IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 November 2023, accepted 16 December 2023, date of publication 25 December 2023,
date of current version 5 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347346

RESEARCH ARTICLE

Application of Image Classification Based on Improved LSTM in Internet Reading Therapy Platform

JIANXIN XIONG[1], HUI YIN[2], AND MEISEN PAN[1]
[1]Hunan University of Arts and Science, Changde, Hunan 415000, China
[2]Hunan University of Medicine, Huaihua, Hunan 418000, China

Corresponding author: Hui Yin (yinhui_hh@163.com)

This work was supported in part by the Hunan University of Arts and Sciences Professor and Doctor Research Start-Up Project Fund under Grant 22BSQD01; and in part by the General Project ''Research on Psychological Health and Ideological and Political Education of College Students Based on Reading Therapy'' of the Hunan Provincial Social Science Achievement Evaluation Committee, in 2024.

ABSTRACT Reading therapy is an effective approach for improving mental states or addressing disabilities associated with individuals' dyslexia. In traditional approaches, this process is performed through the intervention and supervision of an expert, which incurs time and cost. However, by utilizing artificial intelligence technologies, the reading therapy process can be automated. This article focuses on presenting an internet-based automated platform for reading therapy. In this method, audio and visual features during the reading therapy are processed using deep learning techniques to identify the individual's emotional state based on their reading status. In this state, two separate convolutional neural networks are used to describe the facial image features and speech characteristics of the individual. Then, the described features from these two models are merged to determine the individual's mental states using LSTM layers. Finally, a reinforcement learning model is used to provide feedback and design subsequent exercises. This reinforcement approach leads to continuous improvement of the evaluated process and plays a significant role in enhancing the efficiency of the internet-based reading therapy system. The performance of the proposed method has been evaluated based on information from 20 volunteers. According to the results, the proposed method can effectively improve individuals' mental states and compete with the conventional supervisor-based approaches. The performance of the proposed deep learning models in identifying emotional states has also been investigated. The results indicate that this model achieves a minimum improvement of 9.71% in emotional state recognition compared to previous research, with an average correlation coefficient of 0.64485.

INDEX TERMS Internet-based reading therapy, reading therapy framework, deep learning, fuzzy logic, reinforcement learning.

I. INTRODUCTION

Reading, a cornerstone of personal growth and development, holds immense potential to enhance individual well-being and cognitive abilities. It serves as a potent cognitive enhancer, effectively boosting concentration and fostering deeper comprehension [1]. Studies have demonstrated the efficacy of reading in improving sleep quality and alleviating anxiety, with daily reading practices consistently reducing anxiety levels by over 60% in most individuals [2]. This remarkable effect stems from the ability of reading to induce a state of tranquility akin to meditation, leading to profound relaxation and inner peace [3]. Moreover, reading contributes to the expansion of general knowledge and the cultivation of creativity, empowering individuals to navigate the complexities of the world with greater insight and innovation.

Amidst these well-established benefits, there is a concerning trend of declining per capita reading rates across various

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Da Lin.

societies [4]. This decline in reading habits coincides with a rising prevalence of mental health disorders, with depression rates in the United States exhibiting a steady increase, from 7.3% in 2015 to 8.6% in 2019 and 9.2% in 2020 [5]. The stark contrast between the benefits of reading and the growing mental health crisis has prompted the development of reading therapy, a promising approach to addressing psychological distress through the power of literature.

Reading therapy is a scientific and effective approach for improving mental states or addressing disabilities associated with individuals' dyslexia. This therapeutic approach is often planned by experienced professionals based on interactions with the individual undergoing treatment. Therefore, the process of reading therapy requires careful supervision at all stages of implementation [6]. This requirement makes the reading therapy process time-consuming, costly, and susceptible to errors. On the other hand, the lack of continuous access to a therapist makes it impossible to benefit from reading therapy methods in all circumstances [7]. These conditions indicate the necessity of providing an automated framework for reading therapy to address the shortcomings and limitations of traditional approaches.

Deep learning, a subset of artificial intelligence, holds immense promise for revolutionizing reading therapy by enabling smarter and more personalized experiences [8]. Deep learning models can analyze facial expressions, tone of voice, and other nonverbal cues to assess patients' emotional states during reading sessions [9]. This real-time monitoring allows therapists to dynamically adjust the reading materials and therapeutic interventions based on patients' emotional responses, ensuring a more effective and sensitive approach [10].

This article focuses on presenting a new and automated platform for reading therapy. The proposed system utilizes the global internet network as a communication platform. The widespread penetration of the internet in various communities and its attractiveness to individuals facilitate the ease of use and motivation for adopting the proposed approach. In the proposed reading therapy system, deep learning techniques are used to analyze the individual's emotional states, and reinforcement learning strategies are employed for designing therapeutic programs. The contribution of this article includes the following aspects:

- In this article, a new platform for internet reading therapy, based on machine learning techniques is proposed. In this system, therapy programs are determined reactively based on the individual's emotional states.
- The current article presents a hybrid deep learning model for recognizing individuals' mental states. This model consists of a parallel combination of two CNN models that are used to describe facial and speech emotional characteristics. The extracted features from these two models are merged, and the individual's emotional states are determined using LSTM components.
- In this article, an interactive model based on reinforcement learning is proposed for designing therapy programs. This reinforcement model includes a number of learning automata that determine the next exercises based on the identified emotional states for the individual. In this reinforcement model, reward and penalty operators are used to continuously improve the individual's mental state and avoid incorrect strategies.

These aspects have not been investigated in previous research and can be considered as innovative aspects of the current article. The structure of the remaining sections of this article is as follows: In the second section, some relevant research is studied. The third section presents the details of the proposed platform for internet-based reading therapy. The fourth and fifth sections are dedicated to discussing the research findings and presenting the conclusions, respectively. Table 1, lists the abbreviations used in this article.

TABLE 1. Table of abbreviations.

Abbreviation	Definition
CNN	Convolutional Neural Network
FBP	Factorized Bilinear Pooling
FC	Fully Connected
GAN	Generative Adversarial Network
GUI	Graphical User Interface
KNN	K-Nearest Neighbors
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MFCC	Mel-Frequency Cepstral Coefficients
NLP	Natural Language Processing
RBF	Radial Basis Function
RMSD	Root Mean Square Deviation
RMSE	Root Mean Squared Error
SSL	Secure Sockets Layer
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine

II. RELATED WORKS

The number of studies conducted in the field of reading therapy has been limited, and the application of machine learning techniques in these studies has received partial attention. In [11], an interactive tool based on artificial intelligence techniques is presented for smartphones, which addresses issues related to writing and dyslexia. In this method, the individual's activities in using the software are initially monitored to determine their level of dyslexia using Support Vector Machines (SVM). Then, based on predefined rules, training exercises are suggested. It should be noted that this research only provides a limited design and no information is reported regarding its performance. In [12], a framework for treating speech and hearing disorders in children and adolescents is proposed. This research first introduces a set of therapeutic tools based on existing processing and communication technologies and then suggests suitable strategies for system design based on the strengths and weaknesses of each tool. Accordingly, a set of machine learning techniques is proposed to be employed in this framework. However, similar to [11], this study also limits itself to presenting the design and suggesting strategies for its development.

In [13], the relationship between individuals' blood types and their preferred book genres is studied, and an attempt is made to provide a method for recommending content in reading therapy programs based on that. This study was conducted on 80 individuals with blood types A, B, AB, and O, and each individual indicated their reading preferences through a questionnaire. According to the results of this study, individuals with blood types B and O have more distinct reading preferences, but reliable conclusions cannot be drawn for blood types A and AB. This research suggests that the content used in reading therapy can be selected based on individuals' blood types, although scientific evidence to support this claim is not provided.

In [14], the preference for using digital content compared to printed books for reading therapy purposes was examined. This article investigates which of the mentioned media is preferred by individuals and can have stronger effects on the therapeutic process. The research showed that the use of printed books had a positive impact on the effectiveness of reading therapy and was preferred by individuals. On the other hand, the prioritization presented by individuals has a significant correlation with their education level and their use of internet tools.

In [15], a machine learning-based method for analyzing individuals' behavior and distinguishing between anxiety and depression is presented. In this study, the necessary data for diagnosing anxiety and depression were collected through responses to a set of questions from 125 participants. Decision tree was used to classify the data and differentiate between different levels of anxiety and depression. This method achieved an average accuracy of 71.44% in distinguishing between these two conditions, which does not show significant progress. In [16], a combined deep learning model for emotion detection through simultaneous processing of facial images and speech signals is presented. This model utilizes two CNNs with different architectures for processing facial expressions and speech signals. Feature maps are generated from images using the ResNet network, while feature points of speech are extracted using Mel-Frequency Cepstral Coefficients (MFCC) and a 2D CNN model. By merging the feature maps from these two models, the emotion detection task is performed.

In [17], an online system for measuring the level of anxiety using machine learning techniques is proposed. In this approach, the necessary data for anxiety assessment were collected through wearable sensors such as electrocardiogram, electrodermal, and respiration sensors. Then, various learning models such as K-nearest neighbors, SVM, naive Bayes, and ensemble models were used to identify the level of anxiety. The results showed that the ensemble technique can perform better than other learning models with a maximum accuracy of 89.9%. In [18], heart rate data from wearable sensors were used to predict mental health and determine anxiety and depression. This method employed an LSTM model to process the information of heart rate variations in the time and frequency domains. The research showed that using data from 5-minute and 2-minute time intervals, the detection of individuals' mental health can be achieved with accuracies of 83% and 73%, respectively.

In [19], a chatbot for evaluating mental health based on natural language processing (NLP) techniques and machine learning is presented. This research also organized a textual database for detecting different mental states. In this approach, the user's entered text is preprocessed, and with the help of NLP techniques, content-based features are extracted from the text. Then, different learning techniques such as logistic regression, naive Bayes, decision tree, and random forest are used for evaluating an individual's mental health. In [20], two-dimensional and three-dimensional CNN models are used for processing speech and facial images, respectively. The features from these models are combined using a mutual attention mechanism to identify emotional states.

In [21], a method for emotion detection based on facial and speech features is proposed. In this approach, facial feature maps are extracted using a CNN model and combined with speech features based on the Fisher Vector structure. Then, the resulting feature vectors are classified using a Gaussian Mixture Model, and finally, an SVM with a Radial Basis Function (RBF) kernel is used for emotion recognition. The method presented in [22] is based on the factorized bilinear pooling (FBP) for emotion recognition through simultaneous processing of facial expressions and speech. Initially, a fully connected convolutional network with an attention mechanism is used to identify local emotional states. Then, a global FBP approach is employed to integrate audiovisual features. To enhance this approach, an adaptive weighting mechanism based on attention mechanism is used to weight speech and image features. Emotion recognition is performed by a fully connected layer and a SoftMax layer.

In [23], GAN networks are used for emotion detection through simultaneous processing of facial expressions and speech. This approach addresses the problem of limited training samples for facial expression recognition.

III. PROPOSED MODEL

The accurate analysis of individuals' mood and emotions, the utilization of appropriate textual resources, and the effective design of personalized treatment programs are among the essential prerequisites for achieving an efficient reading therapy system. Additionally, a suitable reading therapy system should provide a high level of accessibility. In the proposed method, an attempt has been made to fulfill these prerequisites through the combination of deep learning and reinforcement learning techniques. The proposed model utilizes the internet as a communication platform for interacting with individuals. The structure of the proposed model is illustrated in Figure 1.

According to the structure depicted in Figure 1, in the proposed model, the internet acts as a communication interface between the user and the reading therapy model. The

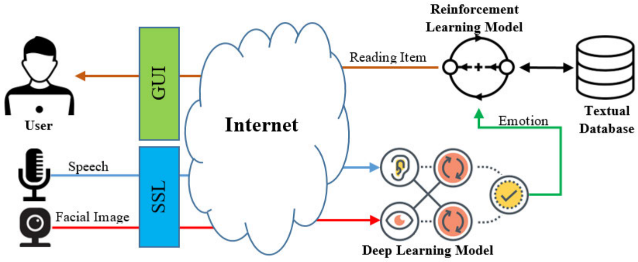

FIGURE 1. Structure of the proposed internet-based reading therapy model.

interaction between the user and the system is facilitated through a Graphical User Interface (GUI). During the execution of reading therapy exercises, the user's facial images and speech are recorded and transferred via the internet to a web server where the proposed reading therapy model is deployed. It's important to note that to protect user privacy, this information is transmitted using the SSL encryption protocol.

The user's speech signals and facial images are processed by the proposed deep learning model to determine the user's emotional states based on the combination of features extracted from these two signals. Subsequently, the output of the deep learning model is transferred to a reinforcement learning model. This reinforcement learning model utilizes a separate learning automaton for each emotional state to determine an optimal strategy for designing the reading therapy program based on the user's current state. The activated learning automata update their probability vector using reward and penalty operators and retrieve the next reading therapy exercise from the database. Finally, the selected item is presented to the user via the GUI. This process is repeated until the completion of the reading therapy program.

With these explanations, the proposed internet-based reading therapy model can be summarized in two phases.

1. Extraction of User Emotional States based on Deep Learning
2. Reinforcement Learning-based Reading therapy Program Design

The first phase of the proposed model utilizes a parallel combination of two CNN models to extract features from the user's speech and facial characteristics. These two feature sets are then merged and processed using a combination of LSTM layers to ultimately identify the user's emotional states through a classification layer.

In the second phase of the proposed model, a combination of reinforcement learning models is used for selecting reading therapy texts. For this purpose, a separate reinforcement learning model is designated for each identifiable emotional state. Upon identifying each emotional state, the corresponding reinforcement learning model in the proposed model becomes active, and the selection process of study items is performed based on the structured probability vector within the model. It should be noted that the proposed reinforcement learning model utilizes reward and penalty operators to update its probability vector after each change in the user's emotional states. This section continues with the details of each phase of the proposed model.

A. EXTRACTION OF USER EMOTIONAL STATES BASED ON DEEP LEARNING

In the proposed approach, a hybrid model based on deep learning techniques is used for the identification of user emotional states. This emotional detection system consists of two CNN models that simultaneously extract visual and auditory features of the user. These features are then combined using LSTM layers and classified by a SoftMax layer. The architecture of the proposed model for emotional state detection is presented in Figure 2.

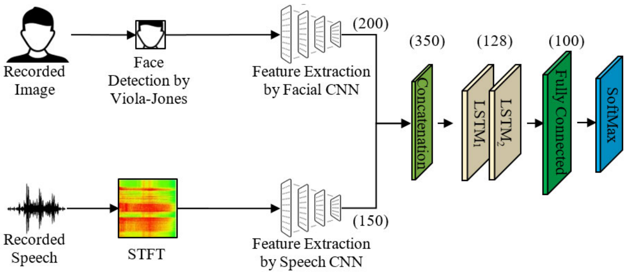

FIGURE 2. The architecture of proposed deep learning model for emotion recognition based on audio-visual features.

According to Figure 2, each speech signal and captured facial image are processed by separate CNN models. The task of these CNN models is to extract features related to emotions from the facial images and speech signals. To extract emotional features from the facial images, the face region needs to be separated from the input image, for which the well-known Viola-Jones algorithm [24] is used. Additionally, to make the speech signals compatible with the CNN model, the Short-Time Fourier Transform (STFT) technique [25] is employed, which allows the description of each speech signal in a matrix form based on time and frequency. Each of these samples is processed by the corresponding CNN model, and their features are extracted in the form of a vector. The facial feature vector has dimensions of 200, and the speech feature vector has dimensions of 150. The structure of the CNN models used in the proposed method will be explained further. After feature extraction by the proposed CNN models, the feature vectors are merged using a concatenation layer with dimensions of 350. These features are then processed by two sequential LSTM layers with dimensions of 128. The output of $LSTM_2$ is transformed using a fully connected layer to a vector with dimensions of 100. Finally, the classification and detection of emotional states are performed using a SoftMax layer. In the following, the preprocessing of facial images and speech signals will be described, followed by the configuration of the CNN models for feature extraction from each of these signals.

1) PREPROCESSING OF FACIAL IMAGES

For the preprocessing of facial images, the image color system is first converted to grayscale. This is because color layers of images may not reflect useful information regarding the person's emotional states. Using color images alone would only increase the complexity of the recognition model. After converting the images to grayscale, the Viola-Jones algorithm is utilized to identify the facial region in the input images. The high speed and acceptable detection accuracy of this algorithm have made it widely used in various research studies, particularly for object detection tasks, including the recognition of facial regions. The Viola-Jones algorithm is a cascading algorithm that learns facial features through the application of the AdaBoost algorithm on a set of weak classifiers. Based on this, the cascading algorithm trains N binary classification models to determine the presence of a face in the input image. In the detection phase, the input image is sequentially passed through all the binary classifiers, and at each stage, if a classifier declares the absence of a face, the result will be the absence of a face. Otherwise, the facial region in the input image will be determined. The computational details of this algorithm are described in [24], and we will not delve into them here. Finally, the facial region is cropped from the image and transformed into a 100 × 100-pixel matrix. This matrix is then used as the input for the CNN model for facial image processing.

2) PREPROCESSING OF SPEECH SIGNALS

For the preprocessing of speech signals, each input sample is first converted into a mono channel signal. Then, the signal frequency is converted to a constant value of 100 Hz. Next, the STFT model is used to describe the features of each signal and form a spectrogram matrix. The STFT model is a suitable solution for representing signal variations in the time domain. It provides a suitable input for deep learning techniques such as CNN and is suitable for describing interpretable features that show signal differences in neighboring points. The spectrogram produced by this model is a time-frequency matrix obtained by normalizing the magnitude of the STFT squares of the signal. To form this matrix, the input signal is first divided into a set of segments of equal length using a windowing function. Then, the Fourier transform is computed for each segment. The STFT can be described as follows [25]:

$$STFT_{S[n]}(m, w) = \sum_{n=-\infty}^{+\infty} S[n] \times w[n-m] e^{-jwn} \quad (1)$$

In the above equation, $S[n]$ represents the normalized signal. Additionally, $w[m]$ denotes the windowing function centered at m. After calculating the STFT of the signal based on the above equation, the spectrogram of the signal can be obtained by squaring it [25].

$$SPG_{S[n]}(m, w) = \left|STFT_{S[n]}(m, w)\right|^2 \quad (2)$$

The result of the above equation is a spectrogram matrix that describes the temporal and frequency dimensions of the signal. In the proposed method, a 3-second interval of the speech signal is used. The resulting spectrogram matrix is then used as the input for the CNN model for the speech signal.

3) CONFIGURING CNN MODELS FOR AUDIO-VISUAL FEATURE EXTRACTION

The two CNN models employed in the proposed method utilize a similar structure for audio-visual feature extraction. They have the same number and combination of layers, but different configurations for each layer's parameters. The overall structure of these two CNN models is illustrated in Figure 3. According to this figure, each CNN model accepts the face region/spectrogram matrix of the speech signal through an input layer, with the dimensions determined based on the input sample. Therefore, for the CNN model of the speech signal, the input layer dimensions are 24 × 300, and for the CNN model of face images, the dimensions are 100 × 100. The input sample is processed by three consecutive convolution components to extract feature maps of the signal. Each convolution component consists of a 2D convolution layer, a ReLU layer as an activation function, and a pooling layer. In the CNN model of the speech signal, the pooling layer is of the Max pooling type, while in the CNN model of facial images, it is of the Average pooling type. Finally, the feature maps obtained from the last pooling layer in these two CNN models are transformed into vector form by fully connected (FC) layers. Thus, the weight vector generated by the FC2 layer is considered as the extracted features for each sample.

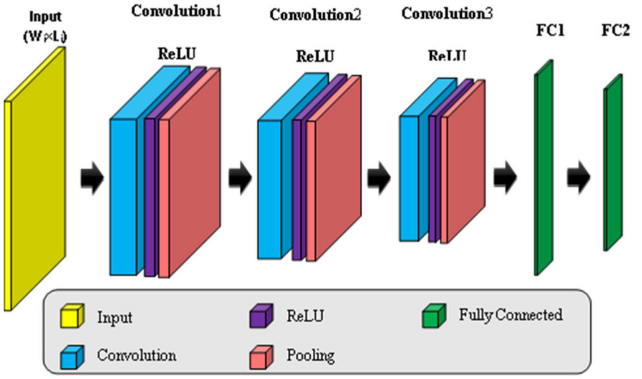

FIGURE 3. Structure of the CNN models used in the proposed method for facial and speech feature extraction.

As mentioned, each of the proposed CNN models for face and speech feature extraction has a different configuration for its layers. It should be noted that the configuration of these two CNN models was done using the BayesOpt tool [26], and the results are presented in Table 2. According to this table, generally, face images require a more complex CNN model to extract efficient features related to emotions, resulting in the need for more convolution filters with smaller dimensions. This is because facial image patterns are more complex compared to STFT matrices. On the other hand,

TABLE 2. Configurations set for each layer of the two CNN models for face and speech feature extraction.

Layer	Facial CNN Parameters	Speech CNN Parameters
Input	100×100	24×300
Convolution1 (W×H×N)	7×7×64	8×8×32
Pooling1	Average pooling	Max pooling (2×2)
Convolution2 (W×H×N)	5×5×32	5×5×32
Pooling2	Average pooling	Max pooling (2×2)
Convolution3 (W×H×N)	3×3×32	4×4×16
Pooling3	Average pooling	Max pooling (2×2)
FC1	500	400
FC2	200	150

face images require a larger number of features for more accurate description compared to speech signals. Therefore, the dimensions of each extracted feature vector from face images will have a length of 200, which is 50 features more than the descriptor vectors of speech signals.

4) DESIGNING A REINFORCEMENT LEARNING-BASED READING THERAPY PROGRAM

In the second phase of the proposed model for internet-based reading therapy, the output vector from the deep model (SoftMax layer in Figure 2) representing the recognized emotional states is used as the input for the reinforcement learning model. It is important to note that a person's emotional states can be accompanied by high uncertainty, and in some cases, their mood may be a combination of two (or even multiple) basic emotional states. Therefore, in the proposed model, two emotional states with higher weights are considered as the input for the reinforcement learning model, where the order of states is significant.

The proposed reinforcement learning model is a combination of learning automata models, where each learning automata corresponds to a binary combination of emotional states. Thus, if the number of basic emotional states is N, the proposed reinforcement learning model consists of K = N × (N-1) learning automats models, each responsible for designing the study program based on the currently identified states. This structure is illustrated in Figure 4.

In Figure 4, the set of recognizable emotional states is depicted in the upper part of the image. The output of the CNN model for an input sample is a pair of emotional states, where the first and second elements represent the first and second identified states, respectively. In Figure 4, it is assumed that the CNN model has generated the output pair (Excited, Depressed) for a hypothetical sample. In this case, the corresponding learning automata model is activated, and it is responsible for the process of selecting a study item to be presented to the user.

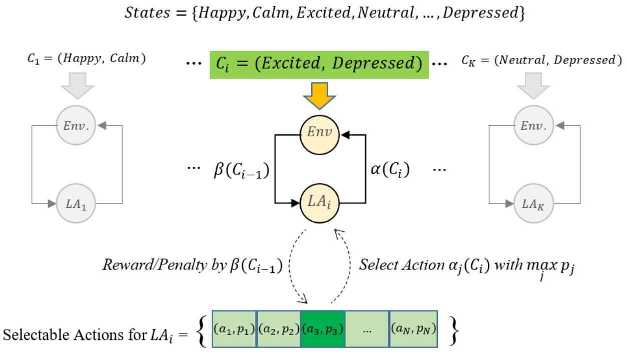
$States = \{Happy, Calm, Excited, Neutral, \ldots, Depressed\}$
$C_1 = (Happy,\ Calm)$
$\cdots$ $C_i = (Excited,\ Depressed)$ $\cdots$
$C_K = (Neutral,\ Depressed)$
$Env.$
LA_1
Env
$\cdots$ $\beta(C_{i-1})$
$\alpha(C_i)$ $\cdots$
LA_i
$Env.$
LA_K
Reward/Penalty by $\beta(C_{i-1})$
Select Action $\alpha_j(C_i)$ with $\max_j p_j$
Selectable Actions for LA_i = $\{ (a_1, p_1) (a_2, p_2) (a_3, p_3) \ldots (a_N, p_N) \}$

FIGURE 4. The structure of the proposed reinforcement learning model for planning the reading therapy program based on the identified emotional states.

In the proposed method, each learning automata is characterized by two sets: the set of actions and the set of probabilities. The set of actions for each learning automata includes the available study items (texts) for presentation to the user. This set is represented as $A = \{a_1, a_2, \ldots, a_N\}$, where N represents the size of the set of actions. Additionally, in the learning automata model, a probability value is assigned to each action, representing the probability of selecting that item. Thus, the set of probabilities for each learning automata is described as $P = \{p_1, p_2, \ldots, p_N\}$. At the start of the system for a new user's reading therapy, all actions in each learning automata have equal probability values of $\frac{1}{N}$, and the selection process is performed randomly. As the reading therapy program progresses, the probability vector is updated, and the selection process is based on the maximum probability value in the automata.

In order to continuously improve the choices in a reinforcement learning model, interaction with the environment is essential. Each learning automata model monitors the continuous changes in the user's mood and updates its probability vector accordingly. In Figure 4, the changes in the user's mood are considered as the environment's response and represented as $\beta\,(C_{i-1})$. In the proposed reinforcement learning model, after each selection and presentation to the user, the changes in the user's mood compared to the previous session are evaluated, and based on reward and punishment operators, the probability vector of the selecting automata is updated. To do this, different mood states are ranked as follows: 1- Tense, 2- Angry, 3- Frustrated, 4- Depressed, 5- Bored, 6- Tired, 7- Neutral, 8- Calm, 9- Relaxed, 10- Content, 11- Happy, 12- Delighted, and 13- Excited. The desired state of the proposed reinforcement learning model is to determine a reading therapy strategy that steers the user's mood towards emotional states with higher scores. If we denote the average score of the top two identified emotional states for the user in the i-th iteration as S_i, then the reward and penalty processes in the learning automata models are as follows:

Reward: If $(S_i > S_{i-1})$ or $(S_i = S_{i-1}, S_i \geq 7)$ it means that the recent action taken by the reinforcement learning model has resulted in positive changes or maintaining the desirable state in the user's mood. Therefore, the selecting automata responsible for this action updates its probability vector using the following equation [27]:

$$p_j^{k+1} = \begin{cases} p_j^k + a[1 - p_j^k] & j = i, \\ (1 - a)\, p_j^k & \forall j \neq i. \end{cases} \quad (3)$$

where in the above equation, i represents the index of the recent selected action (which resulted in positive changes in the user's mood). The values p_j^{k+1} and $\mathrm{p}_\mathrm{j}^\mathrm{k}$ represent the probability of action j in the new iteration and the previous iteration, respectively. Additionally, a represents the reward parameter, which is set to 0.5.

Penalty: If $(S_i < S_{i-1})$ or $(S_i = S_{i-1}, S_i < 7)$, it means that the recent action taken by the reinforcement learning model has resulted in negative changes in the user's mood. Therefore, the selecting automata responsible for this action updates its probability vector using a penalty operator [27]:

$$p_j^{k+1} = \begin{cases} (1 - b)p_j^k & j = i, \\ \left(\frac{b}{N - 1}\right) + (1 - b)\, p_j^k & \forall j \neq i. \end{cases} \quad (4)$$

where in the above equation, i represents the recently selected action (which resulted in negative changes in the user's mood). Additionally, b represents the penalty parameter, which is set to 0.5.

It should be noted that to prevent the transfer of repetitive content to the user, after selecting an item by a learning automata, that item will be set aside for T cycles. The process of detecting the user's emotional states and selecting items by the reinforcement learning model will be repeated sequentially until the end of the therapy program. It is also worth mentioning that the learned automata models obtained at the end of each session for each user are stored and used in the next session of the therapy program.

IV. IMPLEMENTATION AND RESULTS

The proposed method was implemented and evaluated using MATLAB 2019a software. After implementation, the proposed deep model was developed using the MATLAB Compiler and executed on a server. The RECOLA database [28] was used to train the proposed deep model for mood detection in users. Additionally, 10 participants were used to evaluate the performance of the proposed method in the reading therapy.

A. EXPERIMENTAL SCENARIOS

The experiments in this research were conducted in two phases. In the first phase, the performance of the proposed deep model in detecting the moods of the evaluated individuals was assessed. For this purpose, the proposed model was first trained using samples from the RECOLA database. This database contains audio and facial images of 46 participants recorded under different mood conditions. Six individuals were used to label the emotional states in the samples of this database. The labeling process was based on two emotional dimensions: Arousal (ranging from low to high) and Valence (ranging from negative to positive). Since the output of the proposed deep model consists of 13 predefined emotional states, in order to align the output with the Arousal and Valence dimensions in the RECOLA database, each emotional state was described as an ordinal variable indicating different degrees of these two dimensions. Figure 5 illustrates the mapping of target labels in the proposed model to the Arousal and Valence dimensions in the RECOLA database.

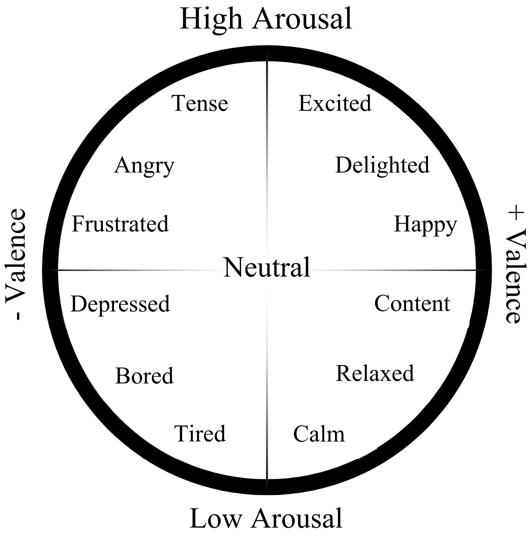

FIGURE 5. Mapping target labels in the proposed model to the arousal and valence dimensions in the RECOLA database.

Using the mapping model depicted in Figure 5, the output of the proposed deep model can be mapped to the Arousal and Valence dimensions for identifying individuals' emotional states. After determining the output of the proposed method based on these dimensions, the performance of the proposed method is evaluated using various evaluation metrics. For this purpose, the proposed model is first trained using 27 training samples from the RECOLA database, and the remaining samples are used for testing its performance.

The second phase of the experiments evaluates the effectiveness of the proposed internet-based reading therapy system in improving the mood of 20 volunteers. In this phase, each participant has utilized the proposed internet-based reading therapy system for 10 sessions. Each session has a minimum duration of 30 minutes, and the session's end time is determined by the user. The results of all sessions are reviewed by an expert. In this phase, changes in the user's mood from the beginning to the end of each session are evaluated, and the results are validated by the expert.

B. PERFORMANCE OF THE PROPOSED DEEP MODEL IN MOOD RECOGNITION

Following the scenario described in section IV-I, the proposed deep learning model was trained using 27 samples from the RECOLA database, and 9 validation samples were used to evaluate its performance. Each video sample contains at least 10 significant mood changes that occur at different time intervals. In each of these time intervals, the mood states identified by the proposed method are compared to the ground truth states, and the performance of the proposed model in mood recognition is evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Concordance Coefficient Correlation (CCC) metrics. Additionally, the results of the proposed hybrid model are compared to those obtained by using only the speech CNN models or the face CNN models for mood recognition. Furthermore, to assess the effectiveness of the proposed model, its performance is compared to the methods presented in [16] and [20].

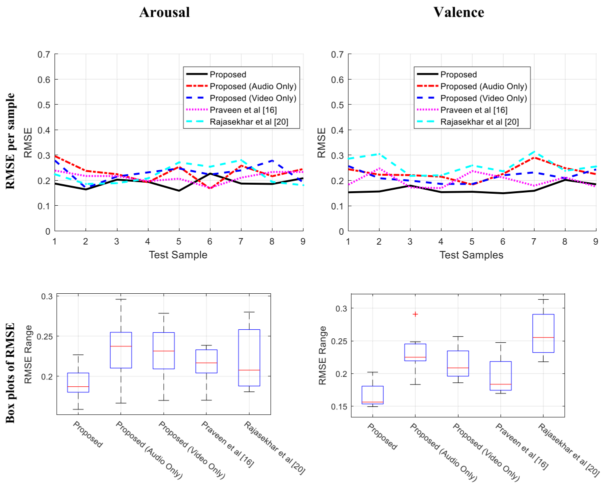

FIGURE 6. Evaluation results of different methods based on the RMSE metric for Arousal (left column) and Valence (right column) in terms of individual samples (first row) and error range (second row).

Figure 6 illustrates the evaluation results of the proposed method based on the RMSE metric for the Arousal and Valence dimensions. In this graph, the RMSE metric is calculated based on the following equation [29]:

$$RMSE = \sqrt{\frac{1}{N}\sum\nolimits_{i=1}^{N}(t_i - p_i)^2} \tag{5}$$

In the above equation, N represents the number of test samples (distinct recognition instances), and t_i and p_i represent the actual and predicted values for the Arousal and Valence dimensions, respectively. In the first row of Figure 6, the errors of different methods in identifying mood changes are presented for each sample. Since each video sample contains multiple changes in the Arousal and Valence dimensions, the average RMSE values are calculated for each sample. Figure 6 shows that the proposed method is capable of more accurate identification of mood changes in most samples. Furthermore, comparing the proposed method with approaches that rely solely on speech or image features for mood recognition indicates the significant superiority of the proposed method. Based on these results, it can be concluded that the combination of speech and facial features in the proposed model has been effective in reducing recognition errors. On the other hand, the second row of Figure 6 represents the range of RMSE variations for all test samples. These plots divide the range of RMSE variations into 4 quartiles, and the middle line of each box represents the median RMSE. The narrower range of RMSE variations and the lower values compared to other methods indicate that the outputs of the proposed method have higher reliability.

Figure 7, shows the performance of different methods in identifying individuals' mood based on the MAE metric. This figure also presents the errors in mood detection for Arousal and Valence dimensions. The MAE metric is calculated using the following formula [29]:

$$MAE = \frac{1}{N}\sum\nolimits_{i=1}^{N}|t_i - p_i| \tag{6}$$

Based on the MAE metric, it is possible to determine how much each method deviates from the actual positive or negative mood of individuals. By examining the plotted graphs in Figure 7, the findings extracted from Figure 6 can be confirmed. These results indicate that the proposed method, by simultaneously analyzing speech and facial features, can predict individuals' mood with lower prediction errors. In the Arousal dimension, the RMSE of the proposed method is 0.1905, and the MAE is 0.1570. The closest performance to the proposed method is achieved by Praveen et al. [16], with RMSE and MAE values of 0.2136 and 0.1748, respectively, in the Arousal dimension. The order of performance of different methods in the Valence dimension is also maintained similarly. Based on the obtained results, the proposed method can reduce the RMSE and MAE metrics by at least 16.51% and 18.7%, respectively. These results confirm the desirable performance of the proposed method in identifying individuals' mood. This superiority of the proposed method can be attributed to the parallel CNN model and the integrated feature processing through LSTM models.

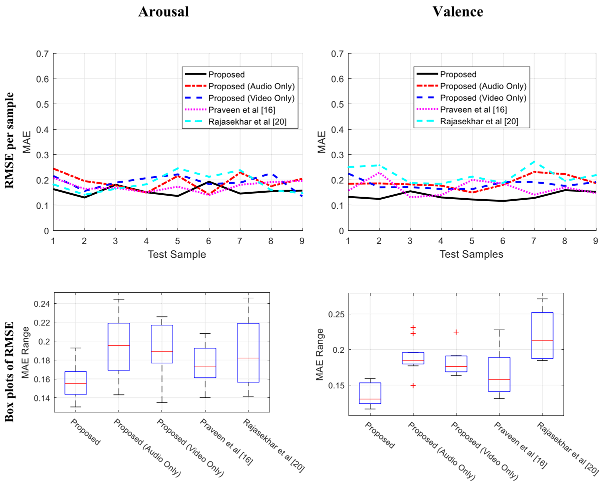

FIGURE 7. Evaluation results of different methods based on the MAE metric for the Arousal aspect (left column) and Valence aspect (right column), separated by sample (first row) and based on the range of absolute error changes (second row).

The comparison of the performance of different methods in identifying the Arousal and Valence dimensions indicates that almost all methods perform more accurately in identifying individuals' mood in the Arousal aspect. This is because distinguishing between different emotional states in the Arousal dimension is easier compared to the Valence dimension, resulting in lower prediction errors.

For a more precise comparison of the performance of different methods in identifying various dimensions of individuals' mood, the Taylor diagram can be used. This diagram demonstrates how close the predictions of each method are to the actual values. The Taylor diagram simultaneously considers Pearson correlation, standard deviation, and root mean square deviation (RMSD) between the model predictions and the actual degree of individuals' mood. Therefore, it provides a better insight for interpreting the results in a more accurate manner. The Taylor diagrams obtained from the identification of various dimensions of individuals' mood using different methods are shown in Figure 8.

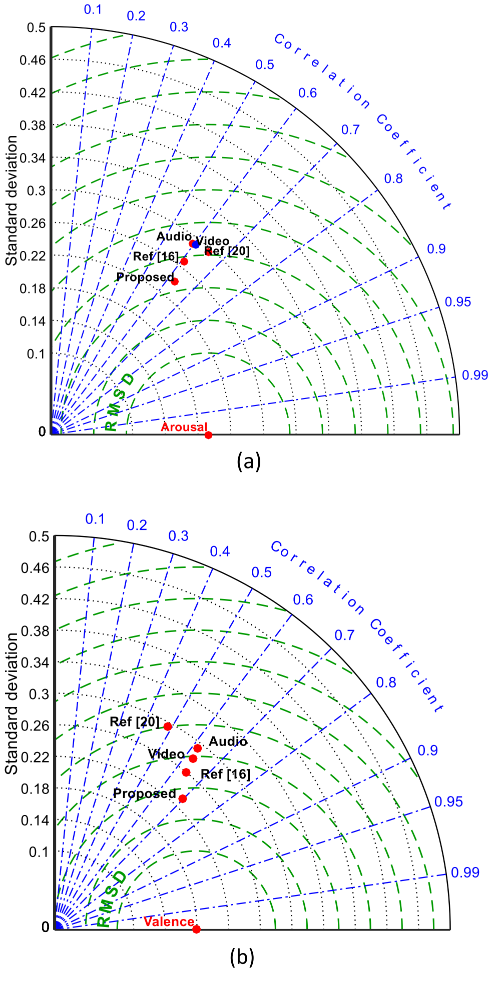

FIGURE 8. Taylor diagrams resulting from the identification of individuals' mood in the (a) Arousal and (b) Valence dimensions.

Based on Figure 8, the predictions made by the proposed method (fusion of speech and image) have a lower standard deviation compared to the actual values of individuals' mood. Additionally, the lower RMSD value of the proposed method compared to other methods indicates its closer proximity and better alignment with the actual values of individuals' mood. Furthermore, the higher correlation of the outputs of the proposed method with the actual values of individuals' mood indicates that this method is capable of more accurately modeling the changes in individuals' mood and provides a more precise estimation compared to other approaches. Based on this figure, the superiority of the proposed method over other methods for both the Arousal and Valence dimensions is clearly observable, confirming its satisfactory performance in identifying individuals' mood.

In Figure 9, regression plots of the proposed method and other compared approaches for predicting individuals' mood in the Arousal dimension are depicted. Similarly, Figure 10 represents these results for the Valence dimension. In these plots, the actual degree of mood is shown on the horizontal axis, and the predicted values by each method are shown on the vertical axis. Based on the presented results in these figures, the values of $R_A = 0.63096$ and $R_V = 69746$ for the Arousal and Valence dimensions in the proposed method indicate a higher correlation between the outputs of this method and the target variable. The regression points for the proposed method in both cases have a higher correlation with the points on the $Y = T$ axis. Therefore, it can be concluded that the outputs of the proposed method exhibit a higher level of alignment with the actual values of individuals' mood. As mentioned, the reason behind this is the better performance of the proposed parallel CNN model in processing speech and facial features simultaneously and the utilization of LSTM layers to provide a more accurate approximation of mood changes over time.

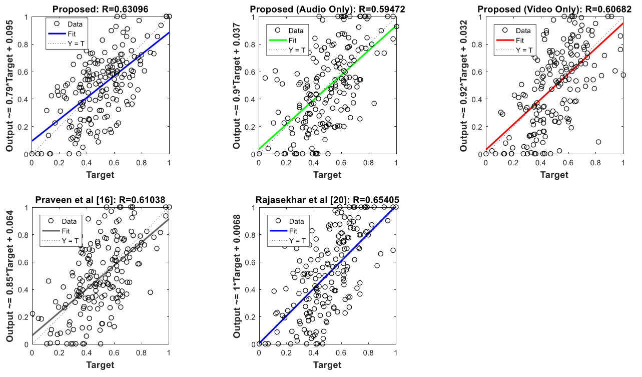

FIGURE 9. Regression plots of the proposed method and other approaches for predicting individuals' mood in the Arousal dimension.

Table 3 provides a summary of the results obtained from the experiments conducted in this section. In this table, in addition to the RMSE and MAE metrics, the performance of the methods is also compared based on the CCC metric. This metric describes the level of correlation coefficient

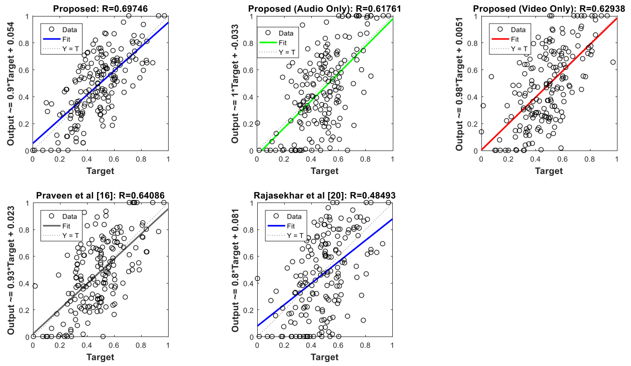

FIGURE 10. Regression plots of the proposed method and other approaches for predicting individuals' mood in the valence dimension.

TABLE 3. Numerical values resulting from the evaluation of mood detection methods based on the RECOLA database.

Methods	Valence				Arousal			
	CCC	MAE	RMSE	R	CCC	MAE	RMSE	R
Proposed Method	0.6751	0.1356	0.1658	0.6975	0.6146	0.1570	0.1905	0.6309
Proposed (Audio only)	0.5461	0.1888	0.2307	0.6176	0.5465	0.1927	0.2321	0.5947
Proposed (Video only)	0.5722	0.1824	0.2156	0.6294	0.5571	0.1914	0.2304	0.6068
Praveen et al [16]	0.5979	0.1667	0.1986	0.6409	0.5776	0.1748	0.2136	0.6104
Rajasekhar et al [20]	0.4293	0.2183	0.2589	0.4849	0.5979	0.1860	0.2206	0.6541

agreement between the output and the actual values and is introduced as a standard metric in the RECOLA database.

$$CCC\,(T,P) = \frac{2 \times Corr(T,P) \times \sigma_T \times \sigma_P}{\sigma_T^2 + \sigma_P^2 + (\mu_T - \mu_P)^2} \tag{7}$$

In the above equation, T and P represent the vectors of actual and predicted mood values, and $Corr(T,P)$ describes the Pearson correlation coefficient between these two vectors. σ_T and μ_T represent the standard deviation and mean of the T values, respectively.

The results presented in Table 1 confirm the superiority of the proposed method in mood detection. The higher CCC values in the proposed method indicate that the predictions provided for the Arousal and Valence mood coefficients have both a higher correlation with the actual values and a smaller deviation from these values. As a result, the approximations provided by the proposed method are more accurate compared to the compared methods.

C. PERFORMANCE OF THE PROPOSED METHOD IN INTERNET READING THERAPY

In the second phase of evaluating the performance of the proposed method, its effectiveness in improving individuals' mood through the implementation of internet-based therapy programs was examined. Following the procedure described in section IV-I, this phase of the experiments involved the participation of 20 volunteers in the internet-based therapy study based on the proposed model. Each volunteer participated in 10 therapy sessions for a minimum of 30 minutes, and changes in their mood were monitored from the beginning to the end of the sessions. It is worth mentioning that in order to eliminate the effects of users' unfamiliarity with the system environment, the initial 5 minutes of each session were disregarded. Additionally, all sessions were reviewed by an expert, and any discrepancies related to individuals' mood were corrected.

In Figure 11, the average changes in participants' mood during the 10 therapy sessions implemented by the proposed model are shown. The mood changes are presented based on the following scoring: 1- Tense, 2- Angry, 3- Frustrated, 4- Depressed, 5- Bored, 6- Tired, 7- Neutral, 8- Calm, 9- Relaxed, 10- Content, 11- Happy, 12- Delighted, and 13- Excited. Additionally, to better display the effectiveness of each session, the mood changes at the beginning and end of each session are provided. It should be noted that the values presented for each session are calculated based on the average scores extracted from all volunteers.

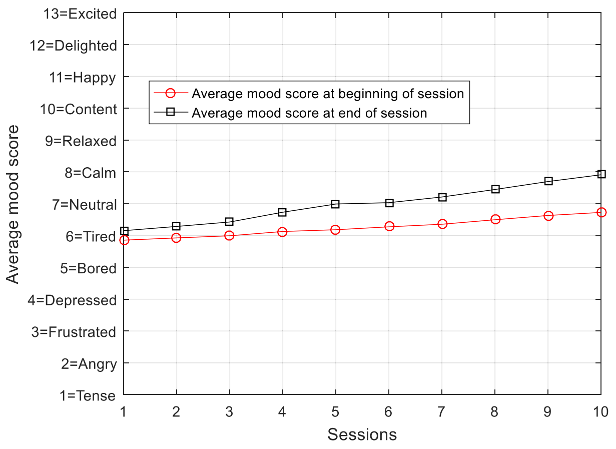

FIGURE 11. Average changes in participants' mood during the 10 therapy sessions implemented by the proposed model.

Analyzing the presented graph in Figure 11 reveals several key features regarding the proposed method. The comparison of the average mood scores at the beginning and end of the sessions demonstrates that, on average, participants had a better mood state at the end of each session compared to the beginning. Therefore, it can be concluded that the proposed method had a positive and tangible effect on individuals' mood in each session. Furthermore, over time, the internet-based therapy approach has been more beneficial for individuals, as indicated by the increasing difference between the average mood scores at the beginning and end of the sessions. This upward trend can be attributed to two main factors. Firstly, over time, individuals become more adapted to the internet-based therapy environment and benefit more from its advantages. Secondly, the reinforcement learning approach requires modeling users' mood through reward and punishment operators to achieve optimal therapy strategies, which can only be accomplished through time and the selection of an initial set of actions. Another notable inference from Figure 11 is the continuous improvement of individuals' mood during the sessions. On average, each participant's mood at the beginning and end of each session shows significant improvement, indicating a shift from a passive state to an inclined state of enthusiasm in the later sessions. The validity of this claim can be confirmed based on the results presented in Figure 12, which shows the average duration of internet-based therapy sessions.

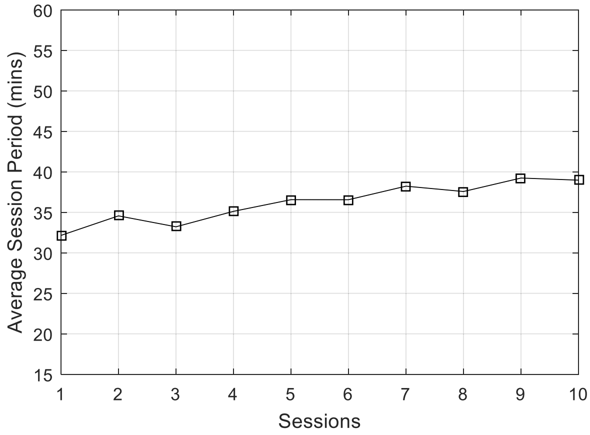

FIGURE 12. **Average duration of internet-based therapy sessions.**

As mentioned, the duration of each session is determined by the user, allowing the evaluation of their enthusiasm to participate in the therapy sessions. Figure 12 illustrates that as the therapy sessions progress, the average duration of the sessions increases. In other words, over time, the participants' enthusiasm for attending the therapy sessions has grown. This can be attributed to the effective performance of the reinforcement learning strategy in the proposed method. This strategy, utilizing multiple learning automata models and based on reward and penalty operators, has successfully contributed to the continuous improvement of individuals' mental states and avoidance of erroneous therapeutic strategies.

V. CONCLUSION

An automated system for online therapeutic interventions can be effective in a wide range of therapeutic and rehabilitation applications, such as improving mood, reducing anxiety, treating depression, and addressing dyslexia. In this paper, an AI-based automated system for online therapeutic interventions was proposed. The proposed method utilizes a parallel combination of CNN models to analyze the audio-visual emotional states of individuals. This hybrid model combines the extracted features from each CNN model and, after processing by LSTM layers, identifies the emotional states of the individual using a classification layer. The results showed that the combination of speech and facial features (compared to using speech or facial features alone) can reduce the average correlation coefficient (CCC) in emotional state recognition by at least 14.2%. Furthermore, this strategy achieved an improvement of at least 9.71% compared to previous studies, with an average CCC of 0.64485. The proposed method utilizes reinforcement learning models to design therapeutic exercises based on the individual's emotional states. This reinforcement model employs reward and punishment operators to continuously improve the individual's emotional well-being and avoid ineffective therapeutic strategies. The effectiveness of the proposed method in determining the effectiveness of therapeutic interventions was studied with the participation of 20 participants. The research results showed that the proposed approach was effective in improving mood and increasing user engagement in using the therapeutic intervention programs, and its performance was competitive with a supervisor-based approach. These results indicate that automated online therapeutic interventions can be utilized in real-world scenarios.

One of the limitations of the current research was considering a relatively small population for evaluating the performance of the proposed method in internet reading therapy. More computation time is another limitation in the proposed method which is the result of utilizing multiple learning models. In future work, these limitations can be addressed by developing cost-efficient computational methods and considering a larger papulation. In future work, the proposed model can be adapted for specific applications in therapy interventions, such as addressing speech problems in individuals. Additionally, since mood states can be subject to high uncertainty, combining the proposed model with a fuzzy model can be explored to improve the model's flexibility.

REFERENCES

[1] X. Wang, L. Jia, and Y. Jin, "Reading amount and reading strategy as mediators of the effects of intrinsic and extrinsic reading motivation on reading achievement," *Frontiers Psychol.*, vol. 11, Oct. 2020, Art. no. 586346.

[2] D. Francis, J. L. Hudson, S. Kohnen, L. Mobach, and G. M. McArthur, "The effect of an integrated reading and anxiety intervention for poor readers with anxiety," *PeerJ*, vol. 9, Feb. 2021, Art. no. e10987.

[3] E. M. E. Koopman, "Effects of 'literariness' on emotions and on empathy and reflection after reading," *Psychol. Aesthetics, Creativity, Arts*, vol. 10, no. 1, p. 82, 2016.

[4] J. Khongtim, "Trends in reading habits of students from school level to higher levels of education: Evidence from the review of literature," *Library Philosophy Pract.*, 2021, Art. no. 5170.

[5] R. D. Goodwin, L. C. Dierker, M. Wu, S. Galea, C. W. Hoven, and A. H. Weinberger, "Trends in U.S. depression prevalence from 2015 to 2020: The widening treatment gap," *Amer. J. Preventive Med.*, vol. 63, no. 5, pp. 726–733, Nov. 2022.

[6] C. Kelly, "Improving empathy of occupational therapy students through reading literary narratives," *J. Occupational Therapy Educ.*, vol. 6, no. 4, p. 4, Jan. 2022.

[7] J. Hart, "Bibliotherapy: Improving patient's health through reading," *Alternative Complementary Therapies*, vol. 27, no. 6, pp. 298–300, Dec. 2021.

[8] A. R. Khan, "Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges," *Information*, vol. 13, no. 6, p. 268, May 2022.

[9] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski, "A survey on facial emotion recognition techniques: A state-of-the-art literature review," *Inf. Sci.*, vol. 582, pp. 593–617, Jan. 2022.

[10] A. Nogales, Á. J. García-Tejedor, D. Monge, J. S. Vara, and C. Antón, "A survey of deep learning models in medical therapeutic areas," *Artif. Intell. Med.*, vol. 112, Feb. 2021, Art. no. 102020.

[11] S. Thelijjagoda, M. Chandrasiri, D. Hewathudalla, P. Ranasinghe, and I. Wickramanayake, "The hope: An interactive mobile solution to overcome the writing, reading and speaking weaknesses of dyslexia," in *Proc. 14th Int. Conf. Comput. Sci. Educ. (ICCSE)*, Aug. 2019, pp. 808–813.

[12] D. A. Salwerowicz, "Design proposal for a software tool for speech therapy," M.S. thesis, Mod. Appl. Struct. Vis. Speech Therapy App Children, UiT Norges Arktiske Universitet, Tromsø, Norway, 2019.

[13] S. Agustina, W. S. W. M. Saman, N. Shaifuddin, and R. A. Aziz, "Reading material selection for bibliotheraphy based on blood type in young adult groups," *Jurnal Kajian Informasi Perpustakaan*, vol. 10, no. 1, pp. 89–106, 2022.

[14] R. De Jesús-Romero, A. Wasil, and L. Lorenzo-Luaces, "Willingness to use Internet-based versus bibliotherapy interventions in a representative U.S. sample: Cross-sectional survey study," *JMIR Formative Res.*, vol. 6, no. 8, Aug. 2022, Art. no. e39508.

[15] T. Richter, B. Fishbain, A. Markus, G. Richter-Levin, and H. Okon-Singer, "Using machine learning-based analysis for behavioral differentiation between anxiety and depression," *Sci. Rep.*, vol. 10, no. 1, Oct. 2020, Art. no. 16381.

[16] R. G. Praveen, W. C. de Melo, N. Ullah, H. Aslam, O. Zeeshan, T. Denorme, M. Pedersoli, A. L. Koerich, S. Bacon, P. Cardinal, and E. Grange, "A joint cross-attention model for audio-visual fusion in dimensional emotion recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2486–2495.

[17] F. R. Ihmig, H. A. Gogeascoechea, F. Neurohr-Parakenings, S. K. Schäfer, J. Lass-Hennemann, and T. Michael, "On-line anxiety level detection from biosignals: Machine learning based on a randomized controlled trial with spider-fearful individuals," *PLoS ONE*, vol. 15, no. 6, Jun. 2020, Art. no. e0231517.

[18] L. V. Coutts, D. Plans, A. W. Brown, and J. Collomosse, "Deep learning with wearable based heart rate variability for prediction of mental and general health," *J. Biomed. Informat.*, vol. 112, Dec. 2020, Art. no. 103610.

[19] N. Ghoshal, V. Bhartia, B. K. Tripathy, and A. Tripathy, "Chatbot for mental health diagnosis using NLP and deep learning," in *Advances in Distributed Computing and Machine Learning*. Singapore: Springer, 2023, pp. 465–475.

[20] R. G. Praveen, E. Granger, and P. Cardinal, "Cross attentional audio-visual fusion for dimensional emotion recognition," in *Proc. 16th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Dec. 2021, pp. 1–8.

[21] E. Ghaleb, M. Popa, and S. Asteriadis, "Metric learning-based multimodal audio-visual emotion recognition," *IEEE MultimediaMag.*, vol. 27, no. 1, pp. 37–48, Jan. 2020.

[22] H. Zhou, J. Du, Y. Zhang, Q. Wang, Q.-F. Liu, and C.-H. Lee, "Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2617–2629, 2021.

[23] F. Ma, Y. Li, S. Ni, S.-L. Huang, and L. Zhang, "Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional GAN," *Appl. Sci.*, vol. 12, no. 1, p. 527, Jan. 2022.

[24] M. M. Hussein and A. H. Mutlag, "Face detection methods: A comparative study between viola-jones and skin color detection," *J. Eng. Appl. Sci.*, vol. 14, no. 14, pp. 4754–4760, Dec. 2019.

[25] J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. DAFx*, vol. 10, Sep. 2010, pp. 397–403.

[26] R. Martinez-Cantin, "BayesOpt: A Bayesian optimization library for nonlinear optimization, experimental design and bandits," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3735–3739, 2014.

[27] B. Anari, J. A. Torkestani, and A. M. Rahmani, "A learning automata-based clustering algorithm using ant swarm intelligence," *Expert Syst.*, vol. 35, no. 6, p. e12310, Dec. 2018.

[28] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–8.

[29] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not," *Geoscientific Model Develop.*, vol. 15, no. 14, pp. 5481–5487, Jul. 2022.

JIANXIN XIONG was born in Yueyang, Hunan, China, in 1977. He received the master's degree in computer technology from Hunan University, in 2008, and the Ph.D. degree in management, in 2021. He is currently a Computer Teacher with the Hunan University of Arts and Sciences. His main research interests include information system construction and digital library.

HUI YIN was born in Shaoyang, Hunan, China, in 1975. She received the master's degree in computer technology from Yunnan University. She is currently a Teacher with the Computer Teaching and Research Section, Hunan University of Medicine. Her research interests include education, computer image processing, and video tracking.

MEISEN PAN was born in 1972. He received the degree from Hunan Normal University, China, in 1995, the M.S. degree from the Huazhong University of Science and Technology, China, in 2005, and the Ph.D. degree from Central South University, China, in 2011. He is currently a Professor with the College of Computer and Electrical Engineering, Hunan University of Arts and Science. He has published more than 40 papers on journals and conferences. His research interests include biomedical image processing, information fusion, and software engineering.