay},i}(t)$ corresponds to transitions out of E_i.

We then redistribute this transition mass according to the matrix P. Let P_{ji} be the probability of moving from E_j to E_i. The stability-adjusted probability of emotion E_i at time t is

$$\tilde{p}_i(t) = P_{\text{stay},i}(t) + \sum_{j \neq i} \left[p_j(t) - P_{\text{stay},j}(t)\right] P_{ji}. \tag{16}$$

E. Constructing P from empirical transitions

The Markov transition matrix P can be interpreted in two complementary ways. First, it can be treated as a *theoretical prior* encoding psychologically plausible shifts (e.g., Surprise → Joy under positive contexts). Second, it can be estimated from model-generated trajectories to summarise how a particular LLM tends to "move" between emotion labels.

Concretely, suppose we collect C_{ij} counts of predicted one-step transitions $E_t = E_i \to E_{t+1} = E_j$ across all prompts and samples. A maximum-likelihood estimate is obtained by row-normalising:

$$\widehat{P}_{ij} = \frac{C_{ij}}{\sum_{k=1}^{8} C_{ik}}. \tag{17}$$

To avoid zero-probability artifacts (common when some transitions are rarely observed), a simple additive smoothing scheme can be used:

$$\widehat{P}_{ij}^{(\alpha)} = \frac{C_{ij} + \alpha}{\sum_{k=1}^{8} (C_{ik} + \alpha)}, \tag{18}$$

where $\alpha > 0$ acts like a symmetric Dirichlet prior and guarantees a well-defined stochastic matrix. In Section III, we primarily use P to (i) generate reference trajectories via Algorithm 1 and (ii) interpret confusion patterns: large off-diagonal mass from a negative emotion into Neutral/Joy-like predictions is consistent with low specificity and reduced AUC for that category.

F. Numerical Simulation Algorithm

To visualize the prediction process, we formalize the simulation steps in Algorithm 1. This algorithm iteratively updates the emotion state vector based on the Markov transition matrix and stability adjustments defined above.

Algorithm 1: Emotion Evolution Simulation

```
Input: Initial state vector $\mathbf{p}(0)$, Transition Matrix $P$,
  Stability parameters $\lambda_i$, Time horizon $T$.
Output: Probability distributions $\tilde{\mathbf{p}}(t)$ for $t = 1 \dots T$.
for $t = 1$ to $T$ do
    // Step 1: Standard Markov Update
    $\mathbf{p}(t) \leftarrow P^{\top}\mathbf{p}(t-1)$;
    // Step 2: Calculate Persistence
    for $i = 1$ to $8$ do
        $P_{\text{stay},i}(t) \leftarrow p_i(t) e^{-\lambda_i t} \cosh(\lambda_i t)$;
    end
    // Step 3: Redistribute Mass
    for $i = 1$ to $8$ do
        $\tilde{p}_i(t) \leftarrow P_{\text{stay},i}(t) + \sum_{j \neq i} [p_j(t) - P_{\text{stay},j}(t)] P_{ji}$;
    end
    // Step 4: Normalize and Store
    $\tilde{\mathbf{p}}(t) \leftarrow \text{Norm}(\tilde{\mathbf{p}}(t))$;
end
return $\tilde{\mathbf{p}}(1 \dots T)$
```

This algorithmic approach ensures that for any given initial emotion detected by the LLM, we can project a probabilistic trajectory of how that emotion might decay or shift, providing a benchmark to compare against the LLM's own predictions.

G. Connection to ROC/AUC metrics and LLM experiments

The model above provides a conceptual bridge between emotional stability and the classification metrics observed in Section III. At the valence level, a larger global λ or larger negative-emotion λ_i produces more frequent sign flips and greater overlap between positive and negative trajectories. In classical detection theory, increased overlap translates into lower AUC: the ROC curve moves closer to the diagonal.

Empirically, we observe that positive emotions (e.g., happiness, surprise) achieve high accuracies and AUC values close to 1, indicating stable, well-separated positive trajectories. Negative emotions, especially disgust, exhibit lower accuracies and smaller AUC, suggesting that their score distributions overlap more with positive classes. This pattern is precisely what the model predicts when negative emotions have larger λ_i (more volatile, shorter dwell times).

III. Experimental Design and Results

Understanding and predicting emotion is a major frontier in conversational AI. By analyzing not just the words people use, but also visual and auditory cues, we can forecast how their feelings will shift throughout a dialogue.

A. Evaluation Metrics

To rigorously assess the model's performance, we utilize standard classification metrics derived from the confusion matrix. Let TP be True Positives, TN be True Negatives, FP be False Positives, and FN be False Negatives.

- **Accuracy:** The proportion of total predictions that are correct.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

- **Sensitivity (Recall):** The ability of the model to correctly identify positive emotional states.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{20}$$

- **Specificity:** The ability of the model to correctly identify negative emotional states.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{21}$$

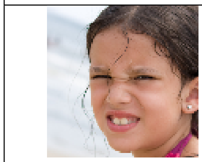
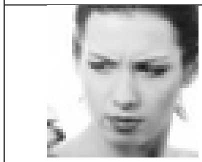
Additionally, we calculate the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which plots Sensitivity against $1 -$ Specificity. An AUC of 0.5 represents random guessing, while 1.0 represents perfect classification.

1) Uncertainty reporting: Point estimates can hide variability across samples and prompts. Where space permits, we recommend reporting uncertainty via nonparametric bootstrap confidence intervals. Specifically, we resample the evaluation set with replacement, recompute Accuracy and AUC for each resample, and report the 2.5/97.5 percentiles as a 95% interval. This is particularly important when comparing models (ChatGPT-4 vs. Doubao) where differences may be concentrated in a small subset of hard negative categories.

B. Emotion Recognition with different situations

For the experimental part, we chose three Data sets from Kaggle which are Emotion Detection, Facial Expressions Training Data, and Natural Human Face Images for Emotion Recognition.

TABLE II
SAMPLE OF FOUR DIFFERENT SITUATIONS

Dataset	Question 1	Question 2	Question 3	Question 4
	What is the emotion of this person? If they are about to be praised by their boss or their parents respectively, what do you think their emotions become?	If they were to be criticized, what do you think their emotions would be?	If they were to receive a $1,000 reward, what do you think their emotions would be?	If they were to break up, what do you think their emotions would be?
	What is the emotion of this person? If they are about to be praised by their boss or their parents respectively, what do you think their emotions become?	If they were to be criticized, what do you think their emotions would be?	If they were to receive a $1,000 reward, what do you think their emotions would be?	If they were to break up, what do you think their emotions would be?
	What is the emotion of this person? If they are about to be praised by their boss or their parents respectively, what do you think their emotions become?	If they were to be criticized, what do you think their emotions would be?	If they were to receive a $1,000 reward, what do you think their emotions would be?	If they were to break up, what do you think their emotions would be?

1) Label harmonisation across datasets and the eight-state model: Different datasets use partially overlapping taxonomies. For consistent reporting, we focus on the shared labels {anger, disgust, happiness, neutral, sadness, surprise} for the six-way experiments. Our stochastic model uses an eight-state affect set inspired by Plutchik; the mapping is summarised in Table III. Neutral is treated as a separate category in evaluation (not one of the eight affect states), which is a common practical compromise when combining categorical theories with "no strong affect" dataset labels.

2) Datasets: **Emotion Dection** This dataset is the same as the FER-2013 [20] dataset. The collection features 35,685 grayscale images, each 48x48 pixels. The images have been categorized by the creators into several emotions, namely anger, disgust, fear, happiness, neutrality, sadness, and surprise.

Facial Expression Training Data The AffectNet [21] database, a substantial compilation of facial images annotated with expressions, serves as the foundation for this dataset. To adapt to typical memory constraints, image resolution is scaled down to 96x96 pixels.

Natural Human Face Images for Emotion Recognition This unique dataset is curated from the Internet, encompassing more than 5,500 images manually labeled for eight emotional expressions. Each image captures real human expressions in grayscale format of 224x224 pixels.

3) Task Definition of Emotion Prediction with Four Situations: To assess ChatGPT-4's capacity for predicting emotional evolution, we performed a zero-shot prompting experiment. We curated a dataset of images spanning six emotions and provided the model with four unique situational prompts.

a) Prompt Engineering Strategy: Crucial to the reproducibility of Large Language Model research is the structure of the prompt. We utilized a zero-shot Chain-of-Thought (CoT) style prompt to encourage the model to reason about the facial features before predicting the emotional shift. The standard prompt template used is shown below:

This structured approach minimizes parsing errors and standardizes the output for automated scoring.

4) LLM querying, output parsing, and scoring pipeline: A practical challenge in LLM evaluation is that outputs are free-form by default. To enable automated scoring, we enforce a structured JSON output (Figure 1) and apply a strict parsing-and-normalisation pipeline:

TABLE III
LABEL HARMONISATION USED IN EXPERIMENTS AND MODELLING.

Source label	Model state E_i	Valence $g(\cdot)$
happiness / happy	Joy	$+1$
surprise	Surprise	$+1$ (often valence-ambiguous in practice)
neutral	(Neutral; evaluation-only)	0 (excluded from binary valence)
anger	Anger	-1
sadness / sad	Sadness	-1
disgust	Disgust	-1

System Prompt: You are an expert psychologist specializing in facial micro-expressions and emotional dynamics.
Input: [Image File]
User Query: 1. Identify the current emotion shown in the image. 2. Consider the following scenario: [Insert Scenario, e.g., "They receive a $1,000 reward"]. 3. Based on the initial emotion and the scenario, predict the most likely subsequent emotional state. 4. Provide a confidence score (1-3) for your prediction.
Output Format: JSON {current_emotion, predicted_emotion, confidence}

Figure 1. Zero-shot prompt template used for emotion prediction.

- **Output normalisation:** Map synonyms (e.g., "happy"→"happiness") and enforce the label set in Table III. If an output label is out-of-set, we map it to the nearest valence-consistent category when possible; otherwise it is marked as invalid.
- **Confidence as a score:** The confidence field (1–3) is treated as an ordinal score used for ROC/AUC where applicable. If confidence is missing, a default mid-score is assigned to avoid discarding samples.
- **Binary valence evaluation:** For valence-only tasks, Neutral is excluded and we map labels to $\{+, -\}$ via Table III.

Algorithm 2 summarises the end-to-end evaluation procedure used to produce confusion matrices and ROC/AUC.

Remark on undefined metrics (NaN). In some one-vs-rest settings, the denominator of Sensitivity ($TP + FN$) or Specificity ($TN + FP$) can be zero (e.g., if no samples of a target class remain after filtering, or if the model never predicts a class under a specific condition). In these cases the metric is mathematically undefined and we report NaN to avoid misleading values.

5) Preliminary Results: Table IV reports ChatGPT-4's predictions of emotion evolution. For images initially labeled negative, accuracy in negative contexts was 79.4%; in positive contexts it was 72.8%. For images initially labeled positive, accuracy was higher in positive than in negative contexts. This aligns with intuition: negative states are less likely to flip to positive under a positive context than to persist under a negative one; similarly, positive states are more stable in positive contexts.

Algorithm 2: Reproducible LLM evaluation pipeline.

```
Input: Dataset $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}$, prompt set $\mathcal{Q}$,
  label map $\phi(\cdot)$, valence map $g(\cdot)$.
Output: Confusion matrices;
  Accuracy/Sensitivity/Specificity; AUC where
  applicable.
foreach $(x_n, y_n) \in \mathcal{D}$ do
    foreach $q \in \mathcal{Q}$ do
        Query LLM with (image $x_n$, prompt $q$) $\rightarrow$ raw
          text $r$;
        Parse $r$ as JSON $\rightarrow$ $(\hat{y}^{\text{cur}}, \hat{y}^{\text{next}}, \hat{c})$;
        Normalise labels: $\hat{y} \leftarrow \phi(\hat{y})$;
        Update task-specific counters (six-way or
          valence-only);
        Store score $\hat{c}$ for ROC/AUC when defined;
    end
end
Compute metrics from confusion matrices; compute
  ROC/AUC from stored scores.
```

TABLE IV
RESULT OF FOUR DIFFERENT SITUATIONS

Emotion	Parameter	Positive Situation	Negative Situation
Anger	accuracy	68.30%	73.30%
	sensitivity	NaN	NaN
	specificity	68.30%	73.30%
Disgust	accuracy	78.30%	85.00%
	sensitivity	NaN	NaN
	specificity	78.30%	85.00%
Happiness	accuracy	91.70%	83.30%
	sensitivity	91.70%	83.30%
	specificity	NaN	NaN
Neutral	accuracy	86.70%	83.30%
	sensitivity	86.70%	83.30%
	specificity	NaN	NaN
Sad	accuracy	71.70%	80.00%
	sensitivity	NaN	NaN
	specificity	71.70%	80.00%
Surprise	accuracy	85.00%	90.00%
	sensitivity	85.00%	90.00%
	specificity	NaN	NaN
Negative	accuracy	72.80%	79.40%
	sensitivity	NaN	NaN
	specificity	72.80%	79.40%
Positive	accuracy	87.80%	85.60%
	sensitivity	87.80%	85.60%
	specificity	NaN	NaN

Given safety considerations, we focus on anger, disgust, and sadness. For negative starting emotions followed by positive events (zero shot), the predictive precision ranks disgust, sadness, anger, with FPR of 78.3%, 71.7% and 68.3%, respectively. Anger appears most resistant to immediate improvement under positive events, whereas disgust—being semantically heterogeneous (e.g., dislike, contempt, displeasure)—shows the highest apparent accuracy.

6) Analysis and Discussion: Two issues emerged during evaluation. First, some dataset images diverge from common real-world interpretations. Second, there is a policy mismatch between ChatGPT-4's open-ended descriptions and the dataset's labeling guidelines: for example, an image tagged as "anger" in the dataset may be read as "sadness" or "confusion" by the model. These observations imply two practical paths. If strict adherence to the dataset taxonomy is not required, performance can be improved via prompt refinement (e.g., enumerating candidate emotions and contextual cues) and human-in-the-loop review. If strict adherence is required, prompt engineering alone is unlikely to suffice; supervised fine-tuning is the more appropriate strategy.

C. Emotion Prediction with Different Categories of Emotional Sentences

1) Dataset: In the second task, we added a dataset called MELD [22]. **MELD** The Multimodal EmotionLines Dataset (MELD) builds upon and enriches the original EmotionLines dataset by incorporating additional modalities such as audio and visual elements alongside text. MELD features over 1,400 dialogue sequences and 13,000 spoken exchanges drawn from the "Friends" TV series.

2) Task Definition: Part Two mirrors Part One by using the same image set, but augments each image with six emotion-conditioned utterances. To assess cross-model diversity, we run the identical protocol with the Doubao large language model [23] and compare outputs.

3) Preliminary Results: Overall accuracy (highest→lowest) is: happiness, surprise, neutral, anger, sadness, disgust. Within the "positive" set, happiness is generally most accurate; the main failure mode is a direct flip from happiness to anger, which yields the lowest accuracy for that class. Surprise and neutral track closely—consistent with ChatGPT-4's descriptions that treat both as valence-ambiguous. Among negative emotions, disgust is hardest to judge, reflected in the highest FPR (per the definition above) and the lowest accuracy. As in earlier tasks, zero-shot prompts are often insufficient for fine-grained negative labels: ChatGPT-4 reliably detects "negative" vs. "positive," but needs richer cues to distinguish specific negative categories.

The comparison model shows similar trends. Table VII contrasts accuracies for ChatGPT-4 and the Doubao LLM [23]. Doubao is notably less accurate on negative emotions, frequently defaulting to neutral or even (in zero-shot) misclassifying negatives as positive—patterns not observed with ChatGPT-4. While ChatGPT-4 may still confuse specific negative types (e.g., disgust vs. anger), it typically identifies that the affect is negative, explaining its stronger performance on emotion-evolution prediction.

Building on the earlier definitions, this section focuses on the Empirical ROC Area. The empirical Area Under the Curve (AUC) measures a model's ability to distinguish positives from negatives. From our data, sensitivities across the three datasets are broadly similar except for prompts expressing disgust. When the initial state varies, ChatGPT-4 finds disgust hardest to identify—e.g., in positive contexts it may reinterpret disgust as banter or a prank, reducing sensitivity. Specificity, however, is consistently strong, especially when the initial sentiment is positive, where predictions are nearly always correct. Taken together with the ROC curves, these results indicate that ChatGPT-4's emotion-conditioned sentence predictions perform better than anticipated.

IV. Discussion and Conclusion

A. Ethical Considerations and Limitations

While the ability of LLMs to predict emotional states offers significant benefits for empathetic human-computer interaction, it raises substantial ethical concerns. First, reliance on facial analysis for emotion detection has been criticized for potential bias; systems often perform poorly on underrepresented demographic groups if the training data is not diverse. In our study, although we used diverse datasets (Natural Human Faces), the underlying LLM's training distribution remains opaque.

Second, the "black box" nature of models like ChatGPT-4 presents a challenge for clinical deployment. If a model predicts a high risk of negative emotional spiraling (e.g., depressive states), the lack of explainability makes it difficult for human practitioners to trust the output without verification. Our Markov-based model attempts to mitigate this by imposing a mathematical structure on the output, but the core inference remains opaque.

Lastly, privacy is paramount. Real-time emotion tracking implies constant surveillance of user expressions. Any implementation of such systems must adhere to strict data privacy standards, ensuring that emotional data is processed locally where possible and not stored without explicit consent.

B. Failure Mode Taxonomy and Practical Implications

Across both tasks, errors are not uniformly distributed; they follow recurring patterns that are useful for both modelling and deployment.

(1) Valence-correct but category-wrong. A common outcome is that the model correctly predicts negative vs. positive affect while confusing specific negative labels (e.g., Disgust vs. Anger, or Disgust vs. Sadness). This directly supports the hierarchical assumption in Section II: a coarse valence layer can be stable even when fine-grained categorical boundaries are blurred.

(2) Ambiguity between Neutral and Surprise. Surprise is frequently treated as valence-ambiguous by the model, especially when facial cues are subtle. In practice, these

TABLE V
EXAMPLE OF SIX DIFFERENT CATEGORIES EMOTIONAL SENTENCES.

Dataset	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6
	What is the emotion of this person? If the next thing they say is, “Well, why don’t you tell her to stop being silly!” What do you think their emotions will become?	If the next sentence they say is, “Say it louder, I don’t think the guy in the back heard you!” What do you think their emotions will become?	If the next sentence they say is, “Guess what, I got an audition!” What do you think their emotions will become?	If the next sentence they say is, “Great. He’s doing great. Don’t you worry about him?” What do you think their emotions will become?	If the next sentence they say is, “Yeah but we won’t be able to like to get up in the middle of the night and have those long talks about our feelings and the future.” What do you think their emotions will become?	If the next sentence they say is, “Look what I got! Look what I got! Can you believe they make these for little people?” What do you think their emotions will become?
	What is the emotion of this person? If the next thing they say is, “Well, why don’t you tell her to stop being silly!” What do you think their emotions will become?	If the next sentence they say is, “Say it louder, I don’t think the guy in the back heard you!” What do you think their emotions will become?	If the next sentence they say is, “Guess what, I got an audition!” What do you think their emotions will become?	If the next sentence they say is, “Great. He’s doing great. Don’t you worry about him?” What do you think their emotions will become?	If the next sentence they say is, “Yeah but we won’t be able to like to get up in the middle of the night and have those long talks about our feelings and the future.” What do you think their emotions will become?	If the next sentence they say is, “Look what I got! Look what I got! Can you believe they make these for little people?” What do you think their emotions will become?
	What is the emotion of this person? If the next thing they say is, “Well, why don’t you tell her to stop being silly!” What do you think their emotions will become?	If the next sentence they say is, “Say it louder, I don’t think the guy in the back heard you!” What do you think their emotions will become?	If the next sentence they say is, “Guess what, I got an audition!” What do you think their emotions will become?	If the next sentence they say is, “Great. He’s doing great. Don’t you worry about him?” What do you think their emotions will become?	If the next sentence they say is, “Yeah but we won’t be able to like to get up in the middle of the night and have those long talks about our feelings and the future.” What do you think their emotions will become?	If the next sentence they say is, “Look what I got! Look what I got! Can you believe they make these for little people?” What do you think their emotions will become?

TABLE VI
RESULT OF SIX DIFFERENT CATEGORIES EMOTIONAL SENTENCES.

Emotion	Anger sentence	disgust Sentence	Happiness sentence	Neutral Sentence	Sad sentence	Surprise sentence
Anger	70.00%	86.70%	86.70%	86.70%	86.70%	83.30%
Disgust	60.00%	70.00%	60.00%	56.70%	83.30%	56.70%
Happiness	70.00%	96.70%	100.00%	96.70%	96.70%	96.70%
Neutral	76.70%	86.70%	96.70%	96.70%	90.00%	90.00%
Sad	63.30%	76.70%	76.70%	76.70%	86.70%	86.70%
Surprise	73.30%	86.70%	96.70%	96.70%	93.30%	96.70%

TABLE VII
ACCURACY OF DIFFERENT LARGE LANGUAGE MODELS.

LLM	Negative Emotion Accuracy	Positive Emotion Accuracy
ChatGPT	68.89%	80.56%
Doubao	26.11%	40%

TABLE VIII
RESULT OF DATASET FOR SIX DIFFERENT CATEGORIES EMOTIONAL SENTENCES

Dataset	Parameter	Anger Sentence	Disgust Sentence	Happiness Sentence	Neutral Sentence	Sad Sentence	Surprise Sentence
Emotion Detection	Accuracy	88.30%	53.30%	93.30%	90.00%	71.70%	91.70%
	Sensitivity	83.30%	30.00%	96.70%	96.70%	70.00%	93.30%
	Specificity	93.30%	76.70%	90.00%	83.30%	73.30%	90.00%
	Empiric ROC Area	0.989	0.837	0.997	0.994	0.92	0.993
Facial Expression	Accuracy	81.70%	58.30%	93.30%	91.70%	78.30%	95.00%
	Sensitivity	83.30%	46.70%	100%	100%	83.30%	96.70%
	Specificity	80.00%	70.00%	86.70%	83.30%	73.30%	93.30%
	Empiric ROC Area	0.967	0.84	1	1	0.956	0.998
Neutral Human	Accuracy	73.30%	58.30%	93.30%	93.30%	79.70%	85.00%
	Sensitivity	76.70%	50.00%	100%	100%	79.30%	100%
	Specificity	70.00%	66.70%	66.70%	86.70%	80.00%	70.00%
	Empiric ROC Area	0.93	0.833	1	1	0.959	1

confusions can inflate six-way errors while leaving valence-level performance relatively strong, depending on the mapping used.

(3) Dataset-label vs. commonsense mismatch. Several images in crowd-sourced datasets encode expression intensity, pose, or occlusion patterns that do not align cleanly with everyday interpretations. This produces "apparent errors" that may actually reflect label noise. In safety-sensitive settings, a conservative design choice is to prioritise reliable detection of *negative valence* over precise negative subtyping, and then escalate ambiguous cases to human review.

(4) Prompt sensitivity. The same image can yield different predicted transitions under small variations in wording, especially for negative emotions. This motivates the use of structured prompts (Figure 1), explicit candidate label sets, and (when feasible) repeated trials with aggregation to reduce variance.

C. Future Work

Our evaluation relies on static inputs (single images or texts), whereas real emotions evolve during interaction. Without real-time feedback to update predictions, immediate applicability to adaptive systems (e.g., conversational agents or monitoring tools) is limited. Although we center on ChatGPT-4 for image-based emotion recognition, future comparisons with other LLMs (e.g., Claude 3) and real-world trials are needed to assess robustness and generalizability. Improving transparency and accuracy may involve prompt refinement or supervised fine-tuning. Because responses are stochastic, single-trial outputs can vary; repeated runs with fixed seeds and averaged results would provide more reliable estimates and reduce variance-driven bias. Finally, judgments based solely on perceived emotional shifts can introduce labeling bias; careful protocol design and human review remain important.

We examined ChatGPT-4's zero-shot performance on image–text emotion interpretation and compared it with the Doubao model. ChatGPT-4 generally achieves higher accuracy, though it can confuse specific negative categories (e.g., classifying disgust as sadness/depressive affect). Targeted prompts and mental-health-aware guidance improve inference quality. Doubao underperforms ChatGPT-4 overall and, in zero-shot settings, more often maps negative affect to neutral or positive. For subjective tasks, we recommend prompt templates with explicit emotion taxonomies and illustrative exemplars; where strict adherence to dataset labels is required, supervised fine-tuning is likely necessary to align outputs with annotation guidelines. Finally, divergences between dataset tags and real-world perceptions can introduce bias; comparing human assessments with model outputs helps surface and correct such mismatches.

References

[1] Clement H. C. Leung and Zhifei Xu. "Predicting Emotion States Using Markov Chains". In: *BRAININFO 2025, The Tenth International Conference on Neuroscience and Cognitive Brain Information*. IARIA. 2025, pp. 7–16.

[2] Clement H. C. Leung, James J Deng, and Yuanxi Li. "Enhanced Human-Machine Interactive Learning for Multimodal Emotion Recognition in Dialogue System". In: *Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence*. 2022, pp. 1–7.

[3] Ben Mann et al. "Language models are few-shot learners". In: *arXiv preprint arXiv:2005.14165* (2020).

[4] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27730–27744.

[5] *Open AI GPT 4*. https://openai.com/gpt-4. 2023.

[6] Tianlin Zhang et al. "Natural language processing applied to mental illness detection: a narrative review". In: *NPJ digital medicine* 5.1 (2022), p. 46.

[7] Davide Ciraolo et al. "Emotional Artificial Intelligence Enabled Facial Expression Recognition for Tele-Rehabilitation: A Preliminary Study". In: *2023 IEEE Symposium on Computers and Communications (ISCC)*. IEEE. 2023, pp. 1–6.

[8] *Fat cat incident*. https://sports.sohu.com/a/776021122_121856967.

[9] Russell Lewis and Joel Rose. *'I'm not okay,' off-duty Alaska pilot allegedly said before trying to cut the engines*. https://www.npr.org/2023/10/24/1208244311/alaska - airlines - off - duty - pilot - switch - off - engines. OCTOBER 25, 2023, 11:55 AM ET.

[10] AV Geetha et al. "Multimodal Emotion Recognition with deep learning: advancements, challenges, and future directions". In: *Information Fusion* 105 (2024), p. 102218.

[11] Rui Zhang et al. "Predicting emotion reactions for human–computer conversation: A variational approach". In: *IEEE Transactions on Human-Machine Systems* 51.4 (2021), pp. 279–287.

[12] Kailai Yang et al. "On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis". In: *arXiv preprint arXiv:2304.03347* (2023).

[13] Weixiang Zhao et al. "Is ChatGPT Equipped with Emotional Dialogue Capabilities?" In: *arXiv preprint arXiv:2304.09582* (2023).

[14] Hoai-Duy Le et al. "Multi-Label Multimodal Emotion Recognition With Transformer-Based Fusion and Emotion-Level Representation Learning". In: *IEEE Access* 11 (2023), pp. 14742–14751.

[15] Wei Zhang, Xuanyu He, and Weizhi Lu. "Exploring discriminative representations for image emotion recognition with CNNs". In: *IEEE Transactions on Multimedia* 22.2 (2019), pp. 515–523.

[16] Paul Ekman. *Facial expressions of emotion: New findings, new questions*. 1992.

[17] Plutchik Robert. *Emotion: Theory, research, and experience. vol. 1: Theories of emotion*. 1980.

[18] Ronak Kosti et al. "Emotion recognition in context". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1667–1675.

[19] Ronak Kosti et al. "Context based emotion recognition using emotic dataset". In: *IEEE transactions on pattern analysis and machine intelligence* 42.11 (2019), pp. 2755–2766.

[20] Lutfiah Zahara et al. "The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi". In: *2020 Fifth international conference on informatics and computing (ICIC)*. IEEE. 2020, pp. 1–9.

[21] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. "Affectnet: A database for facial expression, valence, and arousal computing in the wild". In: *IEEE Transactions on Affective Computing* 10.1 (2017), pp. 18–31.

[22] Soujanya Poria et al. "Meld: A multimodal multi-party dataset for emotion recognition in conversations". In: *arXiv preprint arXiv:1810.02508* (2018).

[23] *Tik Tok*. https://www.doubao.com/chat/?channel=baidu_pz & source = db_baidu_pz_01 & keywordid = weizhi7. August 17, 2023.

Feasibility of an AI-Driven Wearable Ring for Shoulder Mobility Monitoring in Older Adults with and without Dementia

Holly Shannon
Department of Physical Therapy,
University of Manitoba
Manitoba, Canada
Holly.shannon@umanitoba.ca

Jennifer O'Neil
School of Rehabilitation Sciences,
University of Ottawa
Ottawa, Canada
joneil@uottawa.ca

Olga Theou
School of Physiotherapy,
Dalhousie University
Nova Scotia, Canada
olga.theou@dal.ca

Logan Young
Department of Health Sciences,
Carleton University
Ottawa, Canada
loganyoung@cmail.carleton.ca

Makara Rolle
Department of Health Sciences,
Carleton University
Ottawa, Canada
makararolle@cmail.carleton.ca

Jose Carlos Tatmatsu-Rocha
College of Medicine,
Federal University of Ceará-UFC
Ceará, Brazil
Tatmatsu@ufc.br

Dahlia Kairy
École de Réadaptation,
Université de Montréal
Montreal, Canada
Dahlia.kairy@umontreal.ca

Ke Peng
Department Electrical and Computer Engineering,
University of Manitoba
Manitoba, Canada
Ke.peng@umanitoba.ca

Asma Seraj Pour Shooshtari
Department Electrical and Computer Engineering,
University of Manitoba
Manitoba, Canada
serajpoa@myumanitoba.ca

Zahra Moussavi
Department Electrical and Computer Engineering,
University of Manitoba
Manitoba, Canada
Zahra.moussavi@umanitoba.ca

Mirella Veras
Department of Physical Therapy,
University of Manitoba
Manitoba, Canada
Mirella.veras@umanitoba.ca

***Abstract*—Wearable technologies powered by artificial intelligence (AI) can offer a non-invasive method to enhance health monitoring. However, the implementation of such wearable kinematic technologies among older adults with cognitive impairment remains underexplored. This study aims to evaluate the feasibility, usability, and acceptability of a wearable ring sensor powered by AI in long-term care (LTC) residents with and without dementia. A mixed-methods study was conducted with ten LTC residents (five with dementia and five without). Participants engaged in structured shoulder mobility exercises while continuously wearing an AI-integrated ring sensor for one day. Feasibility, usability, and acceptability were assessed through various questionnaires. A post-study focus group was conducted with 6 of the participants, followed by reflexive thematic analysis to identify qualitative themes. No significant differences in feasibility were found between groups**

for device usage adherence, exercise frequency and intensity. Similarly, quantitative data revealed usability, and acceptability did not significantly differ between dementia and non-dementia participants. However, participants without dementia reported a significantly more positive attitude toward the technology. Thematic analysis identified three key themes: high ring comfortability, low ring significance, and ease of use. The AI-integrated wearable ring sensor was well accepted across varying degrees of cognitive impairment, highlighting the non-intrusive nature. Our findings suggest feasibility, usability, and acceptability of the wearable ring device in a LTC setting. Future research should explore its usability in a larger population of individuals with varying cognitive impairment and assess its clinical utility for movement monitoring in older adults.

Keywords—Wearable Devices; Artificial Intelligence; Dementia; Feasibility; Aging; cognitive impairment, Remote Movement Monitoring, Long Term care.

I. Introduction

This is an extended version of the paper published at AIVR 2025 [1]. The current manuscript presents data from the full sample, allowing comparisons between individuals living with and without dementia.

Dementia affects memory, thinking, behavior, and the ability to perform daily activities. The World Health Organization identifies dementia as a critical public health and social care issue of the 21st century [2]. Currently, 35.6 million people worldwide live with dementia, and this number is projected to double by 2030 and triple by 2050 [2], [3]. Historically, people with dementia and cognitive disabilities have been systematically excluded from geriatric research, reflecting a broader pattern of ableism that has marginalized individuals living with dementia [4]. However, this paradigm has started to shift over the past decade, with growing awareness of the importance of addressing these biases and including diverse populations in health technology research to promote equitable opportunities for access, utilization, and benefits from technological advancements [5][6][7][8].

This shift toward inclusivity is especially significant in the context of advancing technologies like wearable devices, which have the potential to improve care for dementia populations [9]. Advancements in kinematic technology, such as accelerometers, GPS trackers, gyroscopes, and motion detection tools integrated into mobile platforms, present a cost-effective means to assess disease burden and deliver personalized care [5]. Likewise, these innovative kinematic technologies enable minimally invasive and real-time monitoring for tailored delivery [9]. Wearable devices (WDs), capable of continuously monitoring physiological metrics in real-world settings such as a patient's home (i.e., smartwatches), provide insights that surpass those of traditional in-clinic assessments [10].

Wearable devices, including smart bracelets, rings, belts, necklaces, glasses, watches, earphones, headbands, and clothing with built-in sensors, are generally used to measure physiological parameters (e.g., heart rate, breathing rate, etc.) or to monitor physical movement [9][11]. Wearable devices for tracking physical movement such as range of motion, are increasingly being used, especially for individuals with neurological or musculoskeletal impairments [12]. These wearable technologies support rehabilitation and address the needs of aging populations, by providing real-time data which informs strategies to help preserve mobility and daily functioning in older adults [12]. Tracking upper body movements can contribute to maintaining mobility and activities of daily living (ADL) in older adults [13].

Various research on the use of wearable technologies for monitoring movement, including upper body functioning, has evolved alongside advancements in the field of kinematic technology. Early studies focused on inertial measurement unit (IMU)-based devices, accurately tracking shoulder joint angles during ADLs [14][15]. With the introduction of smartwatches in subsequent years, research expanded to include wearable IMU-based devices, leveraging their ability to monitor movements and assess rehabilitation progress in real-life situations and over a longer period of time. Wearable IMU-based devices are widely used to assist in tracking movements, making them integral tools in health monitoring [16]. Studies exploring the use of smartwatches using upper extremity rehabilitation exercises measure shoulder function indirectly [17]. Wearable technologies such as wearable rings have emerged as a potential alternative. However, research on the use of wearable rings has largely focused on other health monitoring applications, such as measuring blood pressure or tracking action-planning impairments [18][19].

Artificial intelligence (AI) is significantly changing healthcare, offering innovative solutions for managing dementia [20][21]. AI-driven tools, such as wearables, assistive robots and telepresence systems, provide cognitive support, medication reminders, and opportunities for social interaction, improving both the well-being of patients and the lives of their caregivers. These technologies have demonstrated benefits, including reduced caregiver burden, enhanced patient engagement, and improved mental health [20].

Healthcare services for disease diagnosis and monitoring are often expensive and limited in accuracy, driving interest in wearable health technologies based on flexible electronics. These devices offer benefits such as reduced costs, non-invasive implementation, and real-time access to health data, enabling personalized health monitoring through the accurate measurement of physical and biochemical signals [22]. AI algorithms enhance the functionality of these wearables, analyzing movement patterns and enabling precise tracking of motor activity, early intervention, and tailored care [20]. AI may improve data accuracy, with the potential to facilitate real-time decision-making and promote inclusivity in research through seamless and accessible monitoring [22]. Expanding on these advancements, AI-powered wearable devices, such as a ring sensor designed to monitor shoulder movements, present a novel approach to supporting individuals with dementia. However, the feasibility, usability, and acceptability of such AI-powered wearable devices have not been extensively studied in older adults, especially when considering individuals living with dementia.

This study aims to assess the feasibility, usability, and acceptability of a wearable ring powered with AI designed to track upper body movements, comparing individuals with and without dementia in a long-term care (LTC) facility. It focuses on evaluating how well the device meets the specific needs of both groups and identifying factors that influence its usability and overall acceptance.

II. Methods

A. Study Design

This pilot study employed an explanatory sequential mixed methods to assess the feasibility, usability, and acceptability of wearable sensor technology for older adults in LTC facilities [23]. The initial phase involved using quantitative methods to document feasibility, usability, and acceptability. This provided information into the practicality and potential success of the intervention. Following the quantitative phase, qualitative methods, including a focus group, were used to explore participants' experiences and the factors influencing the adoption of the technology.

B. Participants

Participants were recruited from a LTC facility in a rural area of Nova Scotia, Canada. Convenience sampling was used to select 10 participants, ensuring variability in functional abilities, cognitive function, and health status. Older adults (aged 65 and above) residing in the LTC facility were included if informed consent was obtained, either directly from the resident or from their substitute decision-maker when appropriate. Exclusion criteria included: 1) significant mobility restrictions, or 2) medical conditions that could interfere with sensor use. These conditions included severe hand arthritis, hand tremors, Raynaud's disease, skin conditions (such as dermatitis or eczema), and hand injuries (previous hand injuries or surgeries). Participants with motor impairments were excluded as it would limit their ability to perform the upper body movements required for tracking, preventing meaningful data collection. The potential for discomfort or confusion from using the device could also lead to distress, affecting participant well-being. For these reasons, these individuals were excluded to ensure accurate data collection and to prioritize participant comfort and safety.

C. Intervention

Participants were asked to wear the AI-driven ring sensor to monitor upper-body movements during the one-week intervention period. The LTC facility site coordinator provided instructions to participants to ensure proper use and maintenance of the device, supporting its functionality throughout the study. Participants with dementia were instructed to wear the sensor continuously for one day from 8:30 am until 3:30 pm. This approach was used to assess the feasibility of continuous wearing of the ring device to determine if participants could maintain wearing the device, without removal. Participants without dementia were instructed to wear the device only during exercise or recreational activities and to remove the ring afterwards. This contrasting protocol was implemented as part of a later phase of the study aimed to explore capabilities of the ring device. The site coordinator monitored the residents' use of the device and reviewed collected data daily to assess progress and address any concerns. The intervention prioritized accurate data collection while ensuring participant safety and comfort.

D. Intervention

Each participant was provided with a ring device by XO TECHNOLOGY©, along with information regarding its use [24]. However, the primary focus was on assessing the feasibility, usability, and acceptability of wearing the ring, so participants did not interact with the app themselves during the study period. The XO HEALTH© app, which displayed details such as Participant ID, Start and End Period, Last Data Sync, Average Wear Time, Device ID, and Device Status, was installed on Android tablets running the Android operating system or Apple iPads on iOS. A personal account was created on the XO HEALTH platform for each participant, enabling the device to collect and store data. The software platform utilized AI algorithms and data collection to monitor and analyze everyday shoulder movements. Data collected includes the angle of shoulder flexion, extension, abduction, adduction, internal rotation and external rotation, along with the number of repetitions for each. The collected data are processed by a neural network in order to classify various types of daily activities and quantify the frequency and intensity of these shoulder activities. Employing machine learning techniques, the platform could identify anomalous data points and deliver actionable insights, possibly enabling early detection of potential issues and facilitating proactive health risk mitigation. Further exploration into the ring device capabilities will be addressed in a later phase of the study.

E. Quantitative Data Collection and Measures

Data collection was conducted from October 21st to 25th, 2024, by a research assistant, with support from the site coordinator. Demographic information and cognitive status were obtained from the participant's medical record at the start of the study visit. The demographic questionnaire captured the age, gender, medical history, and functional status of all participants. The Mini-Mental State Examination (MMSE) was used to assess cognitive impairment [25]. Through a data-sharing agreement, the most recent MMSE scores (i.e., within the last 6 months) were obtained for each participant via their records at the LTC facility. For this study, "dementia" classification refers to participants with MMSE scores consistent with up to moderate Alzheimer's disease, using a cutoff of ≤20, whereas "non-dementia" refers to those scoring ≥21. These thresholds align with the following ranges: normal cognition (≥25), mild Alzheimer's disease (21–26), moderate Alzheimer's disease (10–20), and moderately severe Alzheimer's disease (10–14) [25]. Feasibility was assessed by tracking adherence to device usage and monitoring shoulder exercises between participant groups via the observational checklist. These measures allowed for an evaluation of the technical and operational feasibility of the device by recording the time and exercises performed. Usability and acceptability were documented after completing the intervention using the Technology Acceptance Questionnaire (TAQ), and the User Acceptance Questionnaire (UAQ) [26], [27]. The TAQ consists of 12 items on a 7-point Likert Scale and focuses on both perceived usefulness and perceived ease of use of the sensor. The UAQ involves 26 items on a 6-point Likert scale, that comprehensively assess acceptance based on a range of questions about comfort, enjoyment, effort expectancy, attitude toward technology, etc.

F. Qualitative Data Collection and Measures

Approximately one week after the intervention period (November 5, 2024), participants who had completed the intervention were invited to participate in a semi-structured focus group conducted at the LTC facility with a trained staff member. A focus group was used to foster interaction among participants and encourage their expression of their perceptions of the sensor. A research assistant joined the focus group online using Zoom (Zoom Video Communications Inc.) to facilitate participation, while the site coordinator asked predetermined questions to prompt discussion. Focus group questions were developed to explore further comfort, benefits, concerns, and the impact on daily activities (see Supplementary Material for the interview guide). The research assistant transcribed and anonymized the audio recordings of the focus group discussions on Zoom using the qualitative software QSR NVivo 14.

G. Statistical Analysis: Quantitative Analysis

All questionnaire data were presented as mean and standard deviation and initially assessed for normality using the Kolmogorov-Smirnov test. Since the data did not follow a normal distribution, comparisons between groups were made using the Mann-Whitney U test. Categorical variables were reported as absolute and relative frequencies, with group differences analyzed using Fisher's exact test. All statistical analyses were conducted with a 95% confidence interval using SPSS (version 28.0); IBM Corp, Armonk, NY) for Mac. Qualtrics data management system (Qualtrics International Inc.) was used for data capture. These methods were selected to ensure a robust analysis of differences between dementia and non-dementia participants, considering the small sample size and the distribution characteristics of the data.

H. Statistical Analysis: Qualitative Analysis

The qualitative data was analyzed following the Braun and Clarke (2019) reflexive thematic analysis methodology [28]. Our approach followed a constructivist epistemology and an experiential orientation, whereby the three authors (HS, LY, MR) first read all transcripts to become familiar with the full dataset. The authors engaged in reflexive journaling and independently generated initial codes through an approach driven mainly by a latent-coding perspective and inductive analysis. Finally, themes were then generated and refined through discussion among these authors. Our reporting adheres to the Standards for Reporting Qualitative Research (SRQR) guideline, previously done by O'Brien et al. [29].

III. Results

There were no significant differences between participants with dementia and those without dementia across several characteristics, as illustrated in Table 1. In terms of cognitive status, scores on the Mini-Mental State Examination ranged from 5 to 30, with a mean score of 20.90 (SD ±8.84). Both groups had a similar biological sex distribution, with 80% females and 20% males in each group.

Table 1: Sociodemographic Characteristics of participants

Category	**Dementia** (n=5)	**Non-Dementia** (n=5)	**P-value**
Gender			
Women	4 (80.0%)	4 (80.0%)	1.000
Man	1 (20.0%)	1 (20.0%)	
Ethnicity			
White	5 (100.0%)	5 (100.0%)	1.000
Other	0 (0.0%)	0 (0.0%)	
Highest Level of Education			
High School or Equivalent	4 (80.0%)	4 (80.0%)	1.000
Other	1 (20.0%)	1 (20.0%)	
Age (Mean ± SD)	78.60 ± 81.60	81.60 ± 80.10	0.917

Note. P<0.05 indicated statistical significance based on the Mann-Whitney test (mean±SD) or Fisher's exact text (n,%).

Regarding participation in recreational activities involving shoulder exercises, 100% of non-dementia participants and 80% of dementia participants were involved. The majority of participants in both groups reported no shoulder pain or discomfort with the device (*see Table 2*). Overall, the lack of significant differences in these variables suggests that they did not influence the comparison between dementia and non-dementia participants in this study. The participants did not report adverse events.

Table 2: Ring Wearing Characteristics for participants

Category	Dementia (n=5)	Non-Dementia (n=5)	P-value
Duration (in seconds)	1703.00 ± 348.00	1025.00 ± 348.00	0.251
Engaged in Recreational Activities Involving Shoulder Exercises?			
No	0 (0.0%)	1 (20.0%)	1.000
Yes	5 (100.0%)	4 (80.0%)	
Expressed Shoulder Pain Today?			
No	4 (80.0%)	5 (100.0%)	1.000
Yes	1 (20.0%)	0 (0.0%)	
Expressed Discomfort with the Device?			
No	4 (80.0%)	5 (100.0%)	1.000
Yes	1 (20.0%)	0 (0.0%)	

Note. P<0.05 indicated statistical significance based on the Mann-Whitney test (mean±SD) or Fisher's exact text (n,%).

A. Feasibility: Shoulder Exercises

The feasibility of the device was demonstrated, as no residents removed or requested to remove the ring during the intervention period. However, an issue arose when the ring sensor size was too large for one participant, causing it to fall off. For most shoulder exercises, no significant differences were observed between the two groups (*see Table 3*).

Table 3: Comparison of shoulder range of motion exercises between participants

Type of Shoulder Range of Motion	Dementia (n=5)	Non-Dementia (n=5)	P-value
Shoulder Flexion – Number of Sets	1.33 ± 1.67	1.67 ± 1.50	0.796
Shoulder Flexion – Number of Repetitions per Set	10.00 ± 9.00	9.00 ± 9.56	0.699
Shoulder Extension – Number of Sets	1.00 ± 1.33	1.33 ± 1.17	1.000
Shoulder Extension – Number of Repetitions per Set	5.00 ± 7.50	7.50 ± 6.11	0.519
Shoulder Abduction – Number of Sets	1.67 ± 1.33	1.33 ± 1.50	0.796
Shoulder Abduction – Number of Repetitions per Set	8.00 ± 6.50	6.50 ± 7.33	0.502
Shoulder Internal Rotation – Number of Sets	1.33 ± 1.00	1.00 ± 1.17	0.317
Shoulder Internal Rotation – Number of Repetitions per Set	8.60 ± 5.00	5.00 ± 7.25	0.055
Shoulder External Rotation – Number of Sets	1.33 ± 1.00	1.00 ± 1.17	0.317
Shoulder External Rotation – Number of Repetitions per Set	6.60 ± 5.00	5.00 ± 6.00	0.121

Note. P<0.05 indicated statistical significance based on the Mann-Whitney test (mean±SD) or Fisher's exact text (n,%).

Specifically, the number of sets and repetitions for shoulder flexion, extension, abduction, and external rotation showed no significant variation, with p-values ranging from 0.317 to 0.796. However, the number of repetitions for shoulder internal rotation approached significance, with a p-value of 0.055, suggesting a potential trend where participants with dementia performed slightly more repetitions than those without dementia. Despite this, none of the differences reached the standard threshold for statistical significance (p<0.05), indicating that overall, the frequency and intensity of shoulder exercises were similar between the two groups

B. Usability and Acceptability

Overall, for usability, the results of the UAQ (*see Table 4*) indicate that there were no significant differences between the two groups for the total score and most of the questions (p > 0.05). However, one notable exception was found in UAQ_6 (attitude towards technology), where participants with dementia reported a significantly more positive attitude (p = 0.018). These findings suggest that while there may be minor variations in specific areas, the overall technology acceptance and user experience were similar between participants with and without dementia.

Table 4: Comparison of the user acceptance questionnaire between participants

Type of Shoulder Range of Motion	Dementia (n=5)	Non-Dementia (n=5)	P-value
UAQ_1: Ease of use	4.60 ± 4.75	4.75 ± 4.67	0.524
UAQ_2: Usefulness	5.40 ± 4.25	4.25 ± 4.89	0.602
UAQ_3: Perceived usefulness	3.80 ± 2.75	2.75 ± 3.33	0.197
UAQ_4: Likelihood of usage	2.80 ± 2.50	2.50 ± 2.67	0.897
UAQ_5: Interction satisfaction	4.40 ± 4.00	4.00 ± 4.22	0.107
UAQ_6: Attitude toward technology	**4.80 ± 3.50**	**3.50 ± 4.22**	**0.018***
UAQ_7: Interest in future use	2.40 ± 1.75	1.75 ± 2.11	0.618
UAQ_8: Overall satisfaction	2.00 ± 1.75	1.75 ± 1.89	0.694
UAQ_9: Perceived value	1.60 ± 2.50	2.50 ± 2.00	0.530
UAQ_10: Intention to continue use	2.60 ± 2.75	2.75 ± 2.67	0.700
UAQ_11: Likelihood of recommending	3.00 ± 1.00	1.00 ± 2.11	0.121
UAQ_12: Use in future	3.00 ± 3.25	3.25 ± 3.11	0.694
UAQ_13: Usefulness in daily life	2.60 ± 1.75	1.75 ± 2.22	0.521
UAQ_14: Impact on quality of life	2.40 ± 3.75	3.75 ± 3.00	0.258
UAQ_15: Technology frustration	1.20 ± 2.25	2.25 ± 1.67	0.302
UAQ_16: Engagement with technology	2.60 ± 2.00	2.00 ± 2.33	0.706
UAQ_17: Comfort using the technology	4.80 ± 4.25	4.25 ± 4.56	1.000
UAQ_18: Willingness to recommend	4.40 ± 4.75	4.75 ± 4.56	0.893
UAQ_19: Ease of learning technology	5.20 ± 5.25	5.25 ± 5.22	1.000
UAQ_20: Ability of troubleshoot	1.60 ± 2.25	2.25 ± 1.89	0.434
UAQ_21: Overall technology confidence	4.00 ± 4.75	4.75 ± 4.33	1.000
UAQ_22: Understanding of technology features	3.00 ± 3.00	3.00 ± 3.00	1.000
UAQ_23: Motivation to use technology	3.00 ± 2.25	2.25 ± 2.67	0.455
UAQ_24: Technology fits with needs	4.80 ± 5.00	5.00 ± 4.89	0.418
UAQ_25: Satisfaction with technology design	4.00 ± 4.25	4.25 ± 4.11	0.500
UAQ_26: Frequency of use	3.80 ± 2.50	2.50 ± 3.22	0.266
Total UAQ Score	87.80 ± 66.20	66.20 ± 77.00	0.465

Note. P<0.05 indicated statistical significance based on the Mann-Whitney test (mean±SD) or Fisher's exact text (n,%).

For acceptability, there were no significant differences between dementia and non-dementia participants for most of the TAQ items. For example, the ratings on the ease of use (TAQ_1), usefulness (TAQ_2), perceived usefulness (TAQ_3), and other items like interest in future use (TAQ_7) and overall satisfaction (TAQ_8) showed no significant differences between the two groups. Some items had slightly higher or lower scores in one group compared to the other, however, these differences did not reach statistical significance. For instance, participants with dementia rated "ease of use" and "likelihood of usage" slightly higher than those without dementia, but the p-values (0.197 and 0.193, respectively) indicated that these differences were not statistically significant. The overall total TAQ score was also not significantly different between the two groups, with a p-value of 0.251. This suggests that, despite minor variations in individual responses, the overall technology acceptance between participants with and without dementia was similar.

C. Participant Experiences

The focus group comprised six participants, with a mean age of 78.5 years (SD ±10.97). In terms of gender identity, 66.7% identified as women (n=4), and 33.3% identified as men (n=2). All participants (100%) identified as Caucasian. The qualitative analysis yielded 3 themes (see *Figure 1*). No privacy or security concerns were raised during the focus group. Only one participant identified having prior experience with using a wearable device for health or fitness monitoring.

Theme 1:
High Ring
Comfortability
Theme 2:
Low Ring
Significance
Theme 3:
Low Ring
Impact

Figure 1. Schematic summary of themes derived from the qualitative analysis.

Theme 1: High Ring Comfortability

A crucial part of using wearable devices is how comfortable they are for the individual wearing them. A major factor contributing to the comfort of the ring device is its familiarity with the participants. *“I mean, I’ve had a ring on my finger for years; I just put it on top of this one (P1).”* Many participants said that the device's design closely resembled that of a conventional ring they were used to wearing in everyday life. This resemblance made the device non-intrusive while also allowing participants to adjust to wearing it quickly. While some participants expressed worries about swelling, it did not appear to influence general comfort. Many participants expressed *“It didn’t bother me, I was comfortable with it (P2).”* However, size difficulties did arise. One individual stated that the ring felt uncomfortable since it was too large for their finger, pointing out the need to make size adjustments for the best fit.

Theme 2: Low Ring Significance

A theme that participants consistently demonstrated was a perceived low ring significance. One participant noted, *“I couldn’t see any difference when I had it on (P3)”,* underscoring the lack of discerned impact and benefit from the ring. Additionally, participant 2 stated, *“Think I need more information on it,”* when asked how important having a ring to track their shoulder movements and exercises was to them. This statement demonstrates a recurring trend among respondents, as many did not feel they had sufficient information to decide if the ring made a personal difference. Furthermore, several individuals involved in the focus group expressed that they felt the ring had low significance in their lives, as they did not notice a tangible difference after using it. Participant 1 reported, *“I didn’t even really know what the ring was going to do and what we were supposed to do”*, illustrating that multiple participants were under the impression it would provide observable results after completion of the study.

Theme 3: Ease of use

The final emerging theme centered on the ease of use of the ring sensor in participants' daily lives. Several individuals reported that they often forgot they were wearing the ring, which enhanced their confidence and comfort in moving through daily routines without feeling as though they were part of a study. “*It was very easy. You can wash with it on and shower. Go outside. And it's perfect for me*”. Participants were able to complete daily activities like exercising, showering, and recreational activities without any interruption from the ring. Participant 8 explained; “*I don't feel it had any real impact. I used it for most things.”* Overall, the ring did not have any negative outcome on participants.

IV. Discussion

This mixed-methods study assessed the usability and acceptance of an AI-powered wearable ring sensor designed to track upper body movements. This study introduces the novelty of assessing a wearable ring device among individuals with dementia compared to those without, whilst evaluating feasibility, usability, and acceptability. We evaluated how well the device met the specific needs of individuals with and without dementia in a LTC facility. We identified factors that influence its overall usability and acceptance. No significant differences were observed in shoulder exercises between the two groups based on the frequency or intensity of the exercises. Similarly, there were no significant differences in the total scores from the technology acceptance or user acceptance questionnaires. However, when examining the specific questions, attitudes towards technology significantly differed, whereas participants with dementia reported a more positive attitude. Prior literature has identified motivation and positive attitudes as key factors when implementing new technologies for older adults [30]. Furthermore, positive attitudes toward active aging have been found to influence learning and technical skills associated with the implementation of new devices in older adults [31][32]. Recognizing positive attitudes and motivation among participants can be a strength to build on, enhancing feasibility and engagement with wearable technologies.

Prioritizing the assessment of feasibility, usability, and acceptability provides a necessary foundation for the successful integration of new technologies into healthcare and rehabilitation for the aging population. Even if a device demonstrates strong technical performance in later stages, it will not be adopted if it is not considered acceptable to users, practical to implement, or easy to use across diverse populations. Focusing first on these dimensions allows researchers to identify barriers to adoption, cultural or contextual concerns, and potential design improvements that enhance user experience. These outcomes ensure that future research builds on a device that is not only technically promising but also aligned with the lived experiences and needs of its intended users. Understanding differences in adoption and usability between these groups is crucial, as cognitive and functional impairments may influence the device's practicality.

Feasibility, usability, and acceptability were also demonstrated in participant experiences, with three emerging main themes, 1) high ring comfortability, 2) low ring significance, and 3) low ring impact. By integrating both quantitative and qualitative results, this approach enhances the potential for real-world application and informs future advancements in wearable health technologies tailored to individuals with varying cognitive abilities.

Regarding feasibility, both dementia and non-dementia participants wore the ring sensor without removing the device. While the outcomes of this study indicate the high feasibility of implementing such a device among LTC residents, there is still room for improvement regarding the communication of study expectations and end goals between researchers and participants. Based on focus group feedback, it is evident that participant understanding would have been greatly improved had they received more information on the ring’s function, as confusion on this front was the primary reported concern. Although the authors note moderate cognitive impairment in this population could contribute to a misunderstanding of the details of the ring sensor, future research should better target digital literacy in older adults [33][34][35]. Nonetheless, the findings indicate that the wearable ring device is a feasible technology for individuals with cognitive impairment, including dementia [1]. Even in the absence of full understanding, passive compliance was maintained, whereby participants still displayed a high willingness to wear the device. These results align with findings reported by Rocha et al., affirming the use of wearable ring devices in older adult populations [9].

Individuals with dementia frequently have cognitive impairment, which might restrict their ability to utilize and accept wearable technology. As a result, while developing such devices, it is critical to prioritize aspects such as ease of use, adaptability, and intuitiveness [5]. In this study, these core aspects were integrated into the ring’s design, which significantly enhanced the acceptability of the technology. In this population, individuals often remove or avoid using devices that feel out of place or obtrusive [5][9]. The participants were so comfortable with the ring that after putting it on, they were unaware of wearing it throughout the day. The ability to put the ring on the finger and monitor movements without needing constant adjustments makes the technology highly beneficial in this population. Such simplicity reduces the cognitive load, ensuring that the user does not feel overwhelmed or frustrated [5]. These aspects enhance user acceptability and support sustained use of the device among individuals with dementia.

These findings have important implications for telerehabilitation, particularly for older adults in rural, remote and underserved settings where in-person monitoring is limited. Evidence from recent rapid reviews supports the feasibility and effectiveness of wearable and sensor-based monitoring in delivering remote rehabilitation to populations with limited access to in-person care [36]. The high comfort and acceptability of the AI-powered ring among residents with varying cognitive abilities suggest that similar wearable technologies could be integrated into remote rehabilitation programs to support continuous, unobtrusive movement monitoring. Such integration would enable clinicians to receive real-time data on upper limb mobility without requiring complex user interaction, addressing barriers related to geography, mobility limitations, and cognitive impairment, and thereby promoting equity in access to rehabilitation services. From an ethical perspective, the deployment of AI-powered wearables in these contexts must ensure that data collection, storage, and use respect privacy, autonomy, and informed consent, particularly for individuals with cognitive impairment, while also avoiding the risk of exacerbating digital health inequities [37].

A. Limitations and Future Directions

This study had some limitations that should be acknowledged when interpreting the results. First, individuals with significant mobility restrictions or medical conditions that could interfere with sensor use, such as severe hand arthritis, hand tremors, Raynaud's disease, skin conditions, or previous hand injuries, were excluded. These exclusions were made to ensure the accuracy and reliability of data collection, as these conditions could compromise participants' ability to use the wearable ring effectively or lead to discomfort and distress. As a result, the study's findings may not fully represent the experiences of individuals with more advanced physical impairments, limiting the generalizability of the results to a broader population of people with dementia. Additionally, a key limitation of the study was the lack of data from the wearable ring's app and sensor outputs. Although this data would have enhanced the study by offering insights into the device's effectiveness, this study focused on evaluating user experience, comfort, and acceptance of wearing the ring device. Validation of the AI-driven functionalities of the ring device, including AI accuracy, kinematic data reliability and user interactions with the app interface, will be examined in future work. The ring wearing protocol was intentionally designed to explore device capabilities in terms of continuous versus fragmented use of the ring. The dementia group partook in sustained ring use, to assess feasibility, given the potential challenges with adherence. However, we acknowledge that this difference in the ring wearing protocol limits direct group comparisons and therefore should be interpreted cautiously. Finally, as a pilot feasibility study, the small sample size limits statistical power and generalizability; therefore, the findings should be interpreted as preliminary. Future studies should consider a larger sample size of individuals with varying cognitive disabilities to assess how wearable technology can be adapted for their needs, including the use of wearable technology interfaces (i.e., applications). This would expand the generalizability of findings and better address the diverse experiences of people living with dementia. The limitations related to the missing data and exclusion criteria are important to consider but do not detract from the study's contribution to understanding the practical application of wearable technology in dementia care. Furthermore, while this pilot study focused primarily on the feasibility of ring wearability, future work should explore the integration of AI to enhance dementia monitoring capabilities more in depth. Although AI was not directly applied to this study, its potential in wearable data could significantly improve personalized intervention strategies.

V. Conclusions

This study evaluated the feasibility, usability, and acceptability of an AI-enhanced wearable ring for tracking upper-body movements in participants with and without dementia. No significant differences were observed between the two groups in demographics, device-related adverse events, or technology acceptance. Both groups reported similar satisfaction with the device, highlighting its non-intrusive nature and minimal impact on daily routines. Integrating AI capabilities enhances the device's ability to accurately track movement patterns and provide reliable data, making it a valuable tool for real-time monitoring. Given the small sample size, these findings should be interpreted as exploratory, as this pilot study was designed to assess feasibility rather than draw definitive conclusions about group differences. In conclusion, the wearable device was found to be acceptable for both groups. The study underscores its potential for improving care delivery, particularly in dementia care, by leveraging AI-driven data to guide clinical decisions, monitor disease progression, and personalize interventions in LTC facilities.

VI. Ethics

This study received ethical approval from the Carleton University Research Ethics Board-B (CUREBB). Ethics Clearance ID: Project # 12138. All residents and Substitute Decision Makers (SDMs) involved will provide written informed consent to participate in the study and share their personal XO HEALTH account information with LTC staff.

VII. Funding

This research was supported by the National Research Council Industrial Research Assistance Program (NRC-IRAP), Science and Technology to M.V.

VIII. Acknowledgements

We extend our heartfelt gratitude to Michaela Tracy for her exceptional work as the clinical site coordinator of this study. Her dedication, meticulous coordination, and steadfast support throughout all stages of recruitment and intervention were indispensable and instrumental to the success of this study.

We also wish to acknowledge the generous support of Krist Beeler, CEO, and Victoria Gagne, Recreation Director of the Dykeland Lodge. Their contributions and collaboration strongly supported this study and were critical to its successful implementation.

IX. Conflict of Interest

We have no conflict of interests to declare. *XO Technologies* provided the ring device for the study intervention; however the company did not provide any direct financial support to the study.

References

[1] H. Shannon *et al.*, "Acceptability of an AI-Powered Wearable Ring Sensor for Upper Body Mobility in Individuals with Cognitive Impairment: A Pilot Study," presented at the The Second International Conference on Artificial Intelligence and Immersive Virtual Reality, Apr. 2025, pp. 5-7.

[2] World Health Organization, "World report on ageing and health," World Health Organization, 2015. Accessed: Jan. 13, 2025. [Online]. Available: https://iris.who.int/handle/10665/186463

[3] G. Livingston *et al.*, "Dementia prevention, intervention, and care: 2024 report of the Lancet standing Commission," *Lancet Lond. Engl.*, vol. 404, no. 10452, pp. 572–628, Aug. 2024, doi: 10.1016/S0140-6736(24)01296-0.

[4] J. S. Taylor, S. M. DeMers, E. K. Vig, and S. Borson, "The Disappearing Subject: Exclusion of People with Cognitive Impairment and Dementia from Geriatrics Research," *J. Am. Geriatr. Soc.*, vol. 60, no. 3, pp. 413–419, 2012, doi: 10.1111/j.1532-5415.2011.03847.x.

[5] S.-Y. Chien, O. Zaslavsky, and C. Berridge, "Technology Usability for People Living With Dementia: Concept Analysis," *JMIR Aging*, vol. 7, no. 1, p. e51987, July 2024, doi: 10.2196/51987.

[6] A. C. Cote, R. J. Phelps, N. S. Kabiri, J. S. Bhangu, and K. "Kip" Thomas, "Evaluation of Wearable Technology in Dementia: A Systematic Review and Meta-Analysis," *Front. Med.*, vol. 7, Jan. 2021, doi: 10.3389/fmed.2020.501104.

[7] F. Rossetto *et al.*, "A digital health home intervention for people within the Alzheimer's disease continuum: results from the Ability-TelerehABILITation pilot randomized controlled trial," *Ann. Med.*, vol. 55, no. 1, pp. 1080–1091, Dec. 2023, doi: 10.1080/07853890.2023.2185672.

[8] J. S. Yi, C. A. Pittman, C. L. Price, C. L. Nieman, and E. S. Oh, "Telemedicine and Dementia Care: A Systematic Review of Barriers and Facilitators," *J. Am. Med. Dir. Assoc.*, vol. 22, no. 7, pp. 1396-1402.e18, July 2021, doi: 10.1016/j.jamda.2021.03.015.

[9] I. C. Rocha *et al.*, "Monitoring Wearable Devices for Elderly People with Dementia: A Review," *Designs*, vol. 8, no. 4, Art. no. 4, Aug. 2024, doi: 10.3390/designs8040075.

[10] H. S. Kang and M. Exworthy, "Wearing the Future—Wearables to Empower Users to Take Greater Responsibility for Their Health and Care: Scoping Review," *JMIR MHealth UHealth*, vol. 10, no. 7, p. e35684, July 2022, doi: 10.2196/35684.

[11] K. Guk *et al.*, "Evolution of Wearable Devices with Real-Time Disease Monitoring for Personalized Healthcare," *Nanomater. Basel Switz.*, vol. 9, no. 6, p. 813, May 2019, doi: 10.3390/nano9060813.

[12] A. Carnevale *et al.*, "Wearable systems for shoulder kinematics assessment: a systematic review," *BMC Musculoskelet. Disord.*, vol. 20, no. 1, p. 546, Nov. 2019, doi: 10.1186/s12891-019-2930-4.

[13] D. L. Davis, "Shoulder Dysfunction and Mobility Limitation in Aging," *Adv. Geriatr. Med. Res.*, vol. 5, no. 3, p. e230008, 2023, doi: 10.20900/agmr20230008.

[14] B. Kirking, M. El-Gohary, and Y. Kwon, "The feasibility of shoulder motion tracking during activities of daily living using inertial measurement units," *Gait Posture*, vol. 49, pp. 47–53, Sept. 2016, doi: 10.1016/j.gaitpost.2016.06.008.

[15] L.-F. Lin, Y.-J. Lin, Z.-H. Lin, L.-Y. Chuang, W.-C. Hsu, and Y.-H. Lin, "Feasibility and efficacy of wearable devices for upper limb rehabilitation in patients with chronic stroke: a randomized controlled pilot study," *Eur. J. Phys. Rehabil. Med.*, vol. 54, no. 3, pp. 388–396, June 2018, doi: 10.23736/S1973-9087.17.04691-3.

[16] C. Auepanwiriyakul, S. Waibel, J. Songa, P. Bentley, and A. A. Faisal, "Accuracy and Acceptability of Wearable Motion Tracking for Inpatient Monitoring Using Smartwatches," *Sensors*, vol. 20, no. 24, p. 7313, Dec. 2020, doi: 10.3390/s20247313.

[17] S. H. Chae, Y. Kim, K.-S. Lee, and H.-S. Park, "Development and Clinical Evaluation of a Web-Based Upper Limb Home Rehabilitation System Using a Smartwatch and Machine Learning Model for Chronic Stroke Survivors: Prospective Comparative Study," *JMIR MHealth UHealth*, vol. 8, no. 7, p. e17216, July 2020, doi: 10.2196/17216.

[18] J. Kim, S.-A. Chang, and S. W. Park, "First-in-Human Study for Evaluating the Accuracy of Smart Ring Based Cuffless Blood Pressure Measurement," *J. Korean Med. Sci.*, vol. 39, no. 2, Dec. 2023, doi: 10.3346/jkms.2024.39.e18.

[19] E. Rovini *et al.*, "A wearable ring-shaped inertial system to identify action planning impairments during reach-to-grasp sequences: a pilot study," *J. NeuroEngineering Rehabil.*, vol. 18, no. 1, p. 118, July 2021, doi: 10.1186/s12984-021-00913-4.

[20] J. Qi, C. Wu, L. Yang, C. Ni, and Y. Liu, "Artificial intelligence (AI) for home support interventions in dementia: a scoping review protocol," *BMJ Open*, vol. 12, no. 9, p. e062604, Sept. 2022, doi: 10.1136/bmjopen-2022-062604.

[21] A. N. Ramesh, C. Kambhampati, J. R. T. Monson, and P. J. Drew, "Artificial intelligence in medicine," *Ann. R. Coll. Surg. Engl.*, vol. 86, no. 5, pp. 334–338, Sept. 2004, doi: 10.1308/147870804290.

[22] S. Shajari, K. Kuruvinashetti, A. Komeili, and U. Sundararaj, "The Emergence of AI-Based Wearable Sensors for Digital Health Technology: A Review," *Sensors*, vol. 23, no. 23, p. 9498, Nov. 2023, doi: 10.3390/s23239498.

[23] M. K. Das, "An Introduction to Qualitative and Mixed Methods Study Designs in Health Research," *Indian Pediatr.*, vol. 59, no. 5, pp. 416–423, May 2022.

[24] "XO Technology," XO Technology. Accessed: Mar. 12, 2025. [Online]. Available: https://www.xotechnology.ca

[25] I. Arevalo-Rodriguez *et al.*, "Mini-Mental State Examination (MMSE) for the detection of Alzheimer's disease and other dementias in people with mild cognitive impairment (MCI)," *Cochrane Database Syst. Rev.*, vol. 2015, no. 3, p. CD010783, Mar. 2015, doi: 10.1002/14651858.CD010783.pub2.

[26] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Q.*, vol. 13, no. 3, pp. 319–340, 1989, doi: 10.2307/249008.

[27] A. Spagnolli, E. Guardigli, V. Orso, A. Varotto, and L. Gamberini, "Measuring User Acceptance of Wearable Symbiotic Devices: Validation Study Across Application Scenarios," in *Symbiotic Interaction*, G. Jacucci, L. Gamberini, J. Freeman, and A. Spagnolli, Eds., Cham: Springer International Publishing, 2014, pp. 87–98. doi: 10.1007/978-3-319-13500-7_7.

[28] V. Braun and V. Clarke, "Reflecting on reflexive thematic analysis," *Qual. Res. Sport Exerc. Health*, vol. 11, no. 4, pp. 589–597, Aug. 2019, doi: 10.1080/2159676X.2019.1628806.

[29] B. C. O'Brien, I. B. Harris, T. J. Beckman, D. A. Reed, and D. A. Cook, "Standards for reporting qualitative research: a synthesis of recommendations," *Acad. Med. J. Assoc. Am. Med. Coll.*, vol. 89, no. 9, pp. 1245–1251, 2014, doi: 10.1097/ACM.0000000000000388.

[30] T. König *et al.*, "User experience and acceptance of a device assisting persons with dementia in daily life: a multicenter field study," *Aging Clin. Exp. Res.*, vol. 34, no. 4, pp. 869–879, 2022, doi: 10.1007/s40520-021-02013-8.

[31] G. M. Boulton-Lewis, L. Buys, and J. Lovie-Kitchin, "Learning and Active Aging," *Educ. Gerontol.*, vol. 32, no. 4, pp. 271–282, Apr. 2006, doi: 10.1080/03601270500494030.

[32] W. Wilkowska, J. Offermann-van Heek, T. Laurentius, L. C. Bollheimer, and M. Ziefle, "Insights Into the Older Adults' World: Concepts of Aging, Care, and Using Assistive Technology in Late Adulthood," *Front. Public Health*, vol. 9, July 2021, doi: 10.3389/fpubh.2021.653931.

[33] L. Gualtieri, J. Phillips, S. Rosenbluth, and S. Synoracki, "Digital Literacy: A Barrier to Adoption of Connected Health Technologies in Older Adults," *iProceedings*, vol. 4, no. 2, p. e11803, Sept. 2018, doi: 10.2196/11803.

[34] R. Tirado-Morueta, A. M. Duarte-Hueros, A. Infante-Moro, and J. M. Gonzalez Calleros, "Editorial: Patterns of technology-enhanced digital literacy of older adults," *Front. Educ.*, vol. 10, June 2025, doi: 10.3389/feduc.2025.1639857.

[35] C. Wu and G. G. Lim, “Investigating older adults users’ willingness to adopt wearable devices by integrating the technology acceptance model (UTAUT2) and the Technology Readiness Index theory,” *Front. Public Health*, vol. 12, Sept. 2024, doi: 10.3389/fpubh.2024.1449594.

[36] M. Veras *et al.*, “A Rapid Review of Ethical and Equity Dimensions in Telerehabilitation for Physiotherapy and Occupational Therapy,” *Int. J. Environ. Res. Public. Health*, vol. 22, no. 7, p. 1091, July 2025, doi: 10.3390/ijerph22071091.

[37] M. Veras *et al.*, “Ethics and Equity Challenges in Telerehabilitation for Older Adults: Rapid Review,” *JMIR Aging*, vol. 8, no. 1, p. e69660, Aug. 2025, doi: 10.2196/69660.

Fall Prediction in Older Adults: A Model Based on Fall-Trajectory Predictors Collected in Patients' Homes

Amadou M. Djiogomaye Ndiaye
Laboratoire Vie-Santé, UR 24 134, Faculté de Médecine
Chaire d'excellence chez Fondation Partenariale de
l'Université de Limoges I Institut Omega Health
Limoges, France
e-mail: amadou_maguette_djio.ndiaye@unilim.fr

Laurent Billonnet
XLIM-SRI - Systèmes et Réseaux Intelligents
(444315) – XLIM
Limoges, France
e-mail: laurent.billonnet@unilim.fr

Michel Harel
Laboratoire Vie-Santé, UR 24 134, Faculté de Médecine
INSPÉ de Limoges, Université de Limoges
Limoges, France
Institut de Mathématiques de Toulouse, UMR CNRS 5
219, Université Paul Sabatier
Toulouse, France
e-mail: michel.harel@unilim.fr

Achille Tchalla
Laboratoire Vie-Santé, UR 24 134, Faculté de Médecine
Service de Médecine Gériatrique, CHU de Limoges
Dupuytren
Pôle Hospitalo-Universitaire de Gérontologie clinique,
Service de Médecine Gériatrique, Unité de Prévention de
Suivi et d'Analyse du Vieillissement, CHU de Limoges
Dupuytren
Chaire d'excellence chez Fondation Partenariale de
l'Université de Limoges I Institut Omega Health
Limoges, France
e-mail: achille.tchalla@unilim.fr

***Abstract*—Falls among older adults are a major public health concern due to their frequency, consequences and impact on autonomy and mortality. The Risk Of Falling (ROF) is linked to three dimensions: physical/organic, socio-environmental and thymic/cognitive. Identifying individuals at high risk is essential to implementing personalized prevention strategies. While fall history is a well-known predictor, the integration of multi-dimensional health data and interpretable machine learning models may enhance prediction accuracy. We conducted a retrospective analysis of 1,648 older adults who underwent a Comprehensive Geriatric Assessment (CGA) at two time points. Based on clinical, functional, cognitive and psychosocial variables, we developed and compared four supervised classification models: logistic regression, Support Vector Machine (SVM), random forest and eXtreme Gradient Boosting (XGBoost). Predictive performance was evaluated using Area Under the receiver operating characteristic Curve (AUC), F1-score and Brier score. SHapley Additive exPlanations (SHAP) values were used to interpret variable contributions at the individual level. XGBoost and random forest models demonstrated the best performance (AUC = 0.76 and 0.77, F1-score = 0.72 and 0.73, Brier score = 0.19 for both). SHAP analysis confirmed that fall history was a strong predictor but not the sole contributor to the model's decisions. Functional limitations, low Activities of Daily Living (ADL) and low Instrumental Activities of Daily Living (IADL), impaired physical performance (low Short Physical Performance Battery (SPPB)), pathological Single Leg Balance (SLB) and cognitive scores (Mini-Mental State Examination (MMSE)) also played substantial roles. Misclassified cases illustrated the importance of multidimensional balance in the model's outputs. Our findings support the use of interpretable machine learning models, particularly XGBoost, for personalized fall risk prediction in older adults. Beyond fall history, a combination of physical, cognitive and psychosocial variables contributes meaningfully to risk estimation. Such models may help guide targeted preventive interventions in geriatric practice, provided operational complexity is managed to allow real-world clinical integration.**

Keywords-fall; older population; prevention; personalized medicine; AI.

I. INTRODUCTION

This article is an extended version of the international conference paper entitled *"Enhancing Fall Prediction in Older Adults: A Data-Driven Approach to Key Parameter Selection"* [1]. In this extended version, some models have been upgraded by including dyslipidemia, a cardiovascular factor, among the predictive variables for falls. However, we retain XGBoost as our final model, since it remains one of the most effective approaches for ensuring both high predictive performance and interpretability in personalized prediction.

According to the World Health Organization (WHO), older individuals are those aged ≥ 60 years. The proportion of older individuals worldwide is expected to nearly double between 2015 and 2050, increasing from 12% to 22% [2]. The National Institute of Statistics and Economic Studies (INSEE) estimates that one in three individuals in France will be aged ≥ 60 years by 2060, compared to one in four individuals in 2021 [3]. Aging leads to a gradual decline in functional capacity, increasing the ROF [4]. Falls in older adults represent a major public health concern due to their high frequency, their functional, psychological and economic consequences, as well as their impact on mortality. In the study by Tan et al. [5], falling was identified as one of the main predictive factors integrated into a model designed to identify long-term care patients at highest risk of death. Similarly, Shaik et al. [6] highlighted that, in both older and younger individuals, falls, along with bone pathologies, are among the primary causes of hip fractures.

Fall prevention has always been a central focus in medical practice, notably through clinical test batteries or by adjusting specific functions according to identified predictive factors, generally using linear regression models (LRMs), after grouping patients based on shared health characteristics. While traditional regression models have long been the standard tool for analyzing risk factors, machine learning methods now offer improved predictive performance by accounting for complex interactions between variables.

We developed predictive models using as input data the factors identified in various fall trajectories. The objective is to evaluate whether these variables are sufficiently discriminative to power an effective predictive model, among all those tested and thereby contribute to a targeted and personalized fall risk prevention strategy. Early identification of ROF facilitates the administration of personalized interventions for individuals [7].

Most recent studies predict falls using sensors or Electronic Health Records (EHRs). With data collected directly from elderly individuals' homes, our objective is to develop an effective predictive model using the fewest possible features.

In this study, we evaluated and compared several classification algorithms to predict fall risk based on clinical, functional and psychosocial data collected from a CGA. Model interpretability was ensured using SHAP values, in order to facilitate clinical understanding of the results and to precisely identify the factors that most contributed to the prediction of fall risk.

II. Materials and Methods

A. Study Design

Our study is based on a dataset collected between September 2011 and September 2023 through multiple home visits conducted by the Unit for Prevention, Monitoring and Analysis of Ageing (UPSAV – *Unité de Prévention, de Suivi et d'Analyse du Vieillissement*) at Limoges University Hospital, Limoges, France. The UPSAV team comprises nurses, geriatricians and other healthcare professionals. Each patient underwent an initial visit, followed by a second visit six months later and a third visit one year after the second. If the patient remains in the study after the third visit, subsequent visits occur annually. The study includes men and women aged 60 and older. To be eligible, participants had to meet the following criteria:

- Provide written informed consent, either personally or through a legal representative.
- Not be enrolled in a clinical trial that modifies their standard medical management.
- Not have progressive pathologies that could significantly affect short-term prognosis.
- Not reside in a long-term care unit or a nursing home.
- Be covered by social security at 100%.

B. Falls and Comprehensive Geriatric Assessment

During the Follow-up, a fall was defined as unintentionally coming to rest on the ground or other lower level not as a result of a major intrinsic event (e.g., myocardial infarction, stroke, or seizure) or an overwhelming external hazard (e.g., hit by a vehicle) [8], [9]. Each patient underwent a CGA and received a personalized care plan. The CGA is a multidimensional and standardized approach designed to enhance clinical practices in the care of older adults through a comprehensive health assessment. CGA are widely used to evaluate the physical, cognitive, social and medical factors associated with fall risk in older adults [10]. Although they provide valuable clinical information, CGAs often involve numerous variables and can be time-consuming to administer and interpret, particularly in home care settings. This highlights the growing need for efficient and scalable tools that can help prevent falls without increasing the burden on caregivers or patients.

Falls may occur repeatedly within a year. In geriatric practice, individuals who experience at least two falls within a 12-month period are classified as "fallers" [11].

A holistic fall prediction approach considers three key dimensions:

- The physical/organic dimension gathers data related to an individual's medical history and current symptoms, diagnosis of underlying health issues and treatment effectiveness.
- The thymic/cognitive dimension refers to an individual's mental, emotional and cognitive states.
- The socio-environmental dimension refers to age, gender, family and social support, housing conditions, home configuration, the presence of slippery rugs, stairs without railings, uneven surfaces and inadequate lighting.

Evaluating the ROF involves at least a gait and balance assessment of the physical/organic dimension and the age and gender of the socio-environmental dimension. Data involving the thymic/cognitive dimension allow for a comprehensive review of the potential causes of a fall. The term "dimension" refers to the types of factors that contribute to the ROF and their evaluation.

Hospitalized patients often receive incomplete health assessments across all dimensions. Our home-collected data encompass features from all three dimensions.

C. Data Collection and Variable Processing

Covariates included fall occurrences, cardiovascular risk factors, socio-environmental characteristics and the CGA summary. Fall occurrences refer to falls that occurred between visits.

Socio-environmental characteristics assessed in the home included gender, age, lifestyle, housing conditions, presence of an elevator, long-term illness status, leisure activities, social activity, human assistance and pet ownership.

Cardiovascular risk factors considered were hypertension, diabetes, dyslipidemia, obesity and tobacco use.

The CGA summary encompassed multiple functional and cognitive assessments, including:

- Verbal fluency test [12],
- Single Leg Balance (SLB) test, scored 0-60 seconds [13],
- Clock-drawing test (CDT), scored 0-5 [14],
- Instrumental Activities of Daily Living (IADL), scored 0-8 [15],
- Mini-Mental State Examination (MMSE), scored 0-30 [16],
- Mini Nutritional Assessment (MNA), scored 0-30 [17],
- Short Physical Performance Battery (SPPB), scored 0-12 [18],
- Geriatric Depression Scale (GDS), scored 0-30 [19].

For consistency, in the rest of the document, we added 'Pathological' to the feature names SLB test, CDT, Verbal Fluency and GDS to indicate whether the test result is positive or not.

D. Data analysis

In our study, the sample size decreased from 1,648 patients at the first visit to 954 patients followed up at the second visit. A descriptive analysis was conducted to provide an overview of the study variables and their distribution between individuals who had fallen and those who had not. Pearson's Chi-squared test was used for categorical variables, while the Wilcoxon rank-sum test was applied to continuous variables. The significance threshold for all statistical tests was set at a p-value (P) < 0.05 and all reported P-values were two-tailed. The p-value or probability value is a statistical measure ranging between 0 and 1. It expresses the probability of obtaining a result at least as extreme as the one observed under the assumption that the null hypothesis (H_0) is true. The null hypothesis used as the starting point of a statistical test states that there is no effect, no difference, or no relationship between the variables under study. According to the most commonly accepted convention a result is considered statistically significant when $p < 0.05$. In this case, the probability of obtaining the observed data (or more extreme outcomes) under H_0, is less than 5%. The null hypothesis is therefore rejected in favor of the alternative hypothesis (H_1), suggesting the existence of an effect or a difference. All statistical analyses were performed using R software (version 4.4.0, R Foundation for Statistical Computing, Vienna, Austria).

E. Model Development Using Supervised Machine Learning

The construction of a predictive model relies primarily on selecting a limited number of relevant variables. In geriatrics, preventive strategies implemented by geriatricians traditionally rely on "predictive factors" identified using logistic regression models (LRM). These factors correspond to variables significantly associated with fall risk across different patient groups (or clusters), formed based on longitudinal (or panel) data collected at multiple time points during the study.

In our work, after identifying the fall trajectories specific to the study population, we extracted the most explanatory variables for each of these trajectories. These predictive variables then served as the basis for building several predictive models, which we compared in order to evaluate their performance.

We developed a fall risk prediction model by selecting the best-performing algorithm among four classifiers: logistic regression, Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost) and Random Forest. Logistic regression is a linear supervised classification model particularly suited for binary problems . SVM, on the other hand, aims to maximize the margin between classes using an optimal hyperplane [20]. Ensemble models such as XGBoost and Random Forest rely on aggregating multiple decision trees: the former through a sequential boosting process and the latter through a bagging mechanism, both of which enhance model accuracy and robustness [21], [22].

To optimize the performance of each classifier, we used the RandomizedSearchCV method, which randomly explores a subset of hyperparameter combinations within a defined search space. Unlike GridSearchCV, which exhaustively evaluates all possible combinations, this approach reduces computational cost while efficiently exploring influential parameters through cross-validation. Finally, to calibrate the predicted probabilities of the models, we applied 5-fold cross-validation calibration using CalibratedClassifierCV (with cv=5) before evaluating final performance on the test set.

No missing data were observed among the variables included in the analysis. To address class imbalance, the RandomUnderSampler method was applied, which consists of randomly removing observations from the majority class to rebalance the dataset. Given the sensitive and real nature of health data, no synthetic oversampling method was used. The dataset was randomly split into a training set (70%) and a test set (30%).

Model performance was evaluated on both the training and test sets using several metrics: Area Under the Curve (AUC), accuracy, precision, recall, specificity, F1-score and Brier score [23], [24], [25]. Among these, AUC, F1-score and Brier score were selected as the main evaluation indicators. AUC assesses the model's discrimination ability, the F1-score captures the balance between precision and recall, while the Brier score measures the accuracy of probabilistic predictions; it is calculated as the mean squared difference between predicted probabilities and actual outcomes. A high AUC and F1-score, combined with a low Brier score, indicate good classification performance and accurate probability estimation.

For model interpretation, SHAP values were computed to quantify the contribution of each variable to the individual prediction of fall risk. SHAP is an explainable AI method that provides insights into the contribution of each feature both globally (across the entire dataset) and locally (for individual predictions) [26]. All algorithms were implemented in Python 3.10.16 (Python Software Foundation, Wilmington, DE). Variable preprocessing was performed using OneHotEncoder for categorical variables and StandardScaler for numerical variables, via the scikit-learn library.

III. RESULTS

A total of 1,648 individuals met the inclusion criteria for the study. Table I presents the baseline socio-environmental and health characteristics of the sample that significantly differentiate fallers from non-fallers. Among the older adults included, 1,113 (68%) were women and 535 (32%) were men. Additionally, 73% had hypertension and only 288 participants (17%) engaged in social activities. The mean age of participants was 83 ± 6 years. Regarding falls, 823 participants (approximately 50%) had experienced a fall during the previous year. Concerning housing conditions, 991 (60%) were homeowners. Furthermore, 449 participants (27%) were classified as having depression.

TABLE I. OVERVIEW OF BASELINE CHARACTERISTICS ACCORDING TO FALLS OF THE STUDY

Features of the study	Total sample (N = 1,648) n (%)	Falls of the study		p-value*
		No falls (n = 794, 48.2%)	Falls (n = 854, 51.8%)	
Woman	**1,113 (68%)**	**500 (63%)**	**613 (72%)**	**<0.001**
Age, m ± SD, years	**83 ± 6**	**82 ± 6**	**83 ± 6**	**0.001**
Diabetes	**339 (21%)**	**146 (18%)**	**193 (23%)**	**0.035**
Leisure	**1,377 (84%)**	**689 (87%)**	**688 (81%)**	**<0.001**
Social activity	**288 (17%)**	**162 (20%)**	**126 (15%)**	**0.003**
Human assistance	**1,402 (85%)**	**644 (81%)**	**758 (89%)**	**<0.001**
ADL, m ± SD	**5 ± 1**	**5 ± 1**	**5 ± 1**	**<0.001**
IADL, m ± SD	**6 ± 2**	**6 ± 2**	**5 ± 2**	**<0.001**
MMSE, m ± SD	**23 ± 7**	**24 ± 7**	**23 ± 7**	**0.006**
Pathological CDT	**585 (35%)**	**244 (31%)**	**341 (40%)**	**<0.001**
Pathological verbal fluency	**672 (41%)**	**269 (34%)**	**403 (47%)**	**<0.001**
MNA, m ± SD	**24 ± 4**	**24 ± 4**	**23 ± 4**	**<0.001**
SPPB, m ± SD	**7 ± 4**	**7 ± 4**	**6 ± 4**	**<0.001**
Pathological GDS	**449 (27%)**	**176 (22%)**	**273 (32%)**	**<0.001**
Pathological SLB	**708 (43%)**	**261 (33%)**	**447 (52%)**	**<0.001**

**Pearson's Chi-squared test; Wilcoxon rank sum test. Statistically significance (p-value < .05).*
m, mean; SD, Standard deviation; SLB, Single leg balance; CDT, Clock-drawing test; ADL, Activities of Daily Living; IADL, Instrumental Activities of Daily Living; MMSE, Mini-Mental State Examination; MNA, Mini Nutritional Assessment; SPPB, Short Physical Performance Battery; GDS, Geriatric Depression Scale.
Data are shown as the number (percentage) or mean ± SD unless otherwise indicated.

In Table II, which presents the variables included in our predictive models, it is observed that among the 954 participants included in the study, 48.6% reported at least one fall prior to the follow-up period. Fallers exhibited several characteristics that were significantly different ($p \leq 0.05$) from non-fallers. Fallers were predominantly women (74% vs 66%). Their functional and physical abilities were generally more impaired: lower ADL scores, reduced SPPB scores (6 ± 4 vs 8 ± 3) and lower IADL scores. Depression, as indicated by a pathological GDS score, was more frequent among fallers (31% vs 20%) and postural instability, assessed by a pathological one-leg stance test, was observed in 46% of fallers compared to 34% of non-fallers. Participation in leisure activities was also slightly lower among fallers (86% vs 91%), which could reflect behavioral withdrawal or functional restriction.

TABLE II. OVERVIEW OF THE SIX-MONTH INPUT FEATURES USED IN OUR PREDICTIVE MODELS

Features	Falls of the study (N = 954)			p-value*
	Total sample (N = 954) n (%)	No falls (n = 490, 51.4%)	Falls (n = 464, 48.6%)	
Woman	**664 (70%)**	**321 (66%)**	**343 (74%)**	**0.005**
Hypertension	688 (72%)	353 (72%)	335 (72%)	0.96
Dyslipidemia	453 (47%)	237 (48%)	216 (47%)	0.57
Obesity	254 (27%)	122 (25%)	132 (28%)	0.21
Leisure	**843 (88%)**	**445 (91%)**	**398 (86%)**	**0.015**
MMSE, m ± SD	25 ± 6	25 ± 6	25 ± 6	0.13
SPPB, m ± SD	**7 ± 4**	**8 ± 3**	**6 ± 4**	**<0.001**
ADL, m ± SD	**5 ± 1**	**6 ± 1**	**5 ± 1**	**<0.001**
IADL, m ± SD	**6 ± 2**	**7 ± 2**	**6 ± 2**	**<0.001**
Pathological GDS	**238 (25%)**	**96 (20%)**	**142 (31%)**	**<0.001**
Pathological SLB	**381 (40%)**	**167 (34%)**	**214 (46%)**	**<0.001**

**Pearson's Chi-squared test; Wilcoxon rank sum test. Statistically significance (p-value < .05).m,mean; SD, Standard deviation; SLB, Single leg balance; CDT, Clock-drawing test; ADL, Activities of Daily Living; IADL, Instrumental Activities of Daily Living; MMSE, Mini-Mental State Examination; MNA, Mini Nutritional Assessment; SPPB, Short Physical Performance Battery; GDS, Geriatric Depression Scale.*
Data are shown as the number (percentage) or mean ± SD unless otherwise indicated.

In contrast, hypertension, dyslipidemia, obesity and MMSE scores were not statistically associated with falls in this cohort.

These results support the hypothesis of a multifactorial etiology of falls, primarily driven by physical function impairment, loss of autonomy, mood disorders, depression and postural balance issues.

A comparison of Table I and Table II shows that variables such as hypertension, obesity and dyslipidemia are predictive factors of falls but do not significantly differentiate fallers from non-fallers. The remaining variables reported in Table II are also significant in Table I.

Fig. 1 presents the AUC of the four models evaluated for predicting fall risk, namely logistic regression, SVM, XGBoost and random forest.

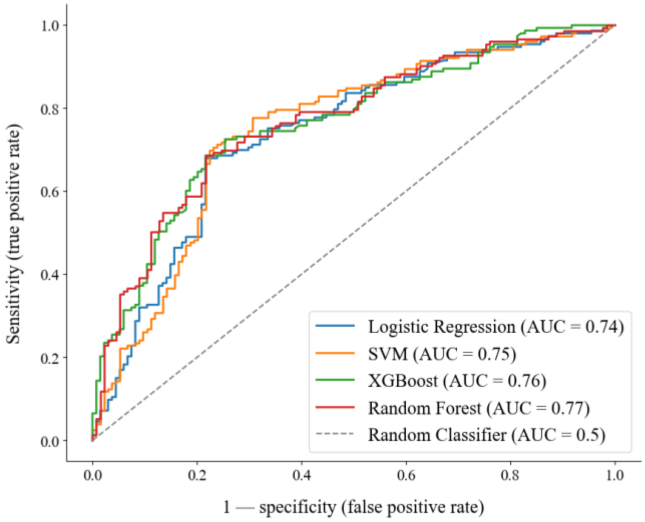


Figure 1. Area Under the Curve (AUC) of the Different Models

Table III reports the performance metrics of the different models. All models achieved an identical precision of 0.78, with balanced F1-scores ranging between 0.71 and 0.73, indicating comparable overall classification performance.

TABLE III. SUMMARY OF PREDICTIVE PERFORMANCE OF THE DIFFERENT MODELS

Metrics	Logistic Regression	SVM	XGBoost	Random Forest
AUC	0.74	0.75	**0.76**	**0.77**
Accuracy	0.73	0.71	0.72	0.73
Precision	0.78	0.78	0.78	0.78
Recall	0.68	0.65	0.67	0.68
Specificity	0.78	0.78	0.78	0.78
F1 score	0.73	0.71	**0.72**	**0.73**
Brier score	0.20	0.20	**0.19**	**0.19**

However, XGBoost and Random Forest show better areas under the ROC curve, with AUC values of 0.76 and 0.77 respectively (see Fig. 1), suggesting higher discriminative ability compared to logistic regression (AUC = 0.74) or SVM (AUC = 0.75). Recall is slightly lower for XGBoost (0.67) than for Random Forest (0.68), which may reflect a tendency to under-detect certain fall cases. Finally, the lowest Brier scores (0.19) are achieved by XGBoost and Random Forest, indicating better probabilistic calibration of predictions. Thus, although all models perform similarly in classification, Random Forest appears to offer the best trade-off between discrimination and calibration.

XGBoost and Random Forest are the models with the best overall performance. Both are tree-based methods; while Random Forest makes binary decisions, XGBoost has the advantage of computing individualized probabilities, which makes it more suitable for personalized care approaches. To better understand the contribution of each variable to the model's predictions, we apply SHAP to XGBoost.

The analysis of SHAP values presented in Fig. 2 highlights both the relative importance and the direction of effect of each variable in predicting fall risk within the XGBoost model.

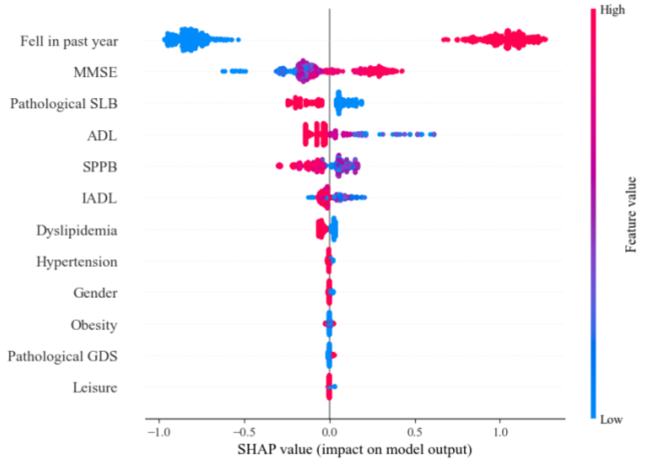


Figure 2. Impact of the Different Variables on the Best Model (XGBoost)

The use of SHAP values provides transparent model interpretation and may help inform priorities for targeted preventive strategies. A low score (values in blue) contributes significantly to risk reduction, whereas a high score (shown in red) is associated with increased predicted risk.

Among all the variables considered, fall history emerges as the most influential factor thereby confirming the strong predictive power of prior fall events. Physical performance, as assessed by the SPPB score also plays a central role in fall-risk prediction, low SPPB values (indicating physical impairment) are strongly associated with higher risk. Pathological single-leg stance reflecting balance impairments does not appear to be correlated with elevated fall risk ; in some cases, it may even be linked to severely limited mobility thereby reducing exposure to risk through restricted movement.

At the cognitive level, the MMSE score shows a more nuanced relationship while low scores are generally considered a risk factor, their impact appears less pronounced in the model. Conversely, higher scores may counterintuitively be associated with increased risk possibly due to overconfidence or engagement in unsafe physical activities.

The ADL and IADL scores indicators of functional autonomy exhibit patterns consistent with clinical evidence reduced functional capacity is generally associated with increased fall risk. However, very low IADL scores may not strongly correlate with higher risk suggesting that advanced dependency could reduce exposure to hazardous situations.

Dyslipidemia reflecting cardiovascular impairment is unexpectedly associated with a lower risk of falls potentially indicating a tendency to avoid physical activity due to fear of falling.

Other variables including hypertension, obesity, gender, and the presence of depression (pathological GDS), exert a more moderate or marginal influence on model predictions. Participation in leisure activities shows a modest protective effect, although its overall contribution to fall-risk prediction remains limited.

In summary, this analysis underscores that the most influential predictors of fall risk are functional and physical domains, while cognitive and psychosocial dimensions exert secondary effects.

After examining the impact of each variable on the model's predictions, we now turn to some examples of personalized predictions.

The personalized predictions will be evaluated using the final selected XGBoost model. XGBoost is a gradient boosting ensemble algorithm that aggregates multiple weak decision trees to produce a high-performing predictive model [22]. In binary classification, it generates a raw output in log-odds, which is then transformed by the logistic function to obtain a probability. The log-odds (logarithm of the odds) is a way to transform a probability into a value that can range from –∞ to +∞. The raw output value of XGBoost is the weighted sum of the decision trees:

$$f(x) = \sum_{k=1}^{K} T_k(x)$$

where:

- $T_k(x)$ is the output of the k-th tree for the observation,
- K is the total number of trees,
- $f(x)$ is the raw model output, expressed in log-odds.

We then transform the raw output $f(x)$ into a probability $p(x)$ with the sigmoid function :

$$p(x) = \sigma(f(x)) = \frac{1}{1 + e^{-f(x)}}$$

where the sigmoid function is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

The reference value (base value) is the mean of $f(x)$ and the associated probability is the overall prevalence in the training sample.

Fig. 3 below illustrates a correctly predicted low fall risk ($f(x)$ = 0.23) compared with the base value of 0.48. Protective factors such as a lower MMSE score of 18, preserved ADL of 6, and a high IADL of 8 strongly contributed to reducing the predicted risk. A history of falls also contributed to lowering the prediction. Although risk-increasing variables such as the absence of a pathological one-leg stance and a low SPPB score of 4 were present, they were outweighed by the protective factors.

Fig. 4 below illustrates a case classified as high fall risk, with a predicted probability of 0.65. However, the prediction is incorrect; in the collected data, the patient did not fall. Several strong risk factors were present including a history of falls, dyslipidemia, low IADL (6) and a low SPPB score (9) all of which contributed to increasing the predicted risk. Nevertheless, these were insufficiently weighted by the model while mitigating factors such as a relatively high MMSE score (23), a non-pathological SLB and a moderate ADL score (6) overly influenced the output leading to a misclassification. This highlights the model's limitation in edge cases where compensatory features may mask critical risks.

These personalized predictions of two different patients highlight that the model's outputs do not depend solely on fall history, even though it is the strongest predictor among all variables (Fig. 2). The trends observed in the SHAP values of all variables in Fig. 2 are confirmed by the prediction shown in Fig. 3.

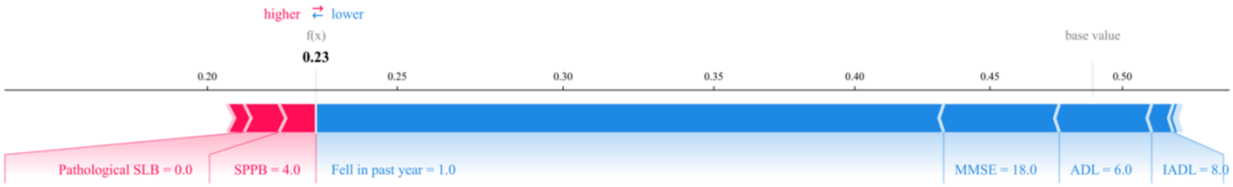


Figure 3. Correct Prediction with SHAP

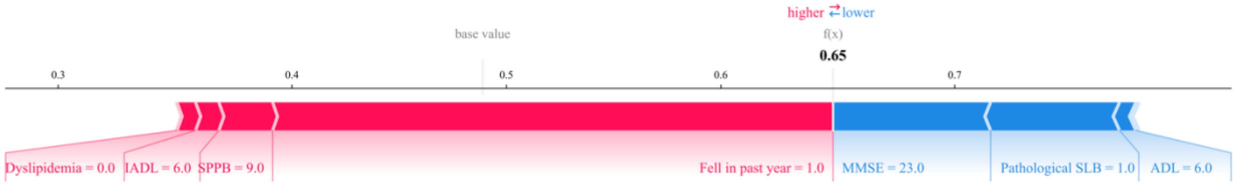


Figure 4. Incorrect Prediction with SHAP

IV. DISCUSSION

The evaluation of various fall risk predictors (see Table II), based on data from patients who completed both the first and second visits, revealed that most variables showed significant differences between fallers and non-fallers. This highlights the importance of identifying predictive factors within the least stable clusters (i.e., those in which falls were observed), as opposed to more stable clusters. Among the variables analyzed, the following were significantly different depending on group membership (fallers vs. non-fallers): gender (female), ADL score, IADL score, SPPB score, presence of a pathological GDS score, pathological SLB and participation in leisure activities.

Among these variables, only sex and participation in leisure activities pertain to the socio-environmental domain and could be collected in other protocols. The remaining variables are scores derived from the CGA conducted at the patients' homes. These findings support the hypothesis that a holistic approach is necessary for predicting fall risk. Specifically, the pathological GDS score reflects the thymic/cognitive dimension, while the ADL, IADL and SPPB scores, along with the pathological one-leg stance, reflect the physical/organic dimension.

Using the variables most significantly associated with fall risk (see Table II) as input data represents a relevant strategy, as the model’s objective is to differentiate fallers from non-fallers in a personalized manner. In order to remain aligned with the clinical approach of identifying predictive factors to develop targeted prevention plans, all variables identified (see Table II) were retained for model training. Fig. 2 confirms the importance of these variables, showing that they rank among the most influential in the XGBoost model, with the exception of gender and pathological GDS score, which were replaced by dyslipidemia and MMSE score in terms of predictive weight. The integration of dyslipidemia, a cardiovascular risk factor and the MMSE score, a marker of cognitive function, further reinforces the model’s holistic approach.

Not every feature within the three ROF dimensions is a predictive factor for falls. The effectiveness of a predictive factor depends on its statistical significance, correlation with fall occurrences and its interaction with other variables across the physical/organic, socio-environmental and cognitive dimensions. In some studies, the identified predictive variables did not encompass all three dimensions of ROF. Kawazoe et al. [27], Ikeda et al. [28] and Cella et al. [29] demonstrated that age category related to socio-environmental was a predictor of falls, suggesting a strong association between age and falls. Bath et al. [30] found that the predictive variables related to the socio-environmental dimension are diverse and varied, contributing to effective prevention. In fact, a higher number of variables related to gait and balance is associated with a more robust predictive model for falls.

In the literature review conducted by Rubenstein, only cognitive impairment was identified as a predictive variable related to the thymic/cognitive [31]. Conversely, Ikeda et al. [28], Kawazoe et al. [27] and Bath et al. [30] identified at least two predictive variables involving the thymic/cognitive dimension, providing a better understanding of the ROF associated with the thymic/cognitive dimension and facilitating preventive measures. In those features, we can find fear of falling, depressive symptoms, self-rated health, impaired consciousness and dementia at admission. Recent studies by Ikeda et al. [28] and Kawazoe et al. [27] achieved Area Under the receiver operating characteristic Curve (AUC) scores of 88% and 85%, respectively, using comprehensive approaches. Ikeda et al. [28] employed a Random Forest-based Boruta algorithm for feature selection, while Kawazoe et al. [27] used a combination of Bidirectional Encoders and Bidirectional Long Short-Term Memory (BiLSTM) networks to process sequential data. These AUC scores indicate strong model performance, reflecting high discriminative ability in classification tasks [25].

Pennone et al. [32] highlighted the difficulty in predicting fall risk among older adults with low levels of daily activity, emphasizing the importance of measuring such activity using standardized indicators. In our predictive model, we included ADL and IADL scores, which are already well-established in the literature as robust predictive factors [33], [34], [35]. A history of falling, which by definition places an older adult at risk of recurrent falls has consistently been identified as a major predictor in recent studies when collected. It is also consistently ranked among the most influential variables in predictive fall models [28], [29], [36], [37]. The cognitive dimension represented here by the MMSE score has also been widely recognized in prior research as an important determinant of fall risk [38], [39], [40]. In addition, Bharadwaz et al. [41] emphasized the influence of depression and sleep disorders on fall risk. Although the pathological GDS score was not among the most influential variables in our final model, it remains relevant when analyzing trajectories. As for sleep disturbances, while not directly measured their impact likely manifests indirectly through reduced performance in activities of daily living further justifying the inclusion of ADL and IADL scores in our predictive approach.

Pathological SLB, combined with the SPPB score, which evaluates gait and balance ability, emerged as one of the strongest determinants in predicting fall risk. Several studies have confirmed that these variables reflecting the physical and organic dimension are essential fall predictors [36], [42], [43], [44]. In the work of Lathouwers et al. [45], it was also shown that maintaining physical, mental, or social activity significantly reduces the probability of falling in older adults, a finding that aligns with our own results.

Indeed, Landers et al. [46] demonstrated that such activities help prevent the onset of fear of falling (FOF) and contribute to maintaining a high level of confidence in one’s balance abilities as measured by the Activities-specific Balance Confidence (ABC) scale, both identified as major risk factors. Similarly, Schumann et al. [47] recently highlighted the role of FOF as a predictor of falling.

The only variable present in our model that is notably absent in recent studies is dyslipidemia, a cardiovascular risk factor. This discrepancy may be explained by the methodological specificity of our study, which was based on data collected directly from patients in their homes, allowing for a more integrative assessment of overall health. The inclusion of dyslipidemia in our model underscores the importance of considering cardiovascular risk as a potential contributor to falls, especially when falls occur suddenly and without prior functional warning signs.

While fall history is consistently identified as one of the most influential predictors of future falls, our analysis shows that the model does not rely exclusively on this variable to make its predictions (Fig. 3 and Fig. 4). SHAP value interpretation reveals that the XGBoost model incorporates a wide range of factors, including physical performance, functional autonomy, cognitive status and psychosocial indicators, when estimating fall risk.

In several correctly classified cases, the presence of a prior fall is counterbalanced by protective factors such as high ADL and IADL scores, preserved cognitive function (as indicated by MMSE) and non-pathological balance performance (e.g., SPPB score or SLB). This demonstrates that the model takes into account the complex interplay between risk and protective variables rather than basing its prediction on fall history alone.

Inversely, certain misclassified cases highlight that a history of falls does not always lead to a high-risk prediction. When other variables present a favorable profile, the model may underestimate the actual risk, suggesting that fall history while important is insufficient on its own to ensure predictive accuracy.

Moreover, the model's use of additional variables such as dyslipidemia and cognitive scores reflects a broader more integrative view of fall risk. These results confirm the necessity of a multidimensional approach and support the implementation of interpretable machine learning models that can provide individualized, clinically meaningful insights beyond any single predictor.

This study confirms the relevance of machine learning models, particularly XGBoost for predicting fall risk in older adults with good discriminative performance and calibration. The analysis of SHAP values enabled a transparent and clinically meaningful ranking of predictive factors. Fall history, impairments in physical performance (SPPB, one-leg stance) and functional limitations (ADL, IADL) emerged as the main determinants. Cognitive and psychosocial factors play a secondary yet non-negligible role. These findings highlight the importance of a multidimensional assessment that incorporates interpretable technological tools to guide personalized prevention strategies. The integration of such approaches into geriatric practice could enhance early identification of at-risk patients and contribute to reducing the incidence of falls.

Nonetheless, our work presents several limitations. First, although the XGBoost model demonstrated good performance (AUC of 0.76, Brier score of 0.19, precision of 0.78), its implementation in clinical practice could be hindered by the time required to perform the assessments, even though the number of variables that significantly influence predictions is relatively low. This complexity may limit its use by healthcare professionals in care settings where workload and time constraints are critical factors. A clinical arbitration process aimed at identifying substitutable or priority variables could facilitate the operational integration of the model.

Moreover, the model was built using all variables identified as predictive, without applying a selection procedure based solely on significant differences between fallers and non-fallers. Such a selection approach might optimize the trade-off between predictive performance and ease of use.

From a methodological standpoint, the study did not include a control group. A randomized design comparing a control group (receiving no care) and an intervention group (receiving personalized follow-up) would have allowed for a more detailed analysis of the impact of care on the dynamics of fall risk factors and would have helped to better identify common or distinguishing predictive variables between the two groups.

Finally, the data used were exclusively collected from patients in France. This geographical limitation restricts the generalizability of the findings to other cultural and socio-environmental contexts. Since falls are a multifactorial phenomenon strongly influenced by lifestyle, home environment and care practices, significant variations may exist in other countries. In particular, the socio-environmental dimension deserves to be examined through a multicenter international approach.

Overall, while our model is grounded in a realistic approach aimed at clinical integration, these limitations open avenues for improvement in both methodological robustness and the transferability of results.

V. CONCLUSION

This study contributes to advancing fall prevention by leveraging a 12-year dataset collected in home settings to develop an AI-based predictive model. Our approach integrates the three dimensions of ROF, optimizing model performance while reducing the number of required input features.

By applying explainable AI techniques, we identified the contribution of each feature to fall risk, thereby supporting the development of more targeted and effective intervention strategies. These findings may help enhance the quality of elderly care by informing personalized prevention efforts and guiding future research in geriatric risk assessment.

As with most AI models, ours can be continuously refined with additional data over time. In our case, improving the model also provides an opportunity to collect data from patients' homes while offering them personalized fall prevention advice. During the intervals between practitioner visits, necessary adjustments to home configurations can also be made if needed.

The clinical utility of the final model could be explored in future studies using Decision Curve Analysis (DCA). This method helps identify the clinical range in which the model provides a net benefit, thereby allowing practitioners to determine the optimal threshold for patient management while taking available resources into account.

Acknowledgment

AUTHOR CONTRIBUTIONS

Amadou M. Djiogomaye Ndiaye: Contributed to the study concept and design, drafted the manuscript, conducted statistical analyses, built the models, interpreted the data and approved the final version for publication.

Michel Harel: Contributed to the study concept and design, critically revised the manuscript and approved the final version for publication.

Laurent Billonnet: Contributed to the study concept and design, critically revised the manuscript and approved the final version for publication.

Achille Tchalla: Contributed to the study concept and design, analyzed and interpreted the data, critically revised the manuscript and approved the final version for publication.

SPONSOR'S ROLE

Data grant by *"CARSAT Centre-Ouest",* Limoges, France were collected by UPSAV and URCI. The analysis and manuscript preparation were funded by a grant from chair of excellence in Artificial Intelligence and Healthy Ageing in the Nouvelle-Aquitaine Region (IABV-NA – *Intelligence Artificielle et Bien-Vieillir en Nouvelle-Aquitaine), Limoges, France.*

References

[1] A. M. D. Ndiaye, M. Harel, L. Billonnet, and A. Tchalla, "Enhancing Fall Prediction in Older Adults: A Data-Driven Approach to Key Parameter Selection," presented at the eTELEMED 2025, The Seventeenth International Conference on eHealth, Telemedicine, and Social Medicine, May 2025, pp. 37–39. Accessed: Sept. 14, 2025. [Online]. Available: https://www.thinkmind.org/library/eTELEMED/eTELEMED_2025/etelemed_2025_1_80_40047.html

[2] WHO, "Ageing and health." Accessed: Mar. 28, 2023. [Online]. Available: https://www.who.int/fr/news-room/fact-sheets/detail/ageing-and-health

[3] Insee, "Life Expectancy at Various Ages | Insee." Accessed: Mar. 27, 2023. [Online]. Available: https://www.insee.fr/fr/statistiques/2416631#graphique-figure1

[4] X. Thierry, "Accidents and physical assaults among the elderly: less frequent than among the young, but more severe," Population & Sociétés, vol. 468, no. 6, pp. 1–4, 2010, doi: 10.3917/popsoc.468.0001.

[5] H.-C. Tan et al., "Deep learning model for the prediction of all-cause mortality among long term care people in China: a prospective cohort study," Sci Rep, vol. 14, no. 1, p. 14639, June 2024, doi: 10.1038/s41598-024-65601-4.

[6] A. Shaik et al., "A Staged Approach using Machine Learning and Uncertainty Quantification to Predict the Risk of Hip Fracture," ArXiv, p. arXiv:2405.20071v1, May 2024.

[7] US Preventive Services Task Force, "Interventions to Prevent Falls in Community-Dwelling Older Adults: US Preventive Services Task Force Recommendation Statement," JAMA, vol. 319, no. 16, pp. 1696–1704, Apr. 2018, doi: 10.1001/jama.2018.3097.

[8] Kellogg, "The prevention of falls in later life. A report of the Kellogg International Work Group on the Prevention of Falls by the Elderly," Dan Med Bull, vol. 34 Suppl 4, pp. 1–24, Apr. 1987.

[9] S. L. Wolf, H. X. Barnhart, N. G. Kutner, E. McNeely, C. Coogler, and T. Xu, "Reducing frailty and falls in older persons: an investigation of Tai Chi and computerized balance training. Atlanta FICSIT Group. Frailty and Injuries: Cooperative Studies of Intervention Techniques," Journal of the American Geriatrics Society, vol. 44, no. 5, May 1996, doi: 10.1111/j.1532-5415.1996.tb01432.x.

[10] A. Pilotto et al., "Three Decades of Comprehensive Geriatric Assessment: Evidence Coming From Different Healthcare Settings and Specific Clinical Conditions," Journal of the American Medical Directors Association, vol. 18, no. 2, p. 192.e1-192.e11, Feb. 2017, doi: 10.1016/j.jamda.2016.11.004.

[11] M. E. Tinetti and M. Speechley, "Prevention of Falls among the Elderly," http://dx.doi.org/10.1056/NEJM198904203201606. Accessed: Apr. 24, 2023. [Online]. Available: https://www.nejm.org/doi/pdf/10.1056/NEJM198904203201606

[12] A. L. Benton, "Development of a multilingual aphasia battery: Progress and problems," Journal of the Neurological Sciences, vol. 9, no. 1, pp. 39–48, July 1969, doi: 10.1016/0022-510X(69)90057-4.

[13] T. H. Trojian and D. B. McKeag, "Single leg balance test to identify risk of ankle sprains," July 2006, doi: 10.1136/bjsm.2005.024356.

[14] M. Freedman, L. Leech, E. Kaplan, G. Winocur, K. Shulman, and D. Delis, "M. Freedman, L. Leech, E. Kaplan, G. Winocur, K. Shulman and D. Delis. Clock drawing: a neuropsychological analysis. New York: Oxford Press, 1994. | Request PDF," ResearchGate, 1994, doi: 10.1017/S0714980800013398.

[15] M. P. Lawton and E. M. Brody, "Assessment of Older People: Self-Maintaining and Instrumental Activities of Daily Living1," The Gerontologist, vol. 9, no. 3_Part_1, pp. 179–186, Oct. 1969, doi: 10.1093/geront/9.3_Part_1.179.

[16] J. I. Escobar, A. Burnam, M. Karno, A. Forsythe, J. Landsverk, and J. M. Golding, "Use of the Mini-Mental State Examination (MMSE) in a community population of mixed ethnicity. Cultural and linguistic artifacts," J Nerv Ment Dis, vol. 174, no. 10, pp. 607–614, Oct. 1986, doi: 10.1097/00005053-198610000-00005.

[17] B. Vellas et al., "The Mini Nutritional Assessment (MNA) and its use in grading the nutritional state of elderly patients," Nutrition, vol. 15, no. 2, pp. 116–122, Feb. 1999, doi: 10.1016/s0899-9007(98)00171-3.

[18] J. M. Guralnik et al., "A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission," J Gerontol, vol. 49, no. 2, pp. M85-94, Mar. 1994, doi: 10.1093/geronj/49.2.m85.

[19] J. A. Yesavage et al., "Development and validation of a geriatric depression screening scale: A preliminary report," Journal of Psychiatric Research, vol. 17, no. 1, pp. 37–49, Jan. 1982, doi: 10.1016/0022-3956(82)90033-4.

[20] W. S. Noble, “What is a support vector machine?,” Nat Biotechnol, vol. 24, no. 12, pp. 1565–1567, Dec. 2006, doi: 10.1038/nbt1206-1565.

[21] L. Breiman, Classification and Regression Trees. New York: Routledge, 2017. doi: 10.1201/9781315139470.

[22] J. Brownlee, XGBoost With Python: Gradient Boosted Trees with XGBoost and scikit-learn. Machine Learning Mastery, 2016.

[23] T. Schlosser, M. Friedrich, T. Meyer, and D. Kowerko, “A Consolidated Overview of Evaluation and Performance Metrics for Machine Learning and Computer Vision,” ResearchGate. Accessed: May 28, 2025. [Online]. Available: https://www.researchgate.net/publication/374558675_A_Consolidated_Overview_of_Evaluation_and_Performance_Metrics_for_Machine_Learning_and_Computer_Vision

[24] L. Hoessly, “On misconceptions about the Brier score in binary prediction models,” Apr. 23, 2025, arXiv: arXiv:2504.04906. doi: 10.48550/arXiv.2504.04906.

[25] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” Pattern Recognition, vol. 30, no. 7, pp. 1145–1159, July 1997, doi: 10.1016/S0031-3203(96)00142-2.

[26] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Accessed: May 14, 2025. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[27] Y. Kawazoe, K. Shimamoto, D. Shibata, E. Shinohara, H. Kawaguchi, and T. Yamamoto, “Impact of a Clinical Text-Based Fall Prediction Model on Preventing Extended Hospital Stays for Elderly Inpatients: Model Development and Performance Evaluation,” JMIR Med Inform, vol. 10, no. 7, p. e37913, July 2022, doi: 10.2196/37913.

[28] T. Ikeda et al., “An Interpretable Machine Learning Approach to Predict Fall Risk Among Community-Dwelling Older Adults: a Three-Year Longitudinal Study,” J GEN INTERN MED, vol. 37, no. 11, pp. 2727–2735, Aug. 2022, doi: 10.1007/s11606-022-07394-8.

[29] A. Cella et al., “Development and validation of a robotic multifactorial fall-risk predictive model: A one-year prospective study in community-dwelling older adults,” PLOS ONE, vol. 15, no. 6, p. e0234904, juin 2020, doi: 10.1371/journal.pone.0234904.

[30] P. A. Bath, N. Pendleton, K. Morgan, J. E. Clague, M. A. Horan, and S. B. Lucas, “New approach to risk determination: development of risk profile fornew falls among community-dwelling older people by use of a GeneticAlgorithm Neural Network (GANN),” The Journals of Gerontology: Series A, vol. 55, no. 1, pp. M17–M21, Jan. 2000, doi: 10.1093/gerona/55.1.M17.

[31] L. Z. Rubenstein, “Falls in older people: epidemiology, risk factors and strategies for prevention,” Age and Ageing, vol. 35, no. suppl_2, pp. ii37–ii41, Sept. 2006, doi: 10.1093/ageing/afl084.

[32] J. Pennone, N. F. Aguero, D. M. Martini, L. Mochizuki, and A. A. do Passo Suaide, “Fall prediction in a quiet standing balance test via machine learning: Is it possible?,” PLoS One, vol. 19, no. 4, p. e0296355, 2024, doi: 10.1371/journal.pone.0296355.

[33] W.-M. Chu et al., “A model for predicting fall risks of hospitalized elderly in Taiwan-A machine learning approach based on both electronic health records and comprehensive geriatric assessment,” Front. Med., vol. 9, 2022, doi: 10.3389/fmed.2022.937216.

[34] C.-W. Kang, Z.-K. Yan, J.-L. Tian, X.-B. Pu, and L.-X. Wu, “Constructing a fall risk prediction model for hospitalized patients using machine learning,” BMC Public Health, vol. 25, no. 1, p. 242, Jan. 2025, doi: 10.1186/s12889-025-21284-8.

[35] A. K. Mishra et al., “Explainable Fall Risk Prediction in Older Adults Using Gait and Geriatric Assessments,” Front. Digit. Health, vol. 4, 2022, doi: 10.3389/fdgth.2022.869812.

[36] S. Chen et al., “Comparing interpretable machine learning models for fall risk in middle-aged and older adults with and without pain,” Sci Rep, vol. 15, no. 1, p. 17032, May 2025, doi: 10.1038/s41598-025-01651-6.

[37] L. Lin, X. Liu, C. Cai, Y. Zheng, D. Li, and G. Hu, “Urban-rural disparities in fall risk among older Chinese adults: insights from machine learning-based predictive models,” Front Public Health, vol. 13, p. 1597853, 2025, doi: 10.3389/fpubh.2025.1597853.

[38] O. Beauchet et al., “Falls Risk Prediction for Older Inpatients in Acute Care Medical Wards: Is There an Interest to Combine an Early Nurse Assessment and the Artificial Neural Network Analysis?,” J Nutr Health Aging, vol. 22, no. 1, pp. 131–137, 2018, doi: 10.1007/s12603-017-0950-z.

[39] T. Ikeda et al., “An Interpretable Machine Learning Approach to Predict Fall Risk Among Community-Dwelling Older Adults: a Three-Year Longitudinal Study,” J Gen Intern Med, vol. 37, no. 11, pp. 2727–2735, Aug. 2022, doi: 10.1007/s11606-022-07394-8.

[40] A. Kabeshova et al., “Falling in the elderly: Do statistical models matter for performance criteria of fall prediction? Results from two large population-based studies,” Eur. J. Intern. Med., vol. 27, pp. 48–56, 2016, doi: 10.1016/j.ejim.2015.11.019.

[41] M. P. Bharadwaz, J. Kalita, A. Mitro, and A. Aditi, “Utilizing machine learning to identify fall predictors in India’s aging population: findings from the LASI,” BMC Geriatr, vol. 25, no. 1, p. 181, Mar. 2025, doi: 10.1186/s12877-025-05813-z.

[42] G. Cuaya-Simbro, A.-I. Perez-Sanpablo, A. Munõz-Meléndez, I. Q. Uriostegui, E.-F. Morales-Manzanares, and L. Nuñez-Carrera, “Comparison of Machine Learning Models to Predict Risk of Falling in Osteoporosis Elderly,” Found. Comput. Decis. Sci., vol. 45, no. 2, pp. 66–77, 2020, doi: 10.2478/fcds-2020-0005.

[43] T. Deschamps, C. G. Le Goff, G. Berrut, C. Cornu, and J.-B. Mignardot, “A decision model to predict the risk of the first fall onset,” Experimental Gerontology, vol. 81, pp. 51–55, Aug. 2016, doi: 10.1016/j.exger.2016.04.016.

[44] A. K. Mishra et al., “Explainable Fall Risk Prediction in Older Adults Using Gait and Geriatric Assessments,” Front Digit Health, vol. 4, p. 869812, May 2022, doi: 10.3389/fdgth.2022.869812.

[45] E. Lathouwers et al., “Characterizing fall risk factors in Belgian older adults through machine learning: a data-driven approach,” BMC Public Health, vol. 22, no. 1, p. 2210, Nov. 2022, doi: 10.1186/s12889-022-14694-5.

[46] M. R. Landers, S. Oscar, J. Sasaoka, and K. Vaughn, “Balance confidence and fear of falling avoidance behavior are most predictive of falling in older adults: prospective analysis,” Physical therapy, vol. 96, no. 4, pp. 433–442, 2016, Accessed: June 17, 2024. [Online]. Available: https://academic.oup.com/ptj/article-abstract/96/4/433/2686463

[47] P. Schumann et al., “Using machine learning algorithms to detect fear of falling in people with multiple sclerosis in standardized gait analysis,” Mult Scler Relat Disord, vol. 88, p. 105721, Aug. 2024, doi: 10.1016/j.msard.2024.105721.

LEMIP-Net: A Longitudinal Exposure-Aware Deep Learning Framework for Predicting Mental-Health Impact from Pandemic-Related Social Media Content

Mental Health Prediction

1st Sujata Patil
Department of Electronics and Communication Engineering
KLE Technological University
Hubballi, India
sujata.patil@kletech.ac.in

2nd Les Sztandera
Department of Computer Information Systems
Thomas Jefferson University
Philadelphia, PA 19107, USA
les.sztandera@jefferson.edu

3rd Hemalatha K. L
Department of Computer Science and Engineering,
Sri Krishna Institute of Technology
Bengaluru, India
hemalathaklise@skit.org.in

Abstract—**In this era, the rapid increase of health and pandemic-related information on social-media platforms has led to diverse issues that includes public anxiety, stress and behavioral volatility. Even though, prior researchers have explored various techniques still there are challenges due to the dependency on post-level analysis using static embeddings that limits exposure intensity, temporal progression and longitudinal psychological patterns. Therefore, to resolve these challenges a Longitudinal Exposure-Aware Mental-Impact Prediction Network namely LEMIP-Net is proposed, which models evolving emotional and linguistic behavior of users interacting with pandemic discourse. Initially, raw user-generated social-media posts are restructured into temporal sequences that enables observation of gradual psychological drift. Subsequently, individual post is semantically encoded by incorporating Cross-lingual Language Model with Robustly Optimized Bidirectional Encoder Representations from Transformers approach (XLM-RoBERTa). Consequently, exposure-classification module estimates probability of every post related to health or pandemic discourse, and these probabilities are aggregated to quantify user-specific exposure intensity. Further, a hybrid weak-supervision strategy refines mental-health labels through sparse self-reports and lexicon-based cues. Finally, the fused sequences are processed using a Transformer-Bidirectional Long Short Term Memory based architecture to capture global behavioral trends and short-term emotional shifts, which performs multi-task prediction of mental-health class, severity and deterioration risk. Hence, the experimental results illustrate that the proposed LEMIP-Net significantly outperforms state-of-the-art models, achieving robust and generalizable mental-health prediction by jointly modeling longitudinal behavior and exposure intensity.**

Keywords-cross-lingual language model; exposure modelling; mental-health prediction; pandemic-related social media; temporal modelling.

I. INTRODUCTION

In recent years, expanding of social media plays the massive impact on human's mental and physical health especially students. Although social media provides information of politics, education and Information Technology (IT), still impacts the human's self-esteem, mental health and sleep patterns [1]. In particular, social platforms such as twitter, Facebook and Instagram allow the users to share their posts, thoughts and ideas. In addition, studies revealed that link between negative consequences in social media does increase the stress, depression and anxiety [2]. Moreover, due to lack of offline communication, the depressed people have negative thoughts, low confidence and ambiguous issues [3]. However, using the social media negatively, impact the user's health issues such as disturbance of sleep, guilt feelings, difficulty in concentrating and suicidal thoughts [4]. Moreover, number of patients has been increasing every year with mental problems due to problems of social media and people who are already suffering with mental issues or physiological orders, will face more difficulties [5]. Also, there was a survey, where social media created development of fear and panic among the people and also females were affected mentally more than males in content of social media. [6]. To overcome these problems, early detection of stress, depression could prevent the mental health issues. In particular, the computers have the ability to express and recognize the emotions assists give better feedback to the users [7]. Further, sentiment analysis examines the people emotions, feelings, mood and attitude and one of the active types of research area in Natural Language Processing (NLP) [8]. Moreover, detecting the depression in posts has achieved important advancement in identifying the depression from social media posts. Further researchers analyzed the social media data to extract the valuable patterns and insights that related to the mental health problems. Further, by analyzing the huge information on social media, researches understand about mental health problems of users [9]. State of the art methods include Convolutional Neural Network (CNN), Transformer and Bidirectional Long Short-Term Memory (Bi-LSTM) performed the strategies to detect the depression. However, these models have huge training time and transformers models was not able to captured the important content that effect the accuracy [10]. Further, Long Short-Term Memory (LSTM) models have been utilized to examine sequential text data from social media platforms such as Twitter, where they learn contextual feature representations from documents, paragraphs and sentences. However, LSTM struggles with long term dependencies [11] and moreover, NLP techniques used for mental status of a person based on writing or speech and predicting the depression. However, most NLP techniques does not appreciate the variability [12] of depression.

The key contributions of the research are as follows:

- A Longitudinal and Exposure-Aware Mental-Health Prediction Framework (LEMIP-Net) is proposed, which jointly captures user's temporal linguistic behavior and their cumulative exposure to health and pandemic-related content. Therefore, by transforming raw social media posts into structured behavioral timelines, the proposed LEMIP-Net model allows clinically aligned assessment of mental-health risk.
- A domain-specific exposure classification and intensity modelling mechanism is employed that quantifies the likelihood and intensity of health-related content encountered by individual user, which provides an essential dimension for understanding the impact in psychological outcomes.
- The proposed LEMIP-Net improves inherently noisy user-reported mental-health labels by combining linguistic symptom markers, contextual cues and rule-based heuristics. Thus, this hybrid weak-supervision strategy systematically enhances label fidelity and significantly increases model learning robustness when compared to dependency on self-reports.

The overall research is structured as follows: Section II describes the literature review, Section III demonstrates proposed LEMIP-Net framework, Section IV illustrates experimental results, discussion and Section V includes Conclusion.

II. Literature Review

The literature review is performed through an organized selection strategy that assisted to recognize relevant research on mental-health prediction from social media during large-scale health crises. Specifically, peer-reviewed journal and research articles published were retrieved from standard journals using keywords including mental health prediction, social media analytics, pandemic-related sentiment analysis and longitudinal modeling. Then, the studies were filtered based on methodological consistency and relevance with significance on Machine Learning (ML) and Deep Learning (DL) approaches that are applied to mental-health inference from user-generated textual data.

Bashar, Nayak and Balasubramaniam [13] determined a hybrid deep learning model, which integrated Semi-Supervised Neural Topic Model (SNTM) and Informed Neural Network (INN) that evaluated COVID-19 discussions happened in Australian Twitter. Specifically, 2.9 million tweets were the data acquired, which were pre-processed and applied for SNTM, INN for topic discovery and sentiment classification, respectively through lexicon-based prior knowledge. Further, evolving public quires at outbreak time were interpreted by tweet volume, dynamic topic modelling, and semantic brand scoring. Thus, this method captured topic diversity and sentiments connected with real-world actions, but this model had challenges such as dependency on keyword-filtered data, English-only tweets, and lack of multimodal context, which affects the real-world deployment.

Inamdar, Chapekar, Gite & Pradhan [14] recommended a Machine Learning (ML) based framework, which detected mental stress in Reddit posts through NLP techniques. Further, this framework utilized reddit dataset that contains approximately 2800 labelled texts, which were pre-processed by various embedding strategies. Where the pre-processing techniques included Bidirectional Encoder Representations from Transformers (BERT) tokenization, Embeddings from Language Models (ELMo), and Bag-of-Words (BoW) representations. Specifically, these features were utilized to train classifiers such as logistic regression, SVM, XGBoost, and random forest models. Subsequently, this framework determined that with the limited data, the effective stress detection was possible. However, lack of demographic context, and exclusion of multimodal indications limits generalization among various sectors.

Abbas, Munir, Raza, Samee, Jamjoom & Ullah [15] introduced a depression detection model, which was a combination of BERT contextual embeddings and probabilistic features produced by random forest approach. Specifically, a dataset that contained 20,000 labelled tweets was considered and applied pre-processing. Subsequently, extracted contextual BERT embeddings and given to random forest that generated depression-related probability features, then these enhanced features were utilized to train many classifiers. Among which logistic regression attained greater accuracy and evaluated through statistical T-tests and k-fold cross-validation. Therefore, this model improved feature quality for mental health prediction, despite advantages this approach relied more on textual content and lacks user behavioral context that limits the real-world use cases.

Villa-Pérez, Trejo, Moin & Stroulia [16] demonstrated a ML approach, which used English and Spanish Twitter communications to detect nine mental health disorders. Further, two bilingual datasets were created from collected timelines of analyzed users by strict self-report patterns and cross verified with control users. Moreover, pre-processing of tweets were performed, through which linguistic features were extracted, which included, Part-of-speech (POS) tags, Linguistic Inquiry n-grams, q-grams, Word Count (LIWC), and word embeddings. Thus, this method attained greater accuracy through n-gram features, but this method contains limitations such as dependency on unverifiable self-reports, imbalance in dataset, minimized performance for low-frequency disorders and demographic mismatching.

Radwan, Amarneh, Alawneh, Ashqar, AlSobeh & Magableh [17] suggested an advanced approach that utilized Large Language Models (LLMs), ML algorithms and Generative Pre-trained Transformer 3 (GPT-3) embeddings. Specifically, these were to detect and classify social media posts that caused stress disorders and lower the mental health of individuals. Further, through all these considered techniques, a screen tool was generated that used online textual data, whereas posts were converted into vectors by GPT-3 embeddings, which also captured linguistic nuances and semantic meaning. However, there were challenges such as model bias, limited generalizability, dataset imbalance, low performance among populations, and require to improve the pre-processing techniques, which are significant for the further process in the approach.

Recent state-of-the-art approaches signifies that transformer-based embeddings and hybrid learning frameworks assists to effectively capture psychological signals from social media posts [13], [15], [17]. However, most existing methods depend on post-level classification, static representations and sparse self-reported labels, while neglecting cumulative exposure effects and longitudinal behavioral evolution [14], [16]. Thereby, these approaches remain limited in modeling sustained mental-health trajectories and exposure-induced distress. Hence, these limitations motivate the proposed LEMIP-Net framework, which integrates exposure-aware modeling with longitudinal sequence learning that helps for the development of mental-health prediction beyond static post-level baselines.

III. Methodology

The proposed research incorporates LEMIP-Net, which is a deep sequential neural architecture designed to model the temporal progression of user's emotional and linguistic behavior while simultaneously quantifying their exposure levels to pandemic-related content. Initially, the user-generated social-media posts are pre-processed into longitudinal timelines that allows the model to capture gradual psychological patterns instead of isolated expressions. Subsequently, individual post is semantically encoded using XLM-RoBERTa, then an exposure-classification module is utilized to evaluate every post to estimate the probability that belongs to health or pandemic discourse. Consequently, these probabilities are aggregated across temporal windows to compute a quantitative health-content intensity score, which provides a key variable reflecting how frequently and intensely user interacts with pandemic information. Further, the fused sequence of semantic embeddings, exposure intensities and refined mental-health indicators is then processed through the LEMIP-Net architecture, where temporal patterns are learned using Transformer and Bi-LSTM layers. Finally, a multi-task prediction module outputs the user's mental-health status, severity score and risk of future deterioration. Hence, this integrated design as demonstrated in Fig. 1, ensures that both content exposure and temporal behavioral evolution facilitates to provide accurate context-aware prediction of mental-health impact.

A. System Model and Data Description

The research incorporates the pandemic-period mental-health dataset [17] that comprises 32,487 social-media posts, which were collected from 4,216 unique users between March 2020 and July 2022. Specifically, each record includes the post text, timestamp, engagement metadata and sparse self-reported mental-health indicators. Additionally, the dataset includes content in English and Arabic, which reflects multilingual pandemic discourse. The dataset is split into 70% training, 15% validation and 15% testing that facilitates in maintaining user-level separation to avoid leakage. In particular, user-level separation facilitates that all posts from a specified user are allocated effectively to a single subset (training, validation, or testing), which helps in preventing overlap of user-specific linguistic or behavioral patterns across splits. This avoids information leakage and enables a reliable evaluation of the model's generalization to unseen users. Hence, this dataset provides a sufficiently large and a temporally rich resource for modeling longitudinal behavior, as the dataset comprises time-stamped post sequences spanning multiple months per user. Thereby, this operation allows the analysis of psychological changes over time instead of isolated observations. Further, each user u contributes a chronological sequence of posts $P_u = \{p_1, p_2, \ldots, p_T\}$, each associated with a timestamp t_i and enhanced metadata including hashtags, mentions, engagement and contextual cues, which are related to COVID-19, health fear, vaccination, restrictions, anxiety and uncertainty. Thus, the dataset for each user is expressed using (1):

$$D_u = (P_u, T_u, M_u) \qquad (1)$$

Where P_u signifies the post sequence, T_u refers to the corresponding timestamps and M_u defines the available mental-health labels or self-reports. Additionally, the textual content comprises naturally occurring pandemic-related expressions such as cases rising, quarantine, fever symptoms and vaccine fear while the mental-health indicators include user self-assessments, sentiment scores or psychological lexicon matches. Subsequently, data pre-processing is performed, which removes noise, bot-generated content, irrelevant posts and normalizes the text for stable embedding generation. Specifically, to define temporal granularity, a data-driven stability rule is applied, where the posting density is evaluated for individual user and choose the minimum window (Δt) where ≥ 80% of users have at least one post. Hence, the weekly windows assist to maximize temporal continuity while preventing sparse user sequences. Henceforth, to formally select the optimal temporal window Δt, the posting-density stability is assessed through (2):

$$S(\Delta t) = \left(\frac{1}{U}\right) \Sigma_u \, I(n_u(\Delta t) \ \geq \ 1) \qquad (2)$$

Where candidate window size (1,3,7,14 days) is defined as Δt, the proportion of user with ≥ 1 post in each window is demonstrated as $S(\Delta t)$, U signifies the total users and $n_u(\Delta t)$ determines the number of posts of user u in window Δt and $I(.)$ determines the indicator function. Thus, the temporal windowing is selected as $\Delta t = 7$ days since it maximizes the $S(\Delta t)$ while preventing fragmentation in low-activity users.

In particular, the raw dataset is transformed into a structured longitudinal behavioral record using a temporal aggregation technique, where user posts are chronologically organized into definite time window size that assists to reconstruct complete behavioral timelines for proposed LEMIP-Net.

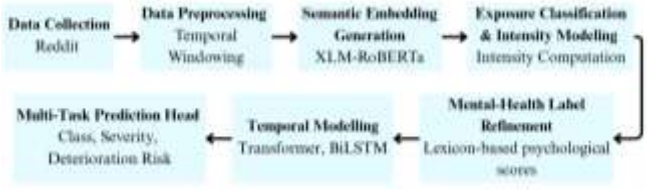


Figure 1. Flow diagram of proposed LEMIP-Net framework

Hence, this step ensures that the raw dataset, which is initially heterogeneous and sparse converted into a structured, time-sensitive representation. Hence, this operation is essential for capturing gradual psychological changes and exposure accumulation that a single-post models struggles to detect, thereby resolves a core limitation of design using only post.

B. Initialization and Semantic Embedding Construction

Further, each post p_t is converted into a dense semantic embedding by employing XLM-RoBERTa model, which is chosen based on the multilingual strength and ability to capture pandemic-related emotional and contextual features. Additionally, due to multilingual and code-switched pandemic discourse, the employed XLM-RoBERTa model applies SentencePiece tokenization with 250k vocabulary, Unicode NFKC normalization and maximum sequence length of 128 tokens where these parameters ensure multilingual consistency. Thus, the embedding process and semantic matrix is defined using (3) and (4), respectively:

$$e_i = f_{\text{XLM-R}}(p_i) \tag{3}$$

$$E = [e_1, e_2, \dots, e_T]^T \tag{4}$$

Where embedding vector of post p_i is defined as e_i, pretrained multilingual encoder is symbolized as $f_{\text{XLM-R}}$, user level embedding matrix is demonstrated as E and T signifies the number of temporal steps. In addition, by enhancing raw text with contextual semantics facilitates to improve data quality beyond the existing static LLM embeddings. Thus, this stage ensures that the model receives psychologically significant and globally representative language patterns which are related to mental-health change. Hence, the use of XLM-RoBERTa model for semantic embedding is essential as pandemic discourse is multilingual and context-sensitive. Also, the XLM-R model, which captures emotional, psychological and health-related semantics across languages which results with richer and more generalizable representations than English-only or traditional LLM embeddings. Henceforth, this semantically enhanced representation formulates the foundation upon which exposure estimation and mental-health inference depend.

C. Exposure Classification and Health-Content Intensity Estimation

Furthermore, to calculate the amount health-related content each user is exposed to, the proposed LEMIP-Net incorporates an exposure classifier F_{exp}, that computes each embedding e_t. Specifically, the Modelling exposure is crucial because psychological stress enhances with frequency of pandemic-related content. Thereby, without exposure modelling, the emotional variation may be misinterpreted. Thus, the classifier outputs the probability that the post concerns pandemic or health-related information as demonstrated in (5):

$$\hat{y}_i^{\varepsilon\xi\pi} = \sigma(W_{exp} e_i + b_{exp}) \tag{5}$$

Where the exposure probability is defined as $\hat{y}_i^{\text{exp}}$, W_{exp}, b_{exp} refers to classifier weight, bias respectively and σ denotes the sigmoid activation function. Hence, the cumulative exposure intensity over a window W_k is assessed through (6):

$$E_k = \frac{1}{|W_k|} \sum_{p_i \in W_k} \hat{y}_i^{\varepsilon\xi\pi} \tag{6}$$

Here, E_k refers to the normalized exposure frequency and intensity, specifically this module identifies posts discussing infection fear, symptoms, lockdown rules, rising cases, medical updates or anxiety triggers. Specifically, the training data for exposure classifier is established from manually tagged COVID-19 posts which includes 4,180 positive, 8,700 negative. Thereby, 2-layer MLP (128–64–1), learning rate = 1e−4, batch size = 32, Adam optimizer and dropout = 0.3 are utilized. In addition, the class imbalance is handled using focal loss ($\gamma = 2$) and random oversampling. Thus, the exposure classifier is trained using weighted binary cross-entropy as demonstrated in (7):

$$L_{exp} = -w_p \, y \, log(p) - w_n \, (1 - y) \, log(1 - p) \tag{7}$$

In particular, the exposure label is assigned through (8):

$$\hat{y} = \begin{cases} 1 & if \ p \geq \tau \\ 0 & otherwise \end{cases} \tag{8}$$

Where class weights for imbalance are defined as w_p, w_n respectively, y signifies true exposure label, predicted probability is defined as p and τ signifies the decision threshold where $\tau = 0.5$ during evaluation which is further optimized on validation. Subsequently, to define domain boundaries for exposure intensity, the predicted probability p is categorized into three regions based on the p range which is as follows: $p < 0.2$ signifies low-exposure, $0.2 \leq p \leq 0.8$ defines medium-exposure and $p > 0.8$ represents high-exposure. Hence, these boundaries are selected on the basis of maximizing inter-class separation on the validation set. Thus, the decision threshold $\tau = 0.5$ is considered for binary exposure assignment because it yields the highest Youden's J-statistic during classifier calibration.

Therefore, the resulting exposure time-series allows modelling the extent and persistence of health-related information a user observe. In particular, the exposure classification is employed because mental-health impact is strongly mediated by the volume, frequency and intensity of health-related content a user encounter. Hence, without explicit exposure modelling, AI systems risk conflating general emotional expression with crisis-driven psychological stress. Henceforth, by computing a quantitative exposure-intensity score for each time window, the proposed LEMIP-Net isolates the effect of health-content saturation, which allows downstream components to differentiate between natural emotional variability and exposure-induced distress.

D. Mental-Health Label Refinement Through Hybrid Weak Supervision

In addition, self-reported mental-health labels in social-media data are infrequent, therefore to obtain dense and usable supervision, the proposed LEMIP-Net integrates self-reported scores s_t with lexicon-derived psychological features l_t using (9):

$$y_t^{\rho\varepsilon\phi} = \alpha y_t^{\sigma\rho} + (1-\alpha) y_t^{\lambda\varepsilon\xi}, 0 \le \alpha \le 1 \qquad (9)$$

Where the refined label is represented as y_t^{ref} and α stands for confidence weight based on self-report presence. For instance, if a user reports self-stress score $s = 0.6$ but the lexicon score is the refined label which is demonstrated in Equation (10):

$$y_t^{\rho\varepsilon\phi} = 0.7(0.6) + 0.3(0.3) = 0.51 \qquad (10)$$

Specifically, if a user provides stress or anxiety self-ratings, the system preserves them. In particular, when such ratings are absent, psychological lexicons detect emotional features related to worry, fear, exhaustion and distress. Hence, these contradictions across self-reports and lexicon features are determined using a noise-aware correction rule as demonstrated where if both signals disagree, confidence-weighted averaging is used, missing values use only lexicon-based features and also lexicon noise is decrease through minimum-support threshold ($\ge 3\ symptom\ terms$).

Thus, the hybrid label facilitates every temporal step in the user's sequence which carries a mental-health estimate. Hence, this step assists to mitigate label sparsity which is a major limitation in the existing research by creating a stable ground truth that enhances learning and prevents temporal gaps in mental-state representation. Henceforth, hybrid label refinement through weak supervision is justified by the inherent sparsity and inconsistency of self-reported mental-health scores in real social-media datasets. Also, the manual labels are insufficient for training deep temporal models. Thus, combining self-reports with lexicon-based psychological indicators provides dense, consistent supervision signals which allows proposed LEMIP-Net the model to learn stable mental-health patterns without oversensitivity to annotation gaps.

E. Longitudinal Sequence Construction and Temporal Feature Encoding

Subsequently, each time step is defined by concatenating semantic embedding, exposure intensity and refined mental-health score. Specifically, for users with missing posts in window t, a padding vector is applied using (11):

$$x_t = [\vec{0}, 0, \mu_y] \qquad (11)$$

Here, the mean refined label of user is defined as μ_y and thereby the padding operation prevents temporal discontinuities. Further, the transformer assists to capture long-range global behaviour, whereas Bi-LSTM helps to capture local fluctuations using (12):

$$x_t = \left[e_t \| \mathrm{I_t} \| y_t^{\rho\varepsilon\phi}\right] \qquad (12)$$

Here, the fused vector is denoted by x_t, embedding is represented as e_t, ‖ refers to concatenation, exposure intensity is defined as I_t and y_t^{ref} signifies the refined label. Further, the transformer layer models long-range behavioral dependencies and hence the fused vector first fed through the transformer, whose output is then sequentially embedded into the Bi-LSTM as illustrated in (13):

$$H^{tr} = Transformer(X) \qquad (13)$$

Where the transformer output is denoted as H^{tr}, X signifies sequence of all x_t. Thus, this operation facilitates global-to-local feature flow that includes global patterns first then short-term variations. Subsequently, a Bi-LSTM layer captures short-term emotional fluctuations and bidirectional mental-health evolution as demonstrated in (14):

$$H^{lstm} = \mathrm{BiLSTM}(H^{tr}) \qquad (14)$$

Here, the BiLSTM output is defined as H^{lstm}. Thus, the fused vectors determine a complete psychological snapshot at individual time step. In particular, the Transformer captures global trends such as steadily increasing anxiety, while the Bi-LSTM models instant changes influenced by daily exposure. Thereby, this longitudinal modelling resolves the inability to account for temporal mental-health progression. Hence, the use of Transformer and Bi-LSTM layers is essential where the transformers learn global behavioral patterns, such as persistent anxiety themes or sustained exposure to crisis information, while Bi-LSTM assists to capture fine-grained emotional shifts between adjacent time steps. Henceforth, this integration ensures that both long-term mental-health evolution and short-term fluctuations are effectively modelled.

F. Multi-Task Mental-Health Impact Prediction

Finally, a multi-task prediction head is employed which processes temporal features to estimate three outcomes as illustrated in (15) – (17):

$$\hat{c} = Softmax(W_c h_T + b_c) \qquad (15)$$

$$\hat{s} = W_s h_T + b_s \qquad (16)$$

$$\hat{r} = \sigma(W_r h_T + b_r) \tag{17}$$

Here, predicted class is defined as $\hat{c}$, severity score is represented as $\hat{s}$, deterioration risk is symbolized as $\hat{r}$, the final Bi-LSTM state is demonstrated as h_T, the classifier weights refer to W_c, W_s, W_r and biases signifies b_c, b_s, b_r. Therefore, by jointly predicting class, severity and risk, the proposed LEMIP-Net model, which captures both immediate mental-health state and future vulnerability which allows a comprehensive assessment. Thus, the Multi-task training loss is demonstrated using (18):

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\chi\lambda\sigma} + \lambda_2 \mathcal{L}_{\sigma\varepsilon\varpi} + \lambda_3 \mathcal{L}_{\rho\iota\sigma\kappa} \tag{18}$$

Where classification loss is defined as $\mathcal{L}_{\text{cls}}$, severity regression loss is symbolized as $\mathcal{L}_{\text{sev}}$ and risk prediction loss is determined as $\mathcal{L}_{\text{risk}}$ and $\lambda_1, \lambda_2, \lambda_3$ signifies chosen weights $(0.5, 0.3, 0.2)$ respectively. Specifically, the multi-task head is optimized using AdamW with learning rate of $1e-5$, 0.01 weight decay and at 1.0 gradient clipping. Therefore, to stabilize multi-task optimization, an equalized gradient scaling g_i' is applied which is expressed through Equation (19):

$$g_i' = \frac{g_i}{\|g_i\|_2} \tag{19}$$

Here, the gradient contribution of each task i is denoted by g_i, specifically the gradient normalization facilitates that no single task dominates the optimization. Thereby, tach task-specific gradient g_i is scaled by its L2-norm $\|g_i\|_2$, which provides a balanced contribution during joint training. In particular, the single-task classification struggles to capture the subtle gradations of mental-health decline or quantify future vulnerability. Hence, the multi-task outputs provide clinically significant insights and enable predictive interpretations aligned with psychological theory.

IV. Experimental Setup

The proposed LEMIP-Net framework is executed using Python with PyTorch and the Hugging Face Transformers library for model development and fine-tuning. Specifically, data pre-processing and evaluation are performed using standard scientific-computing packages such as NumPy, pandas and scikit-learn. Thus, the experiments are implemented on a system with at least 32–64 GB RAM and a multi-core CPU that assists to handle sequence construction and exposure modelling efficiently. Hence, all experiments are executed in a controlled environment with fixed random seeds to ensure reproducibility and the complete training pipeline from embedding generation to multi-task prediction is performed on the same hardware and software configuration. Hence, the proposed LEMIP-Net framework is evaluated in terms of accuracy, precision, recall and F1-Score as demonstrated in the (20) – (23), respectively:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{20}$$

$$Precision = \frac{TP}{TP+FP} \tag{21}$$

$$Recall = \frac{TP}{TP+FN} \tag{22}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{23}$$

Here, TP is true positive, TN is true negative, FP is false positive, and FN is false negative, respectively.

A. Performance Analysis

To evaluate the effectiveness of proposed LEMIP-Net, performance analysis is performed against recent state-of-the-art deep-learning and transformer-based models which are used for mental-health prediction on social-media datasets. Specifically, these models include BERT, RoBERTa and XLNet which are strong baselines for emotional and psychological signal extraction. Thus, each model is fine-tuned under identical experimental conditions and evaluated across the standard metrics as illustrated in Fig. 2.

From Fig. 2, it is depicted that the proposed LEMIP-Net outperforms all the benchmark transformer models across every evaluation metric. Although, the conventional models such as BERT, RoBERTa and XLNet achieved better results due to their robust contextual encoding capabilities, still lacks explicit mechanisms that assists to model exposure intensity or temporal emotional drift which are both essential factors in mental-health prediction. Hence, these results illustrate that integrating exposure signals and temporal dynamics resulted with more reliable and clinically significant mental-health risk estimation.

B. Ablation Study

To compute the individual contribution of each architectural component in the proposed LEMIP-Net, an ablation study is performed conducted by incrementally integrating the major modules into a shared baseline. Specifically, all variants are trained under identical conditions and evaluated as presented in the Table I.

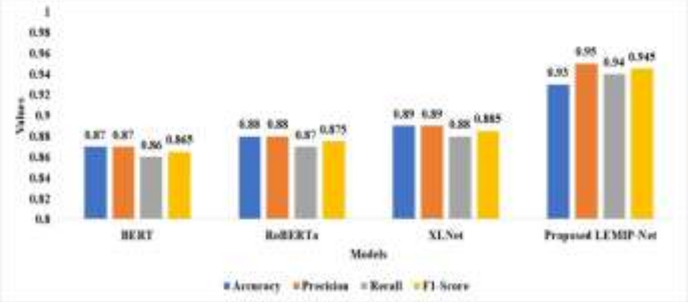


Figure 2. Performance analysis of proposed LEMIP-Net with conventional models

TABLE I. ABLATION STUDY OF PROPOSED LEMIP-NET ACROSS DIFFERENT VARIANTS

Model Variant	Accuracy	Precision	Recall	F1-score
Only XLM-RoBERTa embeddings	0.84	0.84	0.85	0.85
XLM-RoBERTa -Hybrid Weak Supervision (HWS)	0.86	0.86	0.87	0.87
XLM-RoBERTa -HWS-Transformer–Bi-LSTM	0.88	0.89	0.89	0.89
XLM-R + HWS + Temporal (No Exposure)	0.90	0.91	0.91	0.91
Proposed LEMIP-Net	0.93	0.95	0.94	0.945

From Table I, it is observed that the ablation results demonstrates that each component contributes significantly to performance improvement. Specifically, the base model provides only moderate accuracy which signifies that only text embeddings are insufficient. Therefore, by adding HWS, enhances label quality and further introducing temporal modelling improves performance by capturing behavioral changes across time. Additionally, included a no-exposure variant that helps to validate the independent effect of exposure modelling. Thus, the full proposed LEMIP-Net model incorporates exposure classification and intensity modelling that assists to obtain the highest accuracy and F1-Score which determines exposure-aware and longitudinal signals are essential for reliable mental-health impact prediction.

C. Comparative Results

To assess the robustness of the proposed LEMIP-Net framework, the performance is compared against the existing models which are widely used in mental-health prediction from social-media content. Specifically, the models include traditional machine-learning classifiers (SVM), lexicon-augmented gradient boosting (LIWC+XGB) and LLM-enhanced models (GPT-3 + SVM) as demonstrated in Table II. Hence, all comparative models are re-processed with identical tokenization, sequence length and filtering to ensure fair comparison.

Specifically, the existing models such as SVM [14] demonstrates moderate predictive capability, LIWC+XGB [16] and GPT-3 + SVM [17] obtains stronger performance, but still struggles due to the inability to incorporate temporal dynamics and exposure intensity. Hence, the proposed LEMIP-Net outperforms these models by incorporating longitudinal behavioral patterns, refined supervision and explicit modelling of pandemic-related exposure which results with higher accuracy of 0.93, 0.95 precision, 0.94 recall and 0.945 F1-score. Henceforth, these results demonstrates that modelling both the semantic evolution and exposure context significantly improves mental-health prediction compared to static or post-level baselines.

D. Discussion

The experimental results demonstrate that the proposed LEMIP-Net effectively resolves the major limitations in existing post-level mental-health prediction models. Specifically, the existing approaches primarily depended on isolated text embeddings or classical machine-learning classifiers which limited their ability to capture the cumulative psychological effects of prolonged exposure to pandemic-related content. In contrast, the proposed LEMIP-Net integrates diverse approaches including label refinement, temporal behavioral modelling and exposure-intensity quantification which results model that is both context-sensitive and longitudinally effective. Thus, the performance observed in both benchmark comparisons and ablation results determine that mental-health risk is predicted effectively when user behavior is considered as a dynamic trajectory rather than a set of independent posts. Additionally, the results illustrate that incorporating exposure intensity provides substantial predictive advantage over transformer baselines such as BERT, RoBERTa and XLNet. Hence, this defines that psychological distress in digital environments is strongly influenced by the frequency and severity of health-related information encountered which evaluates the conceptual foundation of exposure-aware modelling. Furthermore, HWS significantly enhances label quality which illustrates those self-reports alone lack reliability and benefit from linguistic signal enhancement. In particular, on analyzing the errors, it is determined that existing transformer-based approaches misclassify posts with high exposure but neutral tone, whereas the proposed LEMIP-Net model correctly incorporates exposure signals to avoid such false negatives. Henceforth, the proposed LEMIP-Net framework not only outperforms existing models but also provides a methodology aligned with psychological and behavioral science insights.

TABLE II. COMPARATIVE RESULTS OF PROPOSED LEMIP-NET WITH EXISTING MODELS

Models	Accuracy	Precision	Recall	F1-score
SVM [14]	0.74	0.70	0.74	0.76
LIWC+XGB [16]	0.823	0.997	0.802	0.881
GPT-3 + SVM [17]	0.86	0.84	0.83	0.84
Proposed LEMIP-Net	0.93	0.95	0.94	0.945

V. CONCLUSION

In this research, the proposed LEMIP-Net, which is an exposure-aware and longitudinal deep learning framework that is designed for mental-health prediction. Specifically, pandemic-related social media content is impacting every individual mental health. Further, in the proposed LEMIP-Net model raw user posts are converted into structured behavioral timelines, through which this model captures linguistic evolution, emotional drift, and cumulative effect of exposure to pandemic. Hence, the experimental results demonstrates that proposed LEMIP-Net consistently outperforms the conventional models and transformer-based approaches. This determines the requirement of combining exposure and temporal dimensions into mental health prediction model. Additionally, the ablation analysis shows that each component such as hybrid weak supervision, temporal encoding, and exposure modelling in the proposed LEMIP-Net influences the overall performance of model. In particular, the capability of the proposed LEMIP-Net in integration of refined labels, time-dependent behavioral patterns and quantified exposure signals determined as robust approach for mental-health risk assessment. Thus, the proposed LEMIP-Net outperformed existing model with which results with higher accuracy of 0.93, 0.95 precision, 0.94 recall and 0.945 F1-score. Henceforth, the proposed approach resolves key challenges of existing approaches and contributes a scalable and robust framework for predicting mental-health risks in digital ecosystems. In the future, the proposed LEMIP-Net will explore multi-modal integration including images or engagement behavior, demographic conditioning and real-time deployment for public-health surveillance.

ACKNOWLEDGMENT

The authors thank the Department of Computer Information Systems at Thomas Jefferson University for providing the academic environment and support essential for this research. L. Sztandera thanks Thomas Jefferson University, Philadelphia, PA, USA, for continued guidance and institutional resources that contributed to the successful completion of this work.

REFERENCES

[1] R. Mahevish, A. Khan, H. R. Mahmood, S. Qazi, H. M. Fakhoury, and H. Tamim, “The impact of social media on the physical and mental well-being of medical students during the COVID-19 pandemic,” J. Epidemiol. Global Health, vol. 13, pp. 902-910, November 2023.

[2] F. Benrouba and R. Boudour, “Emotional sentiment analysis of social media content for mental health safety,” Social Network Anal. Min., vol. 13, p. 17, January 2023.

[3] M. H. Al Banna, T. Ghosh, M. J. Al Nahian, M. S. Kaiser, M. Mahmud, K. A. Taher, M. S. Hossain, and K. Andersson, “A hybrid deep learning model to predict the impact of COVID-19 on mental health from social media big data,” IEEE Access, 11, pp. 77009-77022, July 2023.

[4] S. Zhou and M. Mohd, “Mental Health Safety and Depression Detection in Social Media Text Data: A Classification Approach Based on a Deep Learning Model,” IEEE Access, vol. 13, pp. 63284-63297, April 2025.

[5] İ. Aygün, B. Kaya, and M. Kaya, “Identifying patients in need of psychological treatment with language representation models,” Multimedia Tools Appl., vol. 84, pp. 397-418, 2025.

[6] M. E. Lelisho, D. Pandey, B. D. Alemu, B. K. Pandey, and S. A. Tareke, “The negative impact of social media during COVID-19 pandemic,” Trends in Psychology, vol. 31, pp. 123-142, May 2022.

[7] M. Kyrou, I. Kompatsiaris, and P.C. Petrantonakis, “Deep learning approaches for stress detection: A survey,” IEEE Trans. Affective Comput., vol. 16, pp. 499-517, September 2024.

[8] S. A. Alanazi, A. Khaliq, F. Ahmad, N. Alshammari, I. Hussain, M. A. Zia, M. Alruwaili, A. Rayan, A. Alsayat, and S. Afsar, “Public’s mental health monitoring via sentimental analysis of financial text using machine learning techniques,” Int. J. Environ. Res. Public Health, vol. 19, p. 9695, August 2022.

[9] M. Kerasiotis, L. Ilias, and D. Askounis, “Depression detection in social media posts using transformer-based models and auxiliary features,” Social Network Anal. Min., vol. 14, p. 196, September 2024.

[10] L. Ilias, S. Mouzakitis, and D. Askounis, “Calibration of transformer-based models for identifying stress and depression in social media,” IEEE Trans. Comput. Social Syst., vol. 11, pp. 1979-1990, June 2023.

[11] A. Sharaff, T. R. Chowdhury, and S. Bhandarkar, “Lstm based sentiment analysis of financial news,” SN Comput. Sci., vol. 4, p. 584, July 2023.

[12] K. Milintsevich, K. Sirts, and G. Dias, “Towards automatic text-based estimation of depression through symptom prediction,” Brain Inf., vol. 10, p. 4, February 2023.

[13] M.A. Bashar, R. Nayak, and T. Balasubramaniam, “Deep learning based topic and sentiment analysis: COVID19 information seeking on social media,” Social Network Analysis and Mining, vol. 12, no. 1, p. 90, 2022.

[14] S. Inamdar, R. Chapekar, S. Gite, and B. Pradhan, “Machine learning driven mental stress detection on reddit posts using natural language processing,” Hum.-Centric Intell. Syst., vol. 3, pp. 80-91, March 2023.

[15] M. A. Abbas, K. Munir, A. Raza, N. A. Samee, M. M. Jamjoom, and Z. Ullah, “Novel transformer based contextualized embedding and probabilistic features for depression detection from social media,” IEEE Access, vol. 12, pp. 54087-54100, April 2024.

[16] M. E. Villa-Pérez, L. A. Trejo, M. B. Moin, and E. Stroulia, “Extracting mental health indicators from english and spanish social media: A machine learning approach,” IEEE Access, vol. 11, pp. 128135-128152, November 2023.

[17] A. Radwan, M. Amarneh, H. Alawneh, H. I. Ashqar, A. AlSobeh, and A. A. A. R. Magableh, “Predictive analytics in mental health leveraging LLM embeddings and machine learning models for social media analysis,” Int. J. Web Serv. Res (IJWSR), vol. 21, pp. 1-22, January 2024.