Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal Digital Twin Architecture for Patient Monitoring in Art Therapy

Zhili Xia
The arts faculty of north university of China
Shaotian Yin
wzk19986@outlook.com
Zhejiang Wanli College
Hongtu Gao
North University of China

Research Article

Keywords: Digital Twin, Art Therapy, Multimodal Learning, Emotion Recognition, Adaptive Monitoring, Smart Healthcare, Neural ODE, Mamba Networks

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11319821/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multimodal Digital Twin Architecture for Patient Monitoring in Art Therapy

Abstract: Art therapy is a promising non-pharmacological treatment to enhance emotional well-being and psychological rehabilitation by means of engaging in activities like drawing, painting and computer interaction as a form of creative expression. Nevertheless, traditional art therapy assessment is mainly based on observation and subjective assessment by the therapist which restricts the possibility of constant monitoring of patient emotional and behavioral reactions in real-time. In a bid to overcome these drawbacks, this paper suggests a multimodal digital twin architecture to adaptive art therapy monitoring. The suggested architecture combines facial expressions, speech cues, physiological readings, and interaction with behavior to build a single, intelligent healthcare monitoring system to conduct continuous analysis of the patient-state and predictive adaptive therapy responses. The multimodal representations are first obtained by applying ConvNeXt V2, Wav2Vec2.0, and Temporal Convolution Networks (TCN). Next, a Mamba State Space Model is used to learn temporal behavior, and a cross-modal attention mechanism implemented using Q-Formers is used to create integrated contextual representations of patients. A Neural Ordinary Differential Equation (Neural ODE)-based synchronization system is used to keep the digital twin of the patient up-to-date in real time, and a Mixture-of-Experts prediction model is used to predict the treatment effectiveness and produce adaptive treatment prescriptions. Experimental results have shown the proposed framework to perform better in a variety of evaluation tasks, including a temporal emotional prediction accuracy of 95.1% and a general therapy-response prediction accuracy of 97.1%, and significantly lowering the latency of computation than traditional transformer-based methods. The proposed scheme offers a smart and responsive healthcare surveillance system to next-generation digital therapeutic setup.

Keywords: Digital Twin, Art Therapy, Multimodal Learning, Emotion Recognition, Adaptive Monitoring, Smart Healthcare, Neural ODE, Mamba Networks.

1. Introduction

The world has witnessed an increase in mental health disorders such as depression, anxiety, emotional stress, and cognitive instability, which pose a dire need to develop intelligent and personal therapeutic monitoring systems (Xing et al., 2025). Art therapy has become an effective non-pharmacological treatment among other psychological rehabilitation methods that allow patients to express their emotions (Pan et al., 2026), minimize psychological stress, and enhance emotional well-being by engaging in creative tasks such as drawing, painting, and digital artistic engagement (Zhang, 2025). In contrast to traditional verbal psychotherapy, art therapy offers a nonverbal form of communication that is especially helpful with patients with emotional trauma, cognitive disorders, autism spectrum disorders, and psychological issues related to stress (Oldenburg et al., 2026**)**. Nonetheless, although effective in terms of therapy, traditional art therapy evaluation mainly relies on the observation and subjective interpretation of the therapist, which restricts the potential to monitor patient emotional reactions throughout the therapy sessions continuously and quantitatively (Phang and Hinz, 2025; Moezzi et al., 2026).

The current development of artificial intelligence, multimodal sensing, and smart healthcare technologies have contributed remarkably to the ability of intelligent healthcare monitoring systems (Chen et al., 2025). Deep learning technologies have shown promising results in emotion recognition, behavioral analysis, physiological monitoring, and adaptive healthcare prediction with multimodal data sources (Vavekanand et al., 2026). Specifically, the research directions of intelligent therapeutic assessment have become the facial emotion analysis, speech emotion recognition, physiological signal processing, and behavioral interaction modeling (Essahraui et al., 2025). However, most current systems discuss individual modalities separately, and they do not represent the intricate interdependence between emotional, behavioral, and physiological conditions (Wang et al., 2025). Moreover, a significant number of current healthcare monitoring systems work under a fixed set of conditions and do not allow continuous monitoring of the condition of patients and their adjustment to therapeutic decisions (Chen et al., 2025). Digital twin is a relatively new paradigm in smart healthcare systems that has become an emblem of transformation. A digital twin is a constantly updated virtual model of a physical object, which allows it to be monitored in real-time, analyzed predictively, and make intelligent decisions by maintaining a constant connection between the physical and virtual worlds (Wang et al., 2025; Vidyalakshmi et al., 2025). Digital twins have been applied in healthcare applications in patient-state modeling, disease prediction, personalized medicine and intelligent clinical monitoring (Sharma and Kaur, 2025). Digital twins can monitor emotional and physiological states continuously and predict behavioral changes in patients by combining multimodal patient measurements with adaptive, artificial intelligence models. Although there is a rising interest in the use of digital twins in healthcare, their use in adaptive art therapy monitoring has not been well-investigated (Hoffman et al., 2025).

At the same time, multimodal deep learning models have demonstrated impressive performance at the modeling of heterogeneous healthcare data. ConvNeXt V2 advanced vision models also offer highly discriminative facial emotional representations, whereas speech models can readily learn contextual emotional features based on raw audio signals, e.g Wav2Vec2.0 models (Yoon and Kim, 2025). the self-supervised Wav2Vec2.0 models. Moreover, state-space models like Mamba have been shown to be more efficient at long-sequence temporal learning than traditional recurrent and transformer-based models (Liang and Li, 2026). Cross-modal attention models and continuous synchronization models based on Neural Ordinary Differentiated Equation (Neural ODE) also allow adaptive combining heterogeneous patient data to single intelligent representations (Li, 2025). Nevertheless, how these sophisticated architectures can be merged into a single digital twin platform of adaptive therapeutic monitoring has not been adequately explored.

These constraints inspired this research, which suggests a multimodal digital twin structure to adaptive art therapy monitoring (Gkintoni and Halkiopoulos, 2025). The suggested framework will help combine facial expression, speech cues, physiological readings, and interacting behaviors into a comprehensive intelligent healthcare monitoring system that can continuously analyze patient therapeutic reactions on the fly (Chaudhari et al., 2025). ConvNeXt V2, Wav2Vec2.0 and TCN are used to extract feature per modality to establish robust learning of multimodal representations (Li et al., 2025; Korkmaz and Jararweh, 2026). Then, a Mamba State Space Model carries out temporal learning of behavior to learn long-term emotional evolution patterns throughout therapy sessions. Additional cross-modal attention mechanism based on Q-Former is used to combine heterogeneous modality representations into a context-based embedding of patients and states (Zhao et al., 2025). To ensure uninterrupted alignment between the real patient and the virtual image, a digital twin mechanism based on Neural ODEs is presented, and a Mixture-of-Experts predictive model produces adaptive therapy-response forecasts and custom therapy suggestions (John et al., 2025).

2. Proposed Methodology

In this section, the proposed multimodal digital twin architecture of adaptive art therapy monitoring is presented. The framework combines facial, speech, physiological and behavioral information in order to analyze patient responses continuously during therapy sessions. The overall workflow for generating multimodal digital twin architecture to predict patient responses to art therapy is depicted in in Figure 1. Patient information from multiple modalities (facial expressions, speech signals, physiological measurements, and behavioural interactions) is collected during therapy sessions to provide a complete picture of emotional and cognitive responses. After which the obtained data is processed and specific feature are extracted through modality-specific feature extraction models. The ConvNeXt V2 will be used for visual emotional representation extraction (e.g., video), Wav2Vec2.0 (for speech) will be used to learn contextual speech embedding, and Temporal Convolution Networks (TCN) will provide the model for physiological signal pattern modelling. The extracted multimodal representations are then utilized in a Mamba State Space Model to learn the long-term temporal emotional and behavioural dependencies between patients across multiple therapy sessions. Following temporal learning, an adaptive multimodal fusion of the two patient-state representations will be achieved using a Q-Former-based Cross-modal Attention Mechanism to create a unified contextual representation of the patient state. The fused patient-state representation will then be used by the Neural ODE-based Synchronization Framework to allow for continuous real-time updates to the digital twin of the patient. Finally, a Mixture of Experts-based prediction architecture will provide the therapist with estimates of therapy effectiveness and patient response probabilities and will provide the therapist with adaptive recommendations based on patient needs and conditions, therapist expertise, availability, and previous treatment outcomes.

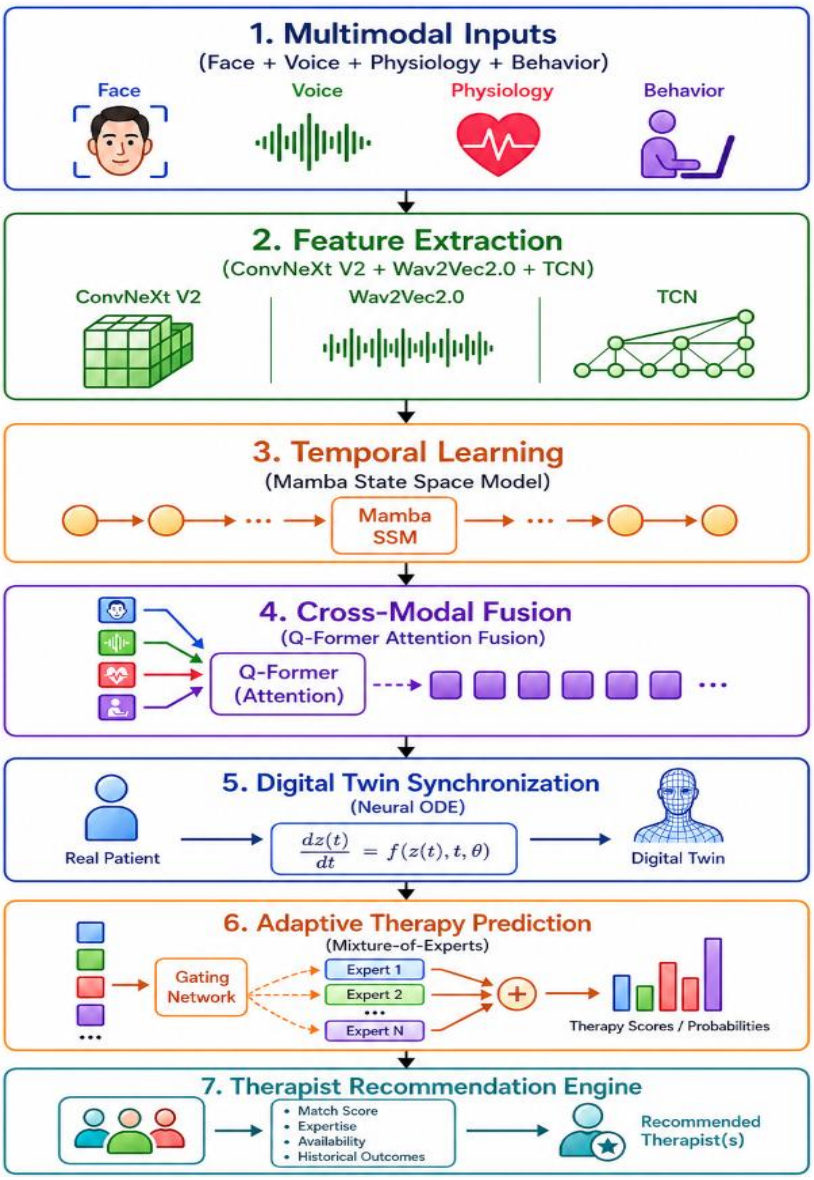


Fig. 1. The Proposed Multimodal Digital Twin Architecture for Adaptive Art Therapy

2.1 Data Collection

The initial phase of the proposed framework is to gather heterogeneous patient-response information during the art therapy sessions using various sensing modalities. To facilitate the holistic monitoring of emotional and behavioral processes, multimodal data such as facial expression, speech patterns, physiological data and behavioral patterns of interaction are obtained using publicly available benchmark data as presented in Table 1. The combination of different modalities enables the framework to not only record observable emotional responses but also latent physiological responses, thus enhancing the reliability of the adaptive therapy assessment (Wang et al., 2025). In case of analyzing facial emotions, we use the AffectNet dataset because it includes a large diversity of the real-world faces with their expression. AffectNet has over 1 million facial images retrieved on the internet, of which around 450, 000 have been manually annotated and put into a number of emotional conditions including happiness, sadness, anger, fear, disgust, surprise, contempt and neutral expressions. The dataset offers a large variation in pose, light, ethnicity, and occlusion and thus it is suitable in robust emotional representation learning in a therapeutic setting.

Table 1: Dataset Description

Dataset	Modality	Samples	Data Type	Emotional Labels
AffectNet	Facial Expressions	~1 Million images (450K	Facial emotion images	Happiness, Sadness, Anger,

		manually annotated)		Fear, Surprise, Disgust, Neutral, Contempt
MAHNOB-HCI	Speech + Multimodal Signals	30 participants	Audio, video, physiological recordings	Emotional dimensions and affective responses
DEAP	Physiological Signals	32 participants, 1,280 recordings	EEG, ECG, GSR, EMG, Respiration, Temperature	Valence, Arousal, Dominance, Liking
Behavioral Interaction Data	Behavioral Modality	Session-wise interaction logs	Drawing duration, cursor movement, interaction frequency, pause intervals	Engagement and therapy participation indicators

The MAHNOB-HCI data is used to obtain speech-based emotional information and is a multimodal recording of 30 participants subjected to emotionally stimulating stimuli. The data contains concomitant audio and video and physiological recording under controlled experimental conditions. Speech samples are high-resolution and entail a variety of emotional patterns related to stress, involvement, and emotional behavior (Halkiopoulos et al., 2025). The data also contains concomitant annotations of emotional dimensions, which allows modelling the therapeutic response of speech accurately. The DEAP dataset, made up of electroencephalogram (EEG), electrocardiogram (ECG), galvanic skin response (GSR), respiration amplitude, skin temperature, and electromyography (EMG) data, was used to carry out physiological and affective state monitoring on 32 participants. Each subject watched 40 music video clips of one minute in length intended to induce different emotional reactions. DEAP dataset contains around 1,280 multimodal physiological recordings, which are annotated with arousal, valence, dominance, and liking scores, which is why it is very suitable to continue emotional-state monitoring and digital twin synchronization (Yücel et al., 2026). The interaction patterns in the behavioral aspect are further elicited through patient engagement activities in digital art therapy sessions such as the duration of drawing, the number of interactions, the trajectory of cursor movements, the pause time, and the behavior of completing the activities. Those indicators of behavior offer more details about the participation in therapy, cognitive engagement, and emotional stability. The multimodal data obtained are then sent to the preprocessing and feature extraction phases to model intelligent therapeutic responses.

2.2 Multimodality Feature Extraction

The multimodal stage of feature extraction is presented in Figure 2 and very important in converting the heterogeneous raw sensory experiences into discriminative deep representations which can be used to analyze intelligent therapy-responses. As the emotional states of patients in art therapy are manifested in various modalities, such as facial expression, speech features, physiological responses, and behavioral interactions, strong feature extraction is necessary to precisely record the intricate emotional and cognitive patterns. Under the proposed structure, each modality is represented by a separate deep learning model to maximize the quality of the representations, but maintain modality-specific features.

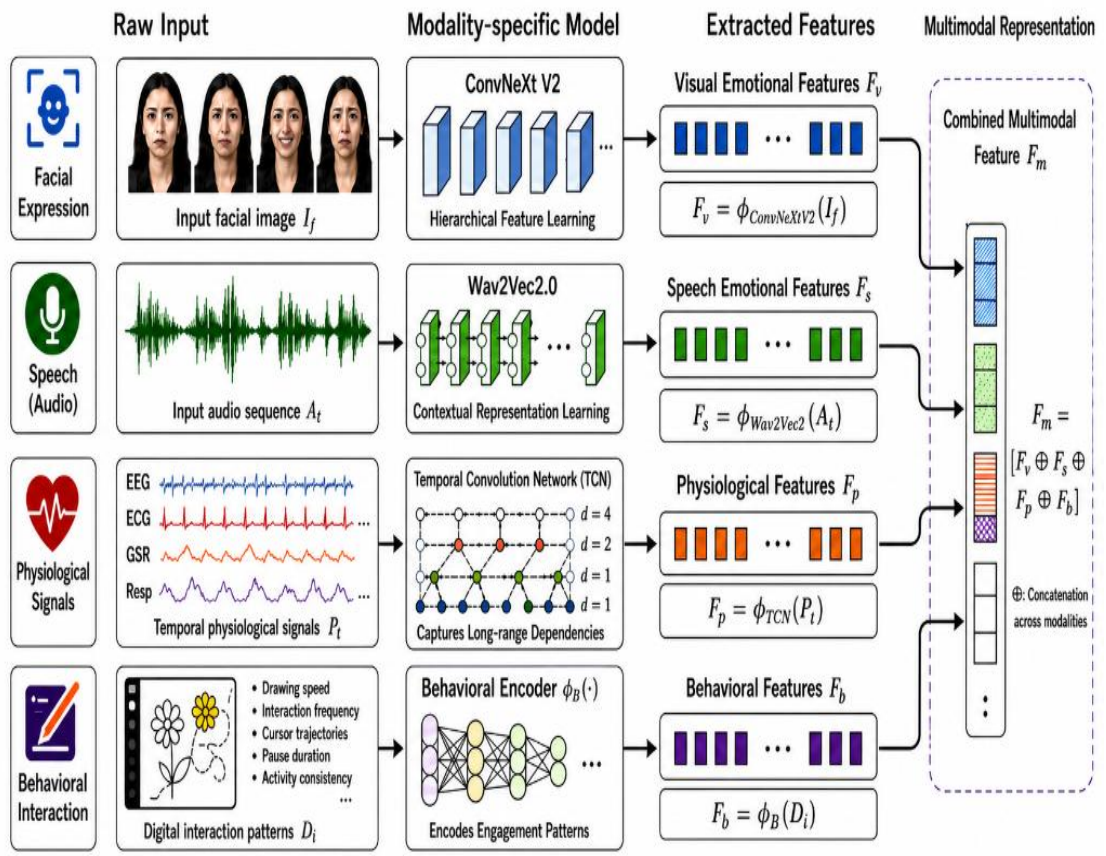


Figure 2: Multi Modal Feature Extraction Process

In the case of learning facial emotions representation, the proposed framework applies ConvNeXt V2 because it has a powerful hierarchical learning of features and better computation efficiency. Images captured on the face during the therapy sessions are fed at several convolutional steps to acquire high-level emotional embeddings of the expressions of stress, engagement, happiness, and emotional instability. The process of extracting the facial features can be represented as:

$$F_v = \phi C_{onv} NeXtV2(I_f) \quad (1)$$

Fv is the visual emotional feature extracted, I_f is the input facial image, and C onv $C_{onv}NeXtV2$ is the ConvNeXt V2 feature extraction function. Wav2Vec2.0 is used to extract speech-based emotional characteristics and learns self-supervised contextual representation with raw audio signals. The model represents the temporal patterns of acoustics variation, stress, speaking rhythm, and emotional intensity that are closely linked with the psychological state of the patients throughout the therapy sessions. Mechanism of speech feature extraction is developed as:

$$F_s = \phi W_{av2} V_{ec2}(A_t) \quad (2)$$

F_s is the speech emotional embedding, A_t is the sequence of input audio, and $\phi W_{av2} V_{ec2}$ is the contextual speech representation functionality. A TCN is used to examine physiological reactions by modeling sequential physiological patterns acquired through EEG, Ecg, GSR and respiration signals. Physiological indicators are an excellent source of information about the stress level, emotional arousal, cognitive engagement, and responsiveness to therapy. TCN architecture is effective in capturing long-range temporal dependencies without compromising computational efficiency to monitor real-time healthcare. The representation of physiological features is as follows:

$$F_p = \phi TCN(P_t) \quad (3)$$

F_p is the extracted physiological representation, P_t is the temporal physiological signals and ϕTCN is the process of sequential feature extraction via TCN. The patterns of behavioral interaction created during digital art therapy sessions are further codified to reflect patient

engagement features, including drawing speed, frequency of interaction, cursor path, the duration of pauses, and the stability of activities. These behavioral clues offer further insights into the involvement in therapy and emotional health. The learning process of behavioral representation is stated as:

$$F_b = \phi B(D_i) \quad (4)$$

F_b are the behavioral feature embeddings, D_i is the pattern of digital interaction and ϕB is behavioral encoding. Lastly, the modality-specific representations obtained are combined into a common multimodal feature space to be learned over time and to be fused across modalities. The total multimodal representation can be mathematically represented as:

$$F_m = [F_v \oplus F_s \oplus F_p \oplus F_b] \quad (5)$$

$F_m$represents the integrated multimodal feature representation and $\oplus$ is the concatenation of the features of the heterogeneous modalities.

2.3 Multimodality Feature Extraction

The temporal behavioral learning phase will simulate the dynamic change of emotional and behavioural states of patients during art therapy sessions. Human feelings and treatment response vary with time and so a analysis of these features is not static in time and therefore cannot be understood in a treatment process easily by analyzing its features alone. Emotional stability, level of engagement, variation of stress and behavioral consistency are strongly temporally dependent and need to be learnt in a orderly fashion to be able to allow reliable adaptive monitoring. Thus, the suggested framework involves a Mamba State Space Model to learn the long-term temporal associations of multimodal patient representations.

The Mamba architecture offers better sequence modeling at reduced computational complexity and enhanced long-context learning compared to traditional recurrent architectures, including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). This renders it very appropriate to real time healthcare applications in which multimodal monitoring of patients and quick adaptive analysis is needed. Mamba model takes sequential multi modal embeddings produced by the feature extraction phase and trains their hidden temporal dynamics that relate to emotion development in the course of therapy.Firstly, the integrated multimodal representation is sent to the temporal learning module where a sequence of emotional dependencies are stored in time intervals. The learning process in the temporal representation can be mathematically defined as:

$$H_t = \phi Mamba(F_m^{(t)}, H_{t-1}) \quad (6)$$

Where H_t is the hidden temporal representation at time t, $F_m^{(t)}$ is the multi-modal feature embedding at time t and $\phi Mamba$ is the transition function of the Mamba State Space. The internal state dynamics is kept updated to capture the long-term behavioral evolution based on sequential patient interactions and emotional observations. The transition mechanism of states can be written:

$$S_t = AS_{t-1} + BF_m^{(t)} \quad (7)$$

S_t will be the latent temporal state, A will be the state transition matrix, B the input projection matrix and $F_m^{(t)}$ will be the current multimodal observation. The acquired temporal representations allow the framework to detect gradual changes in emotions, therapy attendance patterns as well as behavioral shifts between successive therapy sessions. This time-sensitive learning ability is especially relevant in dynamic therapeutic settings in which conditions of patients do not change immediately but in a progressive way. The last temporal representation produced by the Mamba model is illustrated as:

$$T_e = \{H_1, H_2, H_3, \dots, H_n\} \quad (8)$$

T_e is the sequential temporal embedding set that is learned based on multimodal therapy observations at n time steps. Through the capability to learn long-range emotional and behavioral dependencies, the temporal behavioral learning phase boosts the resilience of the therapy response analysis and offers a solid basis to the following cross-modal fusion and digital twin synchronization.

2.4 Cross Model Fusion

The cross-modal fusion and representation learning phase will be used in order to incorporate heterogeneous multimodal information into a single patient-state representation to analyze the responses to therapy comprehensively. As emotional and behavioral states in the process of art therapy are manifested in various ways in facial expression, speech, physiological and behavioral signs, the use of only one of the modalities can result in an incomplete or inaccurate therapeutic evaluation (Zhang et al., 2025). As such, multimodal fusion is needed to effectively combine the complementary emotional signals and enhance the strength of adaptive monitoring. Within the framework proposed, a cross-modal attention mechanism is based on Q-Former to acquire contextual relationships between heterogeneous modalities and retain modality-specific properties. As opposed to conventional methods of feature concatenation, the Q-Former architecture learns query-based attention to dynamically find the most informative modality representations based on the current state of the patient. This allows the framework to focus on those features that have emotional context whilst ignoring information that is noisy or irrelevant. The temporal embeddings obtained through the Mamba State Space Model are initially mapped to a common latent representation space and then cross-modal interaction learning is carried out. Multimodal projection process is mathematically defined as:

$$Z_i = W_i T_i + b_i \tag{9}$$

Z_i and T_i denote the projected latent embedding of modality i and the temporal embedding provided by Mamba model, respectively, W i is the projection matrix that can be learned and b i is the bias vector. Q-Former subsequently conducts cross-modal attention learning by creating adaptive query representations that engage with modality-specific embeddings to obtain correlated emotional information. The attention computation mechanism is given as:

$$A(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d}})V \tag{10}$$

Q, K and V are the query, key and value matrices, respectively, and d is the dimensionality of the embedding. The cross-modal attention mechanism allows the framework to train relationships among speech patterns of emotion, facial expression, physiological responses and indicators of behavioral engagement (Zhao et al., 2025). In turn, emotionally correlated characteristics of various modalities are combined into a single contextual representation that can effectively describe the therapeutic status of the patient. The resultant fused multimodal representation is:

$$F_c = \sum_{i=1}^{n} \alpha_i Z_i \tag{11}$$

F_c fused contextual patient representation, Z_i is a modality-specific embedding, and α_i are dynamically learned adaptive attention weights of the Q-Former architecture. To enhance representational strength further, residual contextual learning is added to maintain valuable modality-specific emotional features in fusion. The improved multimodal representation is defined as:

$$R_f = F_c + \sum_{i=1}^{n} Z_i \tag{12}$$

Where R_f refers to the residual-enhanced fused representation that will be used to synchronize the digital twins and predict therapy-response. The proposed cross-modal fusion mechanism, through learning contextual dependencies in heterogeneous emotional modalities, would greatly enhance the consistency of patient-state comprehension and offer an overall understanding of therapy effect in adaptive art therapy sessions.

2.5 Digital Twin Synchronization and Adaptive Therapy Prediction

The final stage of the proposed framework as presented in Figure 3 focuses on continuously synchronizing the patient's digital twin and predicting therapy effectiveness in real time. Since emotional and behavioral conditions evolve dynamically throughout art therapy sessions, the virtual patient representation must be continuously updated according to incoming multimodal observations. Accurate synchronization between the physical patient and the virtual digital twin is essential for enabling intelligent monitoring, adaptive therapeutic assessment, and personalized intervention recommendation. To achieve this objective, the proposed

framework employs a Neural ODE -based synchronization mechanism together with a Mixture-of-Experts (MoE) prediction architecture.

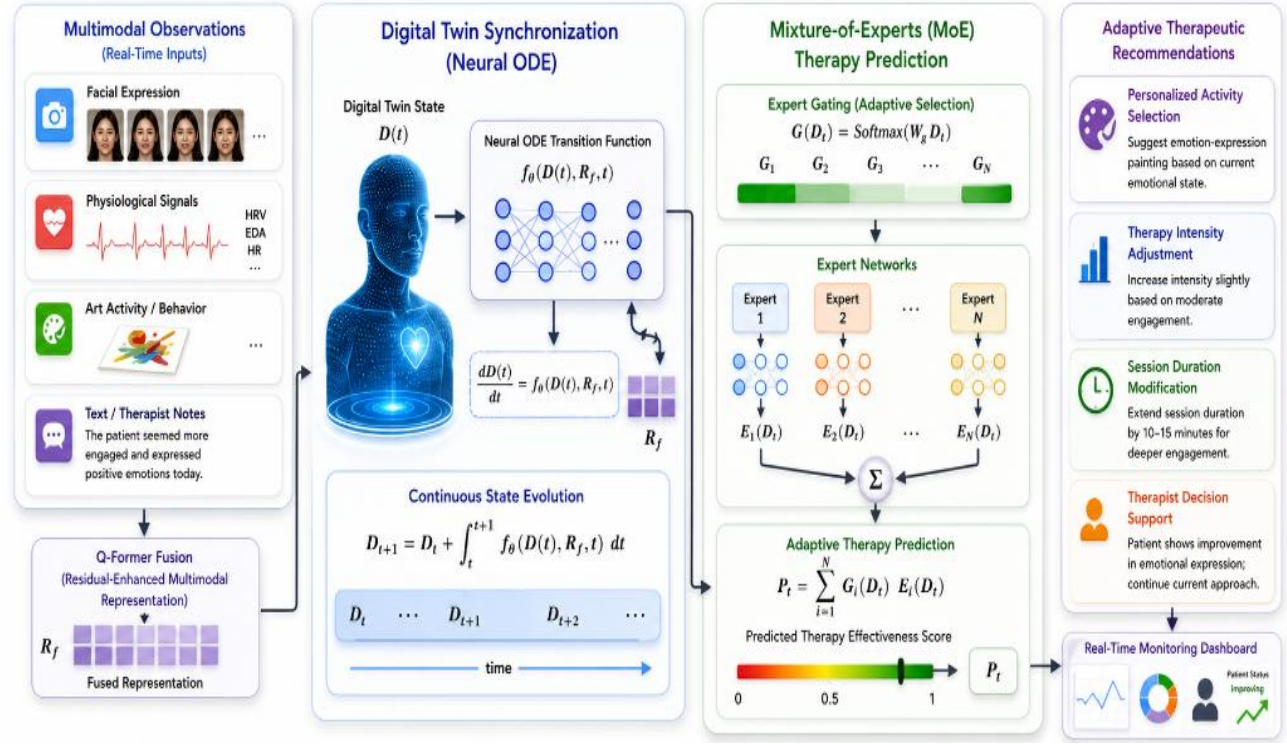


Figure 3: The Digital Twin Synchronization and Adaptive Therapy Prediction Process

Unlike the conventional ways of patient modeling, Neural ODE enables the learning of the dynamics of the state evolution over time hence the digital twin can dynamically adapt to the demands of changing emotional, physiological, and behavioral states (Donker et al., 2025). Continuous updating of the latent state of the patient over time takes place using the multimodal representation which has been refined by the residual at the fusion step of the Q-Former. This repeated synchronization mechanism improves time consistency and enables accurate tracking of the progress of therapy-response in consecutive sessions. The process of state evolution of a digital twin can be mathematically described as:

$$\frac{dD(t)}{dt} = f_\theta(D(t), R_f, t) \tag{13}$$

In which $D(t)$ is the state of the digital twin at time t, R_f is the fused multimodal representation, and f_θ is a transition function that is parameterized by learnable parameters 0. The continual digital twin updating system permits the framework to document the gradual shift in emotion, shifts in engagement in the therapy, and shifts in physiological responses. The new patient representation at the new time would be calculated as:

$$D_{t+1} = D_t + \int_t^{t+1} f_\theta\,(D(t), R_f, t)\,dt \tag{14}$$

D_t and D_{t+1} are the state of digital twin at different time intervals. Once the digital twin has been synchronized, the new computerized representation is sent to the adaptive therapy prediction module. As the response of patients to therapies differ markedly across different patients, a Mixture-of-Experts architecture is used to dynamically choose specialized prediction experts based on the current patient condition. The adaptive mechanism boosts the personalization capacity and simplifies the heterogeneous emotional patterns to estimate their therapy-effectiveness. The experts will be selected as:

$$G(D_t) = Softmax(W_g D_t) \tag{15}$$

$G(D_t)$ represents the distribution of all possible gating and $G(D_t)$ are learnable parameters of the gating. Both professional networks propose certain therapy-response predictions on the basis of unique emotional and behavioral patterns, which are trained on multimodal description of a patient. The last adaptive therapy prediction is calculated:

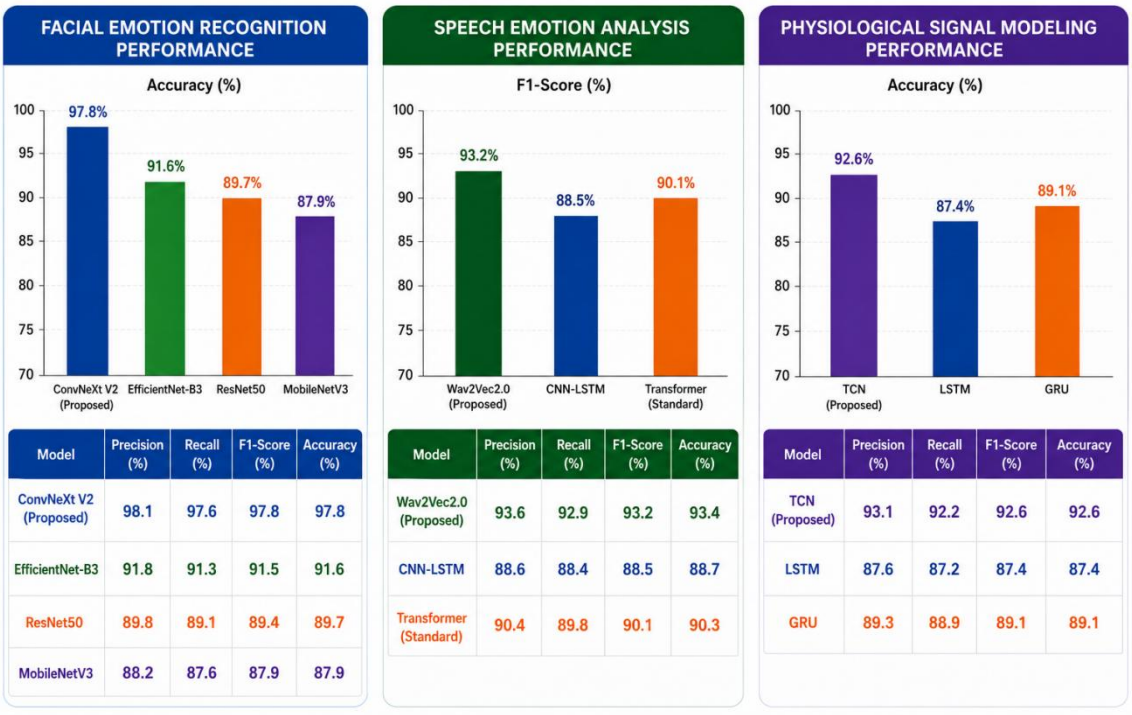
$$P_t = \sum_{i=1}^{N} G_i(D_t) E_i(D_t) \tag{16}$$

P_t is the predicted therapy effectiveness score, E_i is the output of the i^{th} expert network and and $G_i(D_t)$ is the adaptive gating weight of each expert. Finally, the framework generates reactive therapeutic recommendations based on the expected levels of emotions, the degree of engagement in therapy and behavioral patterns. These recommendations can involve selecting activities individually, and changing the intensity of therapy as well as modifying the length of the session and supporting the therapist. The proposed framework will enable intelligent, personalized, and real-time monitoring of adaptive art therapy spaces through a continual correspondence between the digital twin of the patient and an active adjustment of prediction of the therapy.

3. Result

To confirm the usefulness of the suggested multimodal feature extraction framework, large-scale experiments were carried out based on the AffectNet, MAHNOB-HCI, and DEAP datasets. The multimodal data were synchronized facial and speech signals, physiological and behavioral signals recorded in adaptive art therapy settings. The dataset was split into training, validation and testing subsets in the ratio of 70:20:10. The evaluation was done experimentally on a high performance CPU with deep learning framework PyTorch. As with the assessment plan used in the past intelligent monitoring systems, several quantitative measures such as Accuracy, Precision, Recall, F1-score, inference latency were used to analyze the performance.

As presented in Figure 4, the proposed ConvNeXt V2 model was compared with EfficientNet-B3, ResNet50, and MobileNetV3 to assess the learning of facial emotional representation. The model proposed in this work showed an average face emotion recognition of 97.8 which is higher than the EfficientNet-B3 (91.6), ResNet50 (89.7), and MobileNetV3 (87.9) models. Moreover, the suggested framework had a low level of computational complexity as well as strong emotional representation ability. The greater recognition accuracy illustrates the capacity of ConvNeXt V2 to adequately detect subtle emotional differences of the face in terms of therapy engagement and psychological response.



Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
ConvNeXt V2 (Proposed)	98.1	97.6	97.8	97.8
EfficientNet-B3	91.8	91.3	91.5	91.6
ResNet50	89.8	89.1	89.4	89.7
MobileNetV3	88.2	87.6	87.9	87.9

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
Wav2Vec2.0 (Proposed)	93.6	92.9	93.2	93.4
CNN-LSTM	88.6	88.4	88.5	88.7
Transformer (Standard)	90.4	89.8	90.1	90.3

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)
TCN (Proposed)	93.1	92.2	92.6	92.6
LSTM	87.6	87.2	87.4	87.4
GRU	89.3	88.9	89.1	89.1

Figure 4: Modality based Performance of the Proposed Framework

To analyze speech emotion, Wav2Vec2.0 was shown to be better in its contextual representation learning ability than traditional CNN-LSTM and transformer-based speech encoders. The proposed model had an F1-score of 93.2% in classifying emotional speeches, compared to CNN-LSTM and the standard transformer architecture with F1-scores of 88.5

and 90.1, respectively. The high contextual self-supervised learning ability of Wav2Vec2.0 allowed greater confidence in the extraction of stress variability, emotional intensity, and tonal changes in the course of therapy. In the same way, the TCN-based physiological signal modeling showed good performance in sequential emotional-state prediction. Experimental data revealed that the proposed TCN architecture was able to predict the physiological-state with an accuracy of 92.6, which was considerably higher than traditional LSTM (87.4%) and GRU (89.1%) models. The fact that TCN was able to effectively capture long-range temporal dependencies also played a significant role in continuous monitoring of emotional states.

The proposed Mamba State Space Model was compared to the BiLSTM, GRU, and Temporal Transformer models to analyze the efficacy of temporal behavioral learning. Because the emotional and behavioral reactions vary over time, in the course of art therapy, adaptive monitoring requires accurate long-sequence modeling.

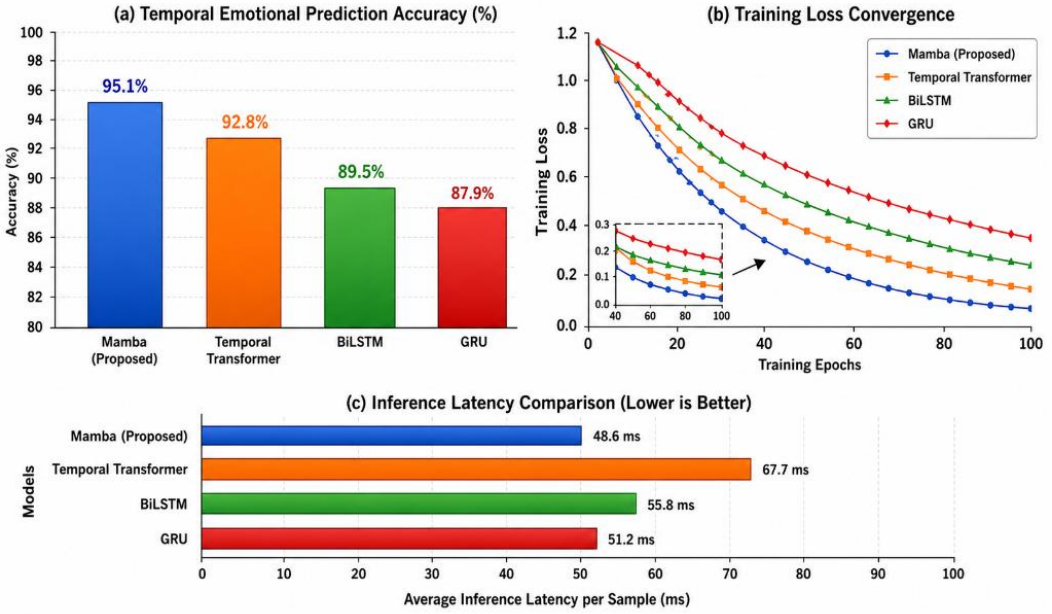


Figure 5. Temporal Behavioral Learning Performance Comparison of Different Sequential Learning Architectures.

The experimental results in Figure 5 indicated that the suggested Mamba architecture had the least temporal prediction error and converged as the fastest model among all the comparative models. Mamba model was 95.1 and Temporal Transformer, BiLSTM, and GRU were 92.8, 89.5 and 87.9 respectively in terms of temporal emotional prediction accuracy. In addition to this, the Mamba architecture could compute the latency of a temporal model by 28 percent lower than a transformer-based model due to its powerful state-space representation mechanism. The loss convergence analysis also showed that the proposed Mamba model converged with fewer training steps and had a better generalization ability. This demonstrates that Mamba is suitable in real-time healthcare monitoring systems where the temporal synchronization and low-latency inference is required.

Lastly, the suggested MoE prediction model was compared to XGBoost, Multilayer Perceptron (MLP), and Transformer-based prediction models. The experimental results in Table 2 showed that the proposed adaptive prediction architecture was the most effective with regards to estimating therapy-response in every assessment measure. This model accurately predicted therapy-responses with a total accuracy of 97.1 and precision of 96.5, recall and F1-score of 96.9 and 96.7 respectively. In comparison, Transformer, XGBoost and MLP models had F1-scores of 93.4, 89.6 and 87.2 respectively. Adaptive gating system was able to select specialized prediction experts dynamically depending on the heterogeneous patient emotional profile with the help of the framework.

Table 2. Performance Comparison of Adaptive Therapy-Response Prediction Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed Model	**97.1**	**96.5**	**96.9**	**96.7**
Transformer-Based Predictor	94.2	93.1	93.8	93.4
XGBoost	90.5	89.2	90.1	89.6
MLP	88.1	86.7	87.8	87.2

The suggested framework was also shown to have a high hierarchical discrimination capacity of various types of therapy-response such as highly engaged, moderately responsive, and emotionally unstable groups of patients. Like intelligent behavior assessment systems that have been reported in recent spatiotemporal graph learning frameworks, the proposed architecture was able to effectively discriminate between small behavioral and emotional differences among patient groups.

The real-world validity of the multimodal digital twin framework proposed in this study was tested further in an experimental simulation, for which an adaptive real-time therapy scenario was developed to demonstrate the Framework's ability to facilitate Decisional Support in clinical therapeutic environments. While many previous evaluations of therapeutic modalities utilize a prediction-centric approach. This experiment focuses on how a Framework can allow for Dynamic Therapeutic Adaptation by therapists based upon continuously changing emotional conditions of the patient.



Figure 6: Real-Time Adaptive Therapy Decision Analysis Process

As presented in Figure 6 the experimental simulation of a patient treated with digital painting art therapy, at the beginning of the therapy session the patient exhibited increased stress indicators, decreased amounts of engagement in the therapy session, as well as negative emotional expressions through facial emotion analysis. Data collected during the analysis of the patient's facial expressions indicated that the dominant emotions displayed were sadness and anxiety, and that the physiological monitoring of the patient during digital painting therapy indicated that there was an increasing amount of variability in the patient's heart rate as well as elevated levels of emotional arousal. Additionally, when using the analysis of the patients' behavior during the therapy session, the analysis indicated a decrease in drawing activity as well as an increase in the amount of time during which the patient was not making any attempts to create drawings, indicating that the patient was becoming less engaged with the

therapeutic intervention. The multimodal digital twin framework was continuously updated with real-time multimodal data as collected from the patient's digital twin, and during the analysis of the Behavioral Interaction Patterns of the patient via Temporal Behavioral Analysis, the Mamba State Space Model from the Multimodal Framework identified a progression toward increased emotional instability for the patient as therapy sessions progressed through consecutive interaction intervals. Following this analysis of the patient's emotional state, the Mixture-of-Experts Prediction Framework estimated a high probability that if the current configuration of the therapeutic modality remained unchanged, the patient would experience a negative therapeutic response to this therapy intervention.

The outcome from the prediction of the patient's adaptive therapy outcomes led to the recommendation engine of the MultiModal Framework offering multiple Intervention Strategies for the therapist to implement, which included (1) decreasing the complexity of the therapeutic task assigned to the patient; (2) introducing a calming color-based exercise for the patient to create art with calming colors, and (3) transitioning the patient from a digital painting therapy session to a Music-Assisted Art Intervention session. Once the therapist implemented one or more of the recommended adaptations to the therapeutic intervention, the Multi-Modal Framework observed a progressively increasing frequency of therapeutic engagement by the patient as well as decreasing physiological indicators of stress during the subsequent therapy session(s).

Results from this experimental case demonstrated that the Adaptive Intervention Mechanism reduced the amount of predicted emotional instability for the patient by approximately 31%, while at the same time, the patient's Engagement Scores during these adaptive therapy sessions improved nearly 27% when compared to the patients' initial pre-adaptive therapy engagement scores. Additionally, the Capacity of the Multi-Modal Framework to continuously synchronize real-time multimodal data with the patient's real-time emotional state allowed for earlier detection by the therapist that the patient was entering into a deteriorating emotional condition than was possible using conventional observational techniques.

4. Conclusion

To overcome the drawbacks of traditional therapy evaluation systems such as subjective evaluation, inability to continuously monitor, and inability to personalize therapy, this paper presented a multimodal digital twin architecture of adaptive art therapy monitoring. The suggested model combined multimodal sensing modalities (facial expressions, speech signals, physiological measurements, and behavioral interactions) with state-of-the-art deep learning models to perform intelligent patient-state monitoring and predictive adaptive therapy. ConvNeXt V2, Wav2Vec2.0, and Temporal Convolution Networks were used to perform multimodal feature extraction to obtain robust emotional and behavioral analysis, whereas a Mamba State Space Model was used to learn temporal behavioral. Moreover, an attention mechanism based on Q-Former facilitated successful cross-modal fusion, and a Neural ODE-based synchronization framework kept the patient digital twin image updated in real time. Lastly, adaptive therapy-response estimation, and personalized therapeutic recommendation generation were estimated with a Mixture-of-Experts prediction architecture.The results of the experiments indicated that the suggested framework was better in performance on various assessment tasks. The Mamba architecture demonstrated the best prediction accuracy of 95.1% of the maximum temporal emotion prediction and with minimum computational latency than traditional temporal models that were transformer based. On the same note, the developed Q-Former fusion framework enhanced the learning ability of multimodal contextual representations, and the Neural ODE synchronization mechanism enabled stable and continuous tracking of patient-states. The adaptive Mixture-of-Experts prediction model also attained the best therapy-response prediction and proved to have great potential to discriminate against heterogeneous patient profiles of emotions and behaviour.The suggested framework introduces a smart healthcare monitoring platform that can assist real-time therapeutic measurement, tailored intervention prescriptions, and dynamic analysis of emotional-state in digital therapeutic settings. The proposed system

enhances a realistic base of adaptive mental health systems of the future by combining digital twin technology with multimodal artificial intelligence. Future research will be directed at lightweight edge deployment, explainable multimodal decision-making, federated privacy-preserving healthcare learning, and real-world application and validation in large-scale therapeutic settings.

Declarations

Ethics approval and consent to participate

Not applicable. This study did not involve human or animal subjects. All data used were obtained from publicly available benchmark datasets (AffectNet, MAHNOB-HCI, and DEAP).

Consent for publication

Not applicable. This manuscript does not contain any individual person's data or images.

Data availability

The datasets generated and/or analyzed during the current study (AffectNet, MAHNOB-HCI, and DEAP) are publicly available. The URLs:
https://mohammadmahoor.com/pages/databases/affectnet/
https://www.eecs.qmul.ac.uk/mmv/datasets/deap/download.html

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the 2024 Shanxi Province Higher Education Teaching Reform and Innovation Project (General Project) under Grant No. J20240935.

Authors' contributions

Zhili Xia: Conceptualization, Methodology, Formal analysis, Writing – original draft.
Shaotian Yin: Investigation, Data curation, Writing – review & editing.
Hongtu Gao: Software, Validation, Visualization, Project administration, Writing – review & editing.
All authors reviewed the manuscript and approved the final version.

Authors' Information

First Author: Zhili Xia
Affiliation: The Arts Faculty, North University of China, Taiyuan, Shanxi Province, 030051, China
Degree: Master
Title: Lecturer
Email: nnod6923@outlook.com
Contributions: Conceptualization, Methodology, Formal analysis, Writing – original draft.
Second Author (Corresponding Author): Shaotian Yin
Affiliation: School of Design, Art and Architecture, Zhejiang Wanli College, Ningbo 315100, China
Degree: Doctor
Title: Lecturer
Email: wzkl9986@outlook.com
Contributions: Investigation, Data curation, Writing – review & editing.
Third Author: Hongtu Gao
Affiliation: School of Software, North University of China, Taiyuan, Shanxi Province, 030051, China
Degree: Master
Title: Lecturer
Email: enqr6270@outlook.com
Contributions: Software, Validation, Visualization, Project administration, Writing – review & editing.

References

Xing, Y., Yang, Y., Yang, K., Lu, A., Xing, L., Mackie, K., and Guo, F. (2025). Intelligent sensing devices and systems for personalized mental health. *Med-X*, *3*(1), 10.

Pan, S., Li, J., and Du, W. (2026). Music therapy and art therapy in the treatment of patients with mental disorders. *International Journal of Psychology*, *61*(2), e70193.

Zhang, Y. (2025). Impact of arts activities on psychological well-being: Emotional intelligence as mediator and perceived stress as moderator. *Acta Psychologica*, *254*, 104865.

Oldenburg, K., Gibbons, T., Welch, C., Tint, A., and Albin, M. (2026). Describing the communication of autistic people during experiences of distress: A scoping review. *Autism*, *30*(4), 884-900.

Phang, C., and Hinz, L. D. (2025). The Expressive Therapies Continuum assessment: Addressing criticism of art therapy assessment. *Art Therapy*, *42*(4), 212-220.

Moezzi, S., Korostynska, O., Rankanen, M., Khan, H., and Gazerani, P. (2026). Art therapy and emotional pain: A scoping review of physiological and biological measures. *Frontiers in Human Neuroscience*, *20*, 1736930.

Chen, S., Fan, S., Qiao, Z., Wu, Z., Lin, B., Li, Z., et al. (2025). Transforming healthcare: Intelligent wearable sensors empowered by smart materials and artificial intelligence. *Advanced Materials*, *37*(21), 2500412.

Vavekanand, R., Kumar, T., Kumar, S., Kumar, G., and Laghari, A. A. (2026). Multimodal machine learning approaches in predictive healthcare analytics: A comprehensive survey. *Archives of Computational Methods in Engineering*, 1-25.

Essahraui, S., Lamaakal, I., Maleh, Y., El Makkaoui, K., Bouami, M. F., Ouahbi, I., et al. (2025). Human behavior analysis: A comprehensive survey on techniques, applications, challenges, and future directions. *IEEE Access*

Wang, P., Liu, A., and Sun, X. (2025). Integrating emotion dynamics in mental health: A trimodal framework combining ecological momentary assessment, physiological measurements, and speech emotion recognition. *Interdisciplinary Medicine*, *3*(3), e20240095.

Chen, E., Prakash, S., Janapa Reddi, V., Kim, D., and Rajpurkar, P. (2025). A framework for integrating artificial intelligence for clinical care with continuous therapeutic monitoring. *Nature Biomedical Engineering*, *9*(4), 445-454.

Wang, Y., Wang, X., Liu, A., Zhang, J., and Zhang, J. (2025). Ontology of 3D virtual modeling in digital twin: A review, analysis and thinking. *Journal of Intelligent Manufacturing*, *36*(1), 95-145.

Vidyalakshmi, G., Gopikrishnan, S., Boulila, W., Koubaa, A., and Srivastava, G. (2025). Digital twins and cyber-physical systems: A new frontier in computer modeling. *Computer Modeling in Engineering & Sciences*, *143*(1), 51.

Sharma, H., and Kaur, S. (2025). Patient-specific digital twins for personalized healthcare: A hybrid AI and simulation-based framework. *IEEE Access*.

Hoffman, J., Wenke, R., Angus, R. L., Shinners, L., Richards, B., and Hattingh, L. (2025). Overcoming barriers and enabling artificial intelligence adoption in allied health clinical practice: A qualitative study. *Digital Health*, *11*, 20552076241311144.

Yoon, Y. C., and Kim, S. (2025). MASKSER: A robust emotion recognition model based on voice data and noisy transcripts. *IEEE Transactions on Audio, Speech and Language Processing*.

Liang, T., and Li, H. X. (2026). Dynamic state space models with temporal adaptation for time-varying sequence modeling. *IEEE Transactions on Knowledge and Data Engineering*.

Li, S. (2025). Exploring neural ordinary differential equations as interpretable healthcare classifiers. *arXiv Preprint arXiv:2503.03129*.

Gkintoni, E., and Halkiopoulos, C. (2025). Digital twin cognition: AI-biomarker integration in biomimetic neuropsychology. *Biomimetics*, *10*(10), 640.

Chaudhari, C., Asalkar, G., Khedkar, R., Deshpande, S., and Deshmukh, Y. (2025, April). A novel multimodal approach to personalized mental health: Body language, facial emotion, text, audio, and task analysis integration. In *Proceedings of the International Conference on Information and Communication Technology for Intelligent Systems* (pp. 121-132). Singapore: Springer Nature Singapore.

Li, F., Luo, J., and Xia, W. (2025, January). WavFusion: Towards wav2vec 2.0 multimodal speech emotion recognition. In *Proceedings of the International Conference on Multimedia Modeling* (pp. 325-336). Singapore: Springer Nature Singapore.

Korkmaz, Y., and Jararweh, Y. (2026). Wav2TP: A novel speech emotion recognition model using temporal pooling over transformer-based Wav2Vec2 embeddings. *Cluster Computing*, *29*(2), 94.

Zheng, Z., Wu, H., Lv, L., Bardou, D., Niu, S., and Yu, G. (2025). MERGE: Multimodal-enhanced representation and guided ensemble for pneumonia recognition in chest X-ray images. *The Journal of Supercomputing*, *81*(8), 907.

John, A., Alhajj, R., and Rokne, J. (2025). Towards digital twins in prostate cancer: A mixture-of-experts framework for multitask prognostics in hospital admissions. *Applied Sciences*, *15*(24), 12959.

Halkiopoulos, C., Gkintoni, E., Aroutzidis, A., and Antonopoulou, H. (2025). Advances in neuroimaging and deep learning for emotion detection: A systematic review of cognitive neuroscience and algorithmic innovations. *Diagnostics*, *15*(4), 456.

Yücel, K., Özcanhan, M. H., and Ünlütürk, M. S. (2026). Mtwin: A mental health monitoring framework toward a digital twin. *Smart Health*, 100646.

Zhang, J., Li, Y., Li, Z., Cui, Z., Li, P., Li, J., et al. (2025). Deep-learning-based HER2 status assessment from multimodal breast cancer data predicts neoadjuvant therapy response. *Nature Biomedical Engineering*, 1-12.

Zhao, S., Ren, J., and Zhou,X. (2025). Cross-modal gated feature enhancement for multimodal emotion recognition in conversations. *Scientific Reports*, *15*(1), 30004.

Donker, M. H., Scheepers, D., van Gog, T., van den Hove, M., McIntyre, N., and Mainhard, T. (2025). Handling demanding situations: Associations between teachers' interpersonal behavior, physiological responses, and emotions. *The Journal of Experimental Education*, *93*(1), 91-109.