Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

A Reinforcement Learning-Based Approach for Promoting Mental Health Using Multimodal Emotion Recognition

Amod Pathirana [1,*], Dumidu Kasun Rajakaruna [1], Dharshana Kasthurirathna [2], Ajantha Atukorale [1], Rekha Aththidiye [3], and Maheshi Yatipansalawa [1]

[1] University of Colombo School of Computing (UCSC), Colombo, Sri Lanka; e-mail: amd@ucsc.cmb.ac.lk; 2018is064@stu.ucsc.cmb.ac.lk; aja@ucsc.cmb.ac.lk; 2018is097@stu.ucsc.cmb.ac.lk
[2] Sri Lanka Institute of Information Technology, Sri Lanka; e-mail: dka@ucsc.cmb.ac.lk
[3] District Health Board Bay of Plenty, New Zealand; e-mail: rekha.aththidiye@gmail.com
* Corresponding Author : Amod Pathirana

Abstract: This research aims to enhance mental well-being by addressing symptoms of anxiety and depression through a personalized, culturally specific multimodal emotion prediction system. It employs an emotionally aware Reinforcement Learning (RL) agent to suggest tailored Cognitive Behavioral Therapy (CBT) activities. The study focuses on developing precise, individualized emotion prediction models using facial expressions, vocal tones, and text, and integrates these models with the RL agent for emotionally aware CBT recommendations. The mHealth approach combines deep learning models with RL, achieving accuracies of 72% for facial expressions, 73% for vocal tones, and 86% for text, all fine-tuned for the Sri Lankan context. Validation through real-world use and user feedback consistently demonstrated that each model exceeds 70% accuracy, fulfilling the objective of precise emotion prediction. A weighted algorithm was introduced to refine the emotion prediction experience and personalize forecasts across the three modalities to enhance mental well-being. The RL-enabled agent suggests CBT activities approved by mental health professionals, tailored based on predicted emotions, and delivered through the same mHealth application. The effectiveness of these interventions was assessed using the DASS-21 questionnaire, revealing significant reductions in depression scores (from 21.08 to 13.54) and anxiety scores (from 19.85 to 10.46) in the study group compared to the control group. The study concludes that integrating multimodal emotion prediction models with RL-based CBT suggestions positively impacts mental well-being and contributes to personalized mental health interventions.

Keywords: Anxiety and depression symptoms; Cognitive behavioral therapy; DASS-21 questionnaire; Multimodal emotion prediction; Reinforcement learning.

Received: August, 18th 2024
Revised: September, 8th 2024
Accepted: September, 14th 2024
Published: September, 17th 2024
Curr. Ver.: September, 17th 2024

1. Introduction

Mental health disorders, such as anxiety and depression, pose a significant global challenge, affecting millions of individuals. According to the World Health Organization (WHO), in 2019, approximately 970 million people lived with mental disorders, including 301 million with anxiety and 280 million with depression [1]. In the U.S., 19.1% of adults suffered from anxiety disorders in 2023[2], and 50% of adults aged 18-24 experienced symptoms of anxiety and depression, exacerbated by the COVID-19 pandemic, which saw a 25.6% increase in anxiety and a 27.6% rise in depression globally [3], [4]. Despite these significant statistics, mental health care remains particularly challenging in low and middle-income countries (LMICs). Over 80% of mental disorders are found in LMICs, yet 75% of cases remain untreated due to limited resources and social stigmas[5], [6]. By 2030, depression is expected to be a leading cause of disease burden in these regions[7]. The complexity of identifying mental illness, influenced by changing external and internal factors from adolescence onward, complicates effective treatment[8].

Early intervention for mental health disorders is critical but often constrained by traditional methods that rely on basic technologies and generic solutions, which are also costly and inaccessible. This limitation impacts the effectiveness and availability of personalized care. Recent advancements in artificial intelligence (AI) and machine learning (ML) have introduced new opportunities for improving mental health care, particularly through Deep Learning (DL) and Reinforcement Learning (RL). DL allows for more accurate emotion monitoring by analyzing multimodal data, such as facial expressions, vocal tone, and text sentiment, providing a richer understanding of emotional states than outdated methods. Conversely, RL offers the potential for personalizing mental health care by optimizing therapeutic recommendations based on user interactions. Despite these advancements, current mental health interventions are often generic and lack real-time adaptability. The integration of DL and RL in mental health care is still in its early stages, with many applications not fully exploiting these technologies. This underscores the urgent need for innovative, adaptive solutions that can deliver real-time, personalized interventions to address mental health challenges more effectively.

This study addresses the research gaps by focusing on two main phases. The first phase involves identifying each individual's most effective emotion prediction modality, whether facial expression, vocal tone, or vocal text, to create a precise, individualized emotion prediction mechanism. The second phase evaluates the efficacy of CBT activity suggestions to promote positive emotional states and mitigate negative ones. This evaluation determines how well an RL-enabled agent can use predicted emotions to recommend appropriate CBT and other therapeutic activities to improve mental well-being.

These two phases are focusing on two primary objectives. The first objective is to devise a precise, individualized emotion prediction mechanism that accurately forecasts each individual's emotions across three modalities: facial expression, vocal tone, and vocal text. By creating a personalized emotion prediction model for each individual, this objective seeks to advance emotionallyaware predictive models on a broader scale. The second objective is to develop an RL-enabled, emotionally-aware activity suggestion agent that uses the predicted emotions of individuals to recommend suitable CBT and other activities. This aim is to enhance the mental well-being of individuals showing symptoms of anxiety and depression.

In pursuit of these objectives, the study addresses two central research questions: Given the variation in people's emotions across the three modalities, how accurately can personalized emotion recognition systems predict an individual's specific emotions, and which modality is most effective for each individual? Additionally, can an RL-enabled agent effectively offer emotionally aware CBT and other activity recommendations based on individuals' predicted emotions to improve their mental well-being, specifically considering symptoms of anxiety and depression?

The research contributes in three significant ways: First, it develops a personalized and culturally specific multimodal emotion recognition and monitoring mechanism. Second, it creates an emotionally-aware RL agent for CBT activity suggestions based on identified emotional states. Third, it provides an empirical evaluation of the proposed approach, utilizing the DASS-21 intervention to investigate anxiety and depression symptoms within the context of Sri Lanka. This comprehensive approach aims to enhance mental well-being, particularly among young people who are highly vulnerable to anxiety and depression.

2. Literature Review

Anxiety disorders, characterized by excessive fear and worry, lead to physical symptoms like heart palpitations and sweating[9], along with cognitive symptoms such as racing thoughts and catastrophic thinking[10]. Depression affects mood and well-being, with symptoms including prolonged sadness, loss of interest in activities, appetite changes, and sleep disruptions [11]. Negative attitudes towards one's emotions are linked to the severity of depression[12]. Both anxiety and depression symptoms significantly alter human emotions.

Research in psychology demonstrates humans' ability to identify basic emotions through facial expressions, known as the "six universal expressions" [13]. This ability is crucial for analyzing mental conditions, as individuals with anxiety and depression often exhibit emotional variations. Depressed individuals typically show fewer animated expressions, more negative or neutral expressions, and specific vocal traits like lower pitch variability and increased vocal tension[14], [15]. Anxiety produces distinct facial expressions such as widened eyes,

raised eyebrows, and a tense mouth, with longer speech pauses reflecting uncertainty[16]. Capturing these facial and vocal characteristics provides insights into emotional states, aiding in diagnosis and treatment planning for anxiety and depression symptoms.

Emotional expression is conveyed through facial expressions, vocal tones, and text, each with unique characteristics. Facial expressions communicate both basic and culturally specific emotions[17]. Vocal tones convey emotions like sarcasm or excitement through variations in pitch, volume, and rhythm[18]. By combining various lexical, semantic, syntactic, pragmatic, and contextual features, researchers have developed increasingly sophisticated models for automatically detecting and classifying emotions expressed in text[19], [20].

Facial expression recognition employs Convolutional Neural Networks (CNNs) such as VGGNet, which are noted for their feature-learning capabilities from image feeds[21], [22]. VGG16 models achieve test accuracies ranging from 70% to 90% on standard facial expression datasets[23]. However, challenges arise in context-specific and cross-cultural scenarios, as highlighted by Dr. Rachael Jack's research on differing facial features focusing on East Asians and Western Caucasians[24], [25]. These cultural differences underscore the need for customized facial emotion recognition models, particularly in regions like Sri Lanka, lacking such models.

Emotion prediction can be achieved through vocal tone and text analysis. Vocal tone, characterized by pitch, volume, and timbre, provides essential emotional cues beyond words, classifying emotions like happiness, sadness, anger, fear, and neutrality[26], [27]. Research indicates that depressed individuals show shorter middle vowel durations and distinct vowel area variations[28]. Speech Emotion Recognition (SER) methods, including SMO, MLP, and Logistic Regression, have been effective, with CNN-based approaches demonstrating high accuracy. Mustaqeem et al. reported accuracies of up to 79.5% with RAVDESS and 81.75% with IEMOCAP [29], while another method achieved accuracies of 86.1%, 96.3%, and 91.7% using RAVDESS, Berlin (Emo-DB), and SAVEE datasets, respectively[30].

Text analysis for emotion prediction has also seen significant progress. Techniques like Word2Vec and GloVe are widely used in NLP applications[31], [32]. Studies using SVM and Naive Bayes for classifying emotions in tweets achieved 71.4% accuracy[33], while deep learning approaches, such as deep learning-assisted semantic text analysis (DLSTA), offer improved accuracy[34].

Recent advancements in multimodal emotion recognition systems have demonstrated significant improvements in accuracy by integrating speech and text data. For instance, Cai et al. (2019) combined Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to analyze acoustic emotion features alongside Bidirectional LSTMs for textual features. Their approach yielded a notable 6.70% increase in overall recognition accuracy and a substantial 13.85% improvement, specifically in vocal tone emotion recognition compared to single modality methods[35].

However, despite these advancements, practical challenges in emotion prediction persist due to cultural and individual variations. Matsumoto (2012) highlighted that facial expressions can vary in intensity and cultural context, influencing recognition accuracy[36]. Moreover, the context of emotional expression, such as how a smile differs in social versus professional settings and individual differences in personality and emotional intelligence, further complicates accurate emotion identification[37]. Therefore, developing personalized models that can effectively accommodate these variations is crucial for advancing the field.

Recent research underscores the dual nature of emotions, which encompass both positive and negative elements and vary across cultures and contexts[38]. CBT effectively alters negative thought patterns and promotes positive thinking[39]. Medications, psychotherapy, and lifestyle changes also support mental well-being[40], [41]. CBT techniques, including refocusing attention and engaging in positive activities, are particularly effective in managing anxiety and depression[42], [43].

Emotion prediction through facial expressions, vocal tone, and text analysis provides valuable insights into mental health conditions, as discussed previously. Building on this, emotions associated with anxiety and depression can be treated using CBT through tailored interventions. Reinforcement learning, especially Q-learning, presents a promising approach. Q-learning enables an agent to optimize decision-making by learning from rewards and punishments, focusing on actions based on current states without requiring a predefined model [44]–[46]. This method is well-suited for suggesting personalized CBT activities for individuals with anxiety and depression. Despite its potential, a review of 1,228 PubMed articles revealed

that, although AI/ML-enabled mental health approaches use self-reported data for mood prediction, none have utilized reinforcement learning for personalized CBT recommendations[47].

A seamless interaction mechanism is crucial to providing effective CBT activities and other recommendations. Existing applications like "Depression CBT Self Help Guide" and "eMoods" offer CBT guidance and dynamic mood tracking. However, they lack a multimodal approach for emotion prediction and do not use RL-based agents[48]. This highlights a significant research gap in mobile health (mHealth) approaches for delivering emotionally aware CBT activities to address anxiety and depression symptoms.

A controlled experiment using the DASS-21 questionnaire is appropriate for evaluating the proposed intervention. The DASS-21 measures depression, anxiety, and stress across five severity levels but does not provide formal diagnoses. This questionnaire is validated across various cultures[49]. Previous studies utilizing the DASS-21 include one employing VGG16, which achieved accuracies of 87.2% for depression, 77.9% for anxiety, and 90.2% for stress using facial expressions[50]. Another study found logistic regression to be the most accurate among five machine learning classifiers, with classification accuracies of 90.33%, 92%, and 90.33% for depression, anxiety, and stress, respectively[51].

The literature review identifies several research gaps:

1. Single vs. Multimodal Approaches: Current methods mainly focus on single modalities (e.g., facial expressions or vocal tone) and often lack personalization across multiple modalities. Developing systems that integrate various emotional recognition modalities tailored to individual users is essential.
2. Reinforcement Learning: There is a notable gap in applying reinforcement learning to personalize CBT activity suggestions for anxiety and depression. Such systems could provide more tailored and effective interventions.
3. Evaluation Methodologies: There is a need for robust evaluation methodologies to assess the impact of multimodal interventions on mental health symptoms. Improved methods that combine traditional and innovative metrics, such as the DASS-21, are necessary.

We propose an integrated system that combines facial expression analysis, vocal tone assessment, and text-based emotion prediction to address these gaps, enhanced with RL. This approach aims to improve the accuracy of emotion prediction through personalized methods and provide emotion aware CBT activity suggestions. By leveraging RL, this comprehensive solution is designed to effectively alleviate symptoms of depression and anxiety through timely and personalized interventions.

3. Proposed Method

3.1. High-Level Architecture of the Solution

This study utilizes a multimodal emotion recognition approach, incorporating facial expressions, vocal tone, and vocal text analysis. Three neural network architectures were trained on secondary datasets and fine-tuned with a primary dataset specific to the Sri Lankan context. An RL agent uses the predicted negative emotions to suggest CBT activities tailored to reduce anxiety and depression symptoms.

Figure 1 illustrates the data capturing process, beginning with users recording videos using a mobile application. The input videos are then split into three second segments to capture rapid emotional transitions in the backend. Each segment's frames, audio clips, and transcribed text are processed through their respective emotion prediction models. The final emotion for each modality is determined by majority voting across all segments.

The three modality-based emotion predictions are unified through a weighted function that adjusts based on initial feedback, creating a single aggregated emotion state. The RL agent uses this state to suggest CBT activities. Feedback on these activities helps the RL agent refine its recommendations, eventually identifying suitable activities to promote mental health. An experiment comparing a study group receiving personalized, emotion aware CBT activities via the RL agent with a control group receiving random activities demonstrated the benefits of personalization in enhancing mental well-being.

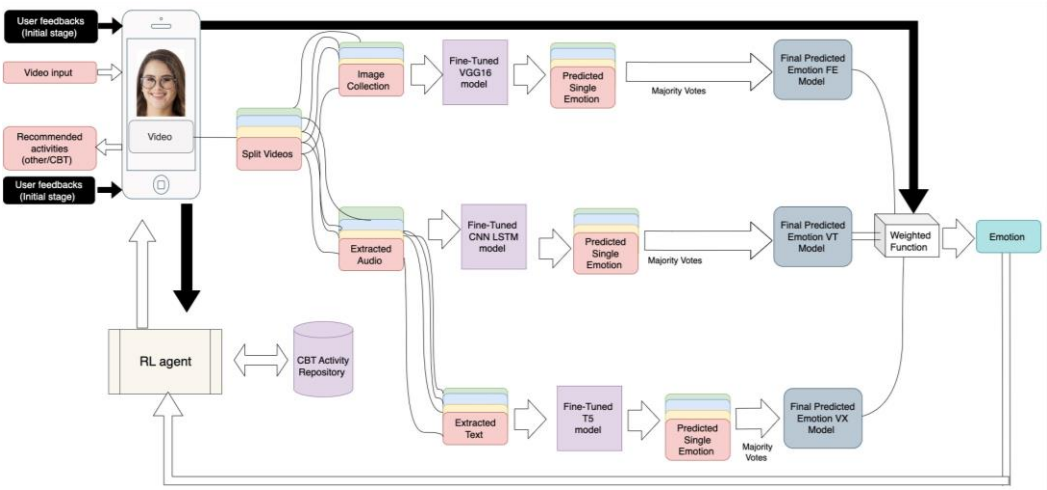

Figure 1. High-Level Architecture of the Solution

3.2. Data Collection

The study's data collection occurred in three stages. Initially, secondary data from trusted and standard sources were used to train the emotion recognition models. In the second stage, a primary dataset specific to the Sri Lankan context was created, featuring labeled images and voice utterances for transfer learning and model fine-tuning to Sri Lankan context. In the final stage, a controlled experiment was conducted, involving further data collection and analysis using the DASS-21 questionnaire to validate the approach.

Stage 1: The initial stage involved collecting secondary data from trusted sources to train the emotion recognition models. For facial emotion recognition, the study used the "2013 Facial Expression Recognition" (FER2013) dataset, which contains over 35,000 labeled facial images and is known for its standardized annotation, making it suitable for model training and evaluation. For speech emotion recognition, the study combined datasets from RAVDESS, SAVEE, TESS-Toronto, and CREMA-D, creating an enhanced dataset with 1,000 utterances for each of the five emotions under study. For text-based emotion recognition, a combined dataset from the "Emotions" and "Emotion-stimulus" datasets was created, encompassing around 20,000 data points with balanced labels for the five emotions.

Table 1. Distribution of Data Across Facial Expression, Vocal Tone, and Vocal Text Datasets for Each Emotion.

Datasets	Facial Expression	Vocal Tone				Vocal Text
	FER-2013	Ravdess	Savee	Tess	Crema-d	Emotions & Emotion stimulus
Angry	4953	384	60	40	516	3350
Happy	8989	384	60	40	516	5200
Sad	6077	384	60	40	516	4600
Neutral	6198	384	60	40	516	4100
Fearful	5121	384	60	40	516	3100
Total (Emotions)	31338	1920	300	200	2580	20350
			5000			

Stage 2: A primary dataset was created (facial expression, vocal tone) to address the issue of relying on foreign datasets, which may not accurately represent the Sri Lankan context when it comes to experiments conducted using young participants in Sri Lanka to evaluate the study. Cultural and demographic differences, as well as variations in facial expressions and vocal tones across cultures, can impact emotion recognition[25]. Therefore, due to the

absence of accessible Sri Lankan datasets, the study opted to create an appropriate primary dataset specifically tailored for research purposes. To fine-tune the model for facial emotion recognition using our primary dataset, over 100 videos featuring professional actors from Sri Lankan television series were collected. Emphasizing local actors helped capture the local context and ensure precise capturing of facial expressions since they are skilled in portraying emotions. Frames with prominent facial visibility and direct camera focus were extracted from permitted videos, isolating the facial regions and eliminating the background to form an image collection. Figure 2 (a) below provides a sample of this dataset. Then, the images were thoroughly checked and labeled by three independent people to guarantee accuracy using majority voting. The labels include five emotions considered in the study. The data were subsequently organized into separate folders according to the five labels, following a structure akin to the FER2013 dataset. A total of 961 images were labeled, as shown in Figure 3(b), with an average of 200 images per emotion, resulting in a balanced distribution.

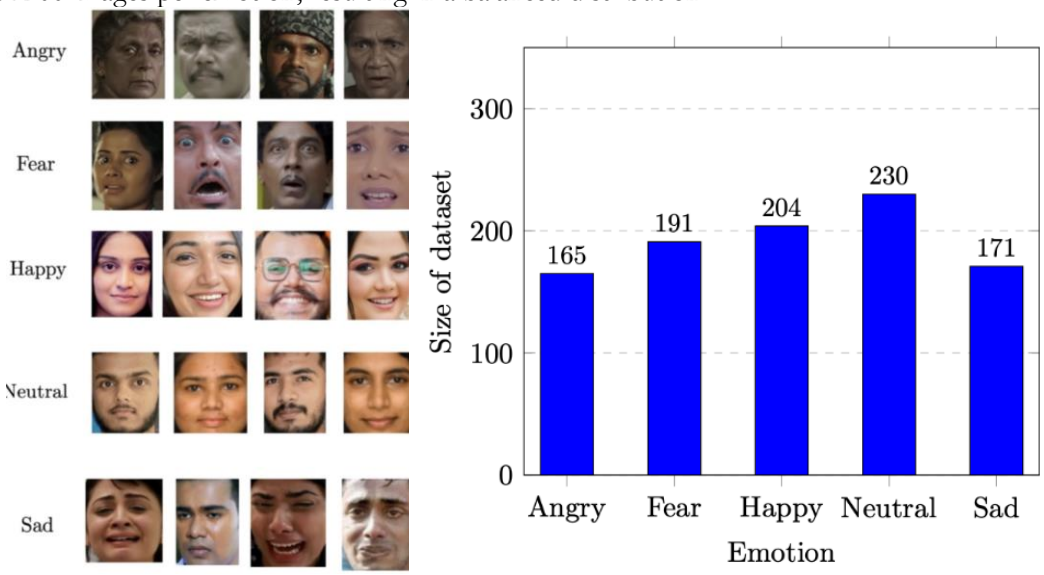

Figure 2. (**a**) Samples from Primary Facial Expression Data Set (**b**) Distribution of the Images in the Primary Facial Expression Dataset.

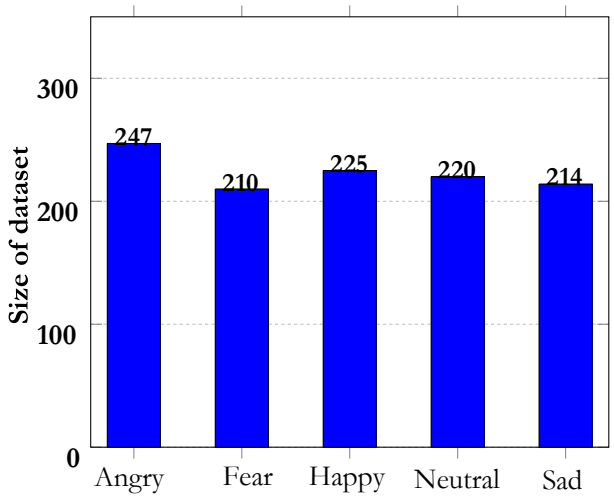

Figure 3. Distribution of the Audio Files in the Primary Vocal Tone Dataset

Another challenge was the prevalence of vocal recordings with foreign accents and contexts in the secondary dataset for speech emotion recognition. Given that the proposed solution aimed specifically at Sri Lankans, it was necessary to employ primary data-gathering methods to enhance accuracy. To fine-tune the model for speech emotion recognition, a sample population of youth was selected, as the study was conducted with a young

demographic, and a balanced dataset covering all five emotions was created, as illustrated in Figure 3. During the dataset creation, participants were asked to read three sentences designed to express five different emotions. The recordings were conducted in a controlled environment with minimal background noise, using clip microphones. A total of approximately 1,200 utterances were collected for all five emotions.

Stage 3: The final stage involved a controlled experiment where participants completed the DASS-21 questionnaire to assess their mental health status. The DASS-21 measures symptoms of depression, anxiety, and stress using 21 items rated on a scale of severity. The questionnaire was administered before and after the intervention to evaluate its impact on mental health outcomes. Ethical considerations were paramount: participants provided informed consent, and their privacy and confidentiality were strictly maintained. Each participant signed a consent form to uphold ethical standards throughout the study.

3.3. Experiment

This research employs a quantitative approach with convenience sampling, involving 30 volunteers. Participants were informed about the study's purpose, risks, and benefits and consented to ethical clearance. They completed the DASS-21 questionnaire via Google Forms to assess initial anxiety, depression, and stress levels.

Participants then downloaded and registered on a mobile app via a Google Play Store link. They were randomly assigned to the study or control groups, with 15 participants in each. The study group received emotionally aware CBT activities from an RL agent, while the control group received generic activities. This randomization minimized bias.

The intervention lasted two weeks, with feedback collected during the first ten iterations to adjust the weighted function and evaluate prediction accuracy. Each user's personalized modality (facial expressions, vocal tone, or vocal text) was identified after these iterations. At the end of the intervention, participants completed the DASS-21 questionnaire again. Of the 30 participants, 26 completed the study. The data collected will be used to analyze and evaluate the research approach further.

3.4. Mobile Application

The mobile application captures facial expressions and vocal information, providing personalized experiences through emotional recognition and CBT activity suggestions. Users record and upload videos via their devices, with the option to crop before submission. Deep learning models analyze these videos to predict emotions.

During the initial training phase, users confirm emotional predictions. If three consecutive negative emotions are detected, the app suggests CBT interventions. Users rate the effectiveness of these activities, training the reinforcement learning (RL) agent to tailor future suggestions. The RL model then continues to recommend activities based on ongoing emotional analysis. Developed with Flutter, the app integrates with Google Firebase for data exchange and connects to a Flask server backend for managing deep learning and RL interactions.

3.5. Multimodal Emotion Prediction

3.5.1. Facial Emotion Recognition

The deep learning technique employed in this study is Convolutional Neural Network (CNN), specifically VGG16 architecture. The CNN model is designed to detect and classify five key emotions (happy, sad, anger, fear, and neutral) based on facial expressions. Several preprocessing steps are implemented, including face detection and cropping, grayscale conversion, histogram equalization, and image augmentation.

The deep learning technique employed in this study is a Convolutional Neural Network (CNN), specifically the VGG16 architecture. The CNN model is designed to detect and classify five key emotions (happy, sad, anger, fear, and neutral) based on facial expressions. The VGG16 model comprises 13 convolutional layers with ReLU activation functions and three fully connected layers. The model was initialized with pre-trained weights from ImageNet, and the output layer was customized to suit the 5-emotion classification task. To enhance model performance and accuracy, several preprocessing steps are implemented:

- **Face Detection and Cropping:** Using the MediaPipe face detection SDK, based on the BlazeFace model, to detect and crop facial regions, isolating them for further processing.
- **Grayscale Conversion:** Applying OpenCV's cv2.cvtColor(image, cv2.COLOR BGR2GRAY) to simplify input data while retaining essential facial features.
- **Histogram Equalization:** Using OpenCV's cv2.equalizeHist(image) to normalize lighting conditions and enhance contrast.
- **Image Augmentation:** Utilizing Keras's ImageDataGenerator with the following parameters: The ImageDataGenerator is configured with the following parameters: rescale = 1./255 to normalize pixel values, validation split = 0.2 to split the dataset into training and validation sets, rotation range = 5 to apply random rotations up to 5 degrees, width shift range = 0.2 and height shift range = 0.2 to apply random shifts in width and height up to 20%, shear range = 0.2 to apply shear transformations, zoom range = 0.2 to apply random zoom, preprocessing function = preprocess fn for custom preprocessing, horizontal flip = True and vertical flip = True to apply random horizontal and vertical flips, and fill mode = 'nearest' to fill in new pixels during transformations.

The VGG16 architecture is used as a base model for transfer learning and model fine-tuning. The VGG16 model comprises 16 weight layers, 13 convolutional layers and 3 fully connected layers. It employs small 3x3 convolutional filters to capture intricate local patterns in the input images. Rectified Linear Units (ReLUs) are used as activation functions to enable faster training and convergence, while max-pooling layers reduce the spatial dimensions of feature maps. In this study, all layers are frozen except for the last four layers, which were fine-tuned to optimize the classification of the 5 target emotions. This allows for preserving pretrained features from the ImageNet dataset and facilitates adaptation for the emotion recognition task[21]. Additional layers are added, including a flattened layer and 4 dense layers, with the last layer utilizing the softmax activation function for the five emotions classification. The model is compiled with an adaptive moment estimation (Adam) optimizer using a $1 \times 10-5$ learning rate. It is trained for 300 epochs with a batch size of 128. The accuracy achieved by the FER2013 dataset is 76.21%. To bridge the cultural gap between application usage and the secondary dataset, the model is fine-tuned using the primary dataset. In this study, all layers in the VGG16 base model are frozen except for the last four, allowing the model to adapt to the specific task while leveraging pre-trained weights. The model is compiled using the Adam optimizer with a learning rate of 1×10^{-5} to minimize overfitting. Training is conducted for 100 epochs. To further prevent overfitting and enhance performance, the following callbacks are implemented: EarlyStopping (monitors validation loss and stops training if no improvement is detected), ModelCheckpoint (saves model weights at specific intervals), and ReduceLROnPlateau(automatically reduces the learning rate if validation loss plateaus). Hyperparameter tuning is performed to enhance performance, resulting in an accuracy of 72.58% on the primary (Sri Lankan) test dataset.

3.5.2. Speech Emotion Recognition

Recent research has shown the effectiveness of Convolutional Neural Networks (CNNs) for Speech Emotion Recognition (SER) due to their high accuracy. SER with CNNs requires converting audio signals into Mel-spectrograms. Preprocessing involves loading the audio dataset with librosa, applying z-normalization, setting the sample rate to 16kHz, and adjusting the signal length to 3 seconds.

The Mel-spectrograms are generated with the following configuration: a sample rate of 16,000 Hz, 512 FFT components, a window length of 256 samples (approximately 16 ms), a hop length of 128 samples (approximately 8 ms), 128 Mel bands, and a frequency range up to 4,000 Hz. Data augmentation enhances the dataset, such as pitch alteration and noise addition. Pitch is altered within a range of -5 to +5 semitones, and Gaussian noise with a mean of 0 and variance of 0.01 is added to the signals. Features are extracted by transforming audio signals into Mel-spectrograms using the Short-time Fourier Transform (STFT) squared magnitude. The secondary SER dataset is split 80:20 for training and testing, and early stopping is applied with patience of 5 epochs to prevent overfitting.

A hybrid CNN-LSTM architecture, including Local Feature Learning Blocks (LFLBs) with Conv2D, BatchNormalization, ReLU, MaxPooling2D, and Dropout layers, is implemented. The first LFLB starts with a Conv2D layer consisting of 64 filters with a kernel size of (2, 2) and a stride of (1, 1). The second LFLB has a Conv2D layer with 128 filters, a kernel

size of (5, 5), and a stride of (2, 2). The third and fourth LFLBs have Conv2D layers with 128 and 256 filters, respectively, with kernel sizes of (4, 4) and strides of (2, 2). L2 regularization is applied to all Conv2D layers to prevent overfitting. After the LFLBs, a Time-Distributed flattened layer reshapes the feature maps into a single dimension. An LSTM layer with 256 units and a dropout of 0.3 follows, capturing temporal dependencies in the input data. The model is compiled with the SGD optimizer, a 0.001 learning rate, and trained for 100 epochs with a batch size of 32, achieving an accuracy of 81.1% on the combined secondary dataset (RAVDESS, SAVEE, TESS-Toronto, and CREMA-D).

Fine-tuning was performed using the primary dataset to adapt the SER model for the Sri Lankan context. Given that our primary vocal tone dataset is not particularly large compared to the secondary combined datasets, we customized the architecture of the pre-trained model before applying transfer learning to achieve optimal performance. This involved reducing some complex layers and simplifying the model. The last two layers were removed and replaced with two Dense layers (ReLU activation) and two Dropout layers (rate 0.4). These changes improved feature extraction and helped prevent overfitting. A final Dense layer with softmax activation was added, and the model was compiled with the SGD optimizer. Fine-tuning achieved an accuracy of 73.26% on the primary (Sri Lankan) test dataset.

3.5.3. Text-based Emotion Recognition

The "Text-to-Text Transfer Transformer" (T5) is used for text emotion classification, and it is a pretrained transformer-based language model developed by Google[52]. In the text emotion recognition task, the secondary datasets (Emotions and Emotion-stimulus, Table 1) are preprocessed using tokenization, encoding, and padding/truncation techniques with the T5Tokenizer provided by the T5 model. For the preprocessing pipeline, the following steps are applied:

- Tokenization: In natural language processing, tokenization is the process of breaking down a sentence or text into individual words or subwords, called tokens. The T5Tokenizer is used for this purpose, which employs a subword-based approach capable of handling out-ofvocabulary words and reducing vocabulary size. This tokenizer breaks words into smaller units called subwords. tokenizer = T5Tokenizer.from pretrained('t5-base')
- Encoding: After tokenization, the text input is encoded into numerical format. The T5Tokenizer uses byte-pair encoding (BPE) to convert tokens into integers. BPE replaces commonly occurring character sequences with single symbols, reducing vocabulary size and improving processing efficiency.
- Padding and Truncation: Input texts vary in length, so padding (adding zeros to the end) and truncation (cutting off excess text) techniques are used to convert texts into a fixed length. The T5Tokenizer handles padding and truncation automatically, accommodating texts of various lengths.

Transfer learning is employed to fine-tune the pre-trained T5 model. Transfer learning leverages pre-trained models to adapt them for new tasks. This study uses transfer learning to finetune the T5 model for text-based emotion recognition, retaining knowledge from the pre-training phase. The fine-tuning process involves using the following hyperparameters: max seq length = 512, learning rate = 0.0003, weight decay = 0.0, adam epsilon = 1e-08, warmup steps= 0, train batch size = 8, eval batch size = 8, num train epochs = 2, gradient accumulation steps = 16. The preprocessed data is then used to fine-tune the T5-base model through transfer learning for text-based emotion recognition. The model was trained using only the preprocessed secondary data, achieving a test accuracy of 85.51%. This indicates that the fine-tuned T5 model did well in recognizing emotions from text-based inputs.

3.6. Weighted Function

A weighted function is designed and implemented to unify the predictions from the three models and implement a personalized emotion prediction mechanism. This function utilizes weights for each modality and adjusts them based on user feedback provided for predictions. By doing so, the study can identify the most appropriate model to consider when predicting emotions for each individual. The weighted function allows for the combination of outputs from all three models while prioritizing the highest-scoring model, which is personalized for each user based on their specific emotional expression. User feedback plays a crucial role within the weighted function, acting as a reward that enhances the model's accuracy over

time. As the function learns from user feedback, it becomes more effective in predicting emotions based on the user's unique characteristics in emotion expressions. The generalized mathematical equation for the weighted function is as follows:

For each instance $i \in \{1, \ldots, n\}$

1. Compute the maximum emotion with the highest confidence for each model in the i-th instance using the Equation (1).

$$\begin{aligned} E_F(i) &= \max(F) \cdot w_1, \\ E_{V_t}(i) &= \max(V_t) \cdot w_2, \\ E_{V_x}(i) &= \max(V_x) \cdot w_3 \end{aligned} \tag{1}$$

E represents the predicted emotion for each model: facial expression (F), vocal tone (V_t), and vocal text (V_x), weighted by w_1, w_2, and w_3 respectively. These weights are adjusted iteratively based on user feedback, starting from default values set according to the models' fine-tuned accuracy.

2. Identify the maximum weighted emotion score, $E_{\max}(i)$, and its corresponding category, $C_{\max}(i)$, for the i-th instance using Equations (2) and (3).

$$E_{\max}(i) = \max\left(E_F(i), E_{V_t}(i), E_{V_x}(i)\right) \tag{2}$$

$$C_{\max}(i) = \operatorname{argmax}\left(E_F(i), E_{V_t}(i), E_{V_x}(i)\right) \tag{3}$$

3. Obtain user feedback, $F(i)$, for the i-th instance and update the weights as Equation (4). If the predicted emotion $C_{\max}(i)$ is equal to the user feedback $F(i)$ update the weight of the relevant model by adding R reward value. $w_k(i)$ is the weight of the model k (F, V_t, V_x) at instance i.

$$w(i)_k = w(i-1)_k + R[F(i) = C_{\max}(i)] \tag{4}$$

Else if user feedback, $F(i)$ is equal to one of the highest percentage valued emotions predicted by another model $\left(E_F(i), E_{V_t}(i), E_{V_x}(i)\right)$, update the weight of that model by adding R reward value using Equations (5).

$$w(i)_k = w(i-1)_k + R\left[F(i) \neq C_{\max}(i) \wedge F(i) = \operatorname{argmax}\left(E_F(i), E_{V_t}(i), E_{V_x}(i)\right)\right] \tag{5}$$

Else (No model predicted the user feedback emotion $F(i)$ as their highest percentage valued emotion $\left(E_F(i), E_{V_t}(i), E_{V_x}(i)\right)$, Check the model that has the highest probability (P) value for the user feedback emotion $F(i)$ and update the weight of that model by adding $R/2$ reward value using Equations (6).

$$\begin{aligned} w(i)_k = w(i-1)_k + \frac{R}{2}\Big[&F(i) \neq C_{\max}(i) \wedge F(i) \neq \operatorname{argmax}\left(E_F(i), E_{V_t}(i), E_{V_x}(i)\right) \wedge \\ &F(i) = \max\left(E_F(i), E_{V_t}(i), E_{V_x}(i)\right)\Big] \end{aligned} \tag{6}$$

For this study, the default weights w_i are set based on the Fine-tuned accuracy of each model. These weights are then adjusted during each user's first 10 iterations of emotion prediction, incorporating user feedback. This adjustment uses a weighted function to provide the user a more personalized emotion prediction experience.

3.7. Reinforcement Learning Agent

In order to recommend CBT activities based on users' predicted emotional status, a RL based approach is utilized. The algorithm follows the Q-learning framework, where the agent learns to select appropriate activities by estimating the expected cumulative reward for each state-action pair. This implementation aims to adapt and improve the recommendation process over time based on user feedback. A Q-table is defined for each user to store state action values to implement the Q-learning algorithm. Hyperparameters such as learning rate, discount factor, exploration rate, and exploration decay are set. The Q-table is updated over time through the learning process. The three negative emotions considered in the study (sadness,

anger, and fear) are defined as states. The learning rate determines how much the Q-value should update based on new experiences, while the discount factor influences the reward value based on feedback. The exploration rate helps balance exploration (random activity selection) and exploitation (choosing the best known action). Exploration decay determines how the exploration rate changes over time. Reward scaling is applied to make the reward more significant and aid in faster learning.

The RL process is triggered upon the user's corrected emotion. The agent suggests an activity based on the emotional status and the current exploration rate. If the user completes the activity, they are asked to rate it on a scale of 1 to 10. The values in the Q-table are updated based on the reward, using the learning rate and reward scaling. This feedback was collected at the initial stage, and over time, the values in the Q-table are adjusted with rewards, considering the change from negative to positive emotional states. To ensure accurate fine-tuning, feedback was taken at random instances.

A database stores the basic activities to be suggested as a collection, which is common for all users. Different sets of activities are associated with each negative emotion. The Q-table and exploration rate update is personalized so that each user can provide emotionally aware activity suggestions.

3.8. CBT Activity Selection

In this study, we focus on providing healthy interventions for negative emotions and exclude CBT activity suggestions for when users are happy and neutral. We consider a subset of 5 out of the seven emotions introduced by Palu Ekaman[13]. Throughout our research process, we have collected a set of Cognitive Behavioral Therapy activities relevant to our study. The activities we have collected for this study have undergone organization and approval by clinical mental healthcare professionals who serve as external advisors in our research process. Their expertise ensures the appropriateness and relevance of these activities in the context of our study.

4. Results and Discussion

This section's primary focus is evaluating two research questions, which are subsequently followed by two research objectives. These objectives consist of several sub-objectives that require thorough evaluation to examine the insights and draw conclusions.

4.1. Model Performance and Generalizability

This section discusses quantitative results for multi-model emotion prediction (facial expression, vocal tone, vocal text) using secondary and primary datasets. Evaluation includes accuracy, precision, recall, and F1 score to address research objectives on having an accurate emotion recognition system to predict emotions in Sri Lankan context.

4.1.1. Assessing Emotion Recognition in Facial Expressions

Transfer Learning with VGG16 Using FER2013 Secondary Data Set: The VGG16 model is used with some modifications to existing architecture and compiled and trained using specific hyperparameter values as described in the 3.5.1 section. After training, the model achieves good performance when evaluated with test data. The comparison Table 2 shows that the proposed model performs well compared to other models trained using the FER2013 dataset and the VGG16 pretrained model combination.

Table 2. Comparison of Facial Emotion Recognition Models with FER2013 dataset.

Study	Model Architecture	Accuracy
Bodavarapu and Srinivas[53]	Custom VGG16 (FERConvNet)	65%
Kusuma et al. [54]	VGG16 + GAP	69.40%
Proposed study	Enhanced VGG16	76.21%

The proposed model performs well on the FER2013 dataset, outperforming other models in previous studies. Specifically, it achieves an accuracy of 76.21% with a precision of 0.76, a recall of 0.74, and an overall F1-score of approximately 0.75. These metrics are significantly

better than the accuracies of 65% and 69.40% obtained by other models. Transfer learning with VGG16 proves to be effective in facial emotion recognition.

Sri Lankan Specific Emotion Recognition with VGG16 Using Primary Data Set: The previously trained VGG16 model with the FER2013 dataset was adjusted by freezing certain layers and fine-tuning it with the primary dataset in this study. The fine-tuned model achieved an accuracy of around 72%, precision of 0.72, recall of 0.62, and F1 score of 0.72 on the primary dataset, indicating its ability to learn context-specific features. Although there was a slight decrease in accuracy compared to the VGG16 model with the FER2013 dataset, it can still be concluded that the model will perform well in the specific context of this study. Therefore, the facial emotion recognition model is well suited for utilization and future analyses of Sri Lankan specific emotion recognition. This study serves as a foundational framework for further exploration and research in this area.

4.1.2 Assessing Emotion Recognition in Vocal Tone

CNN-LSTM Model with Combined Data Sets (RAVDESS, SAVEE, TESS, CREMA-D): The CNNLSTM model, after hyperparameter tuning, achieved favorable results on combined datasets. It demonstrated an 81.10% accuracy with a precision of 0.8188, a recall of 0.8110, and an F1-score of 0.8108. The model effectively distinguished emotional nuances in speech, performing well across various emotions. It showed strong performance in detecting anger (0.87 F1 score), fear (0.78 F1 score), and neutral speech (0.87 F1 score). The model also accurately recognized happiness (0.81 F1 score) and sadness (0.73 F1 score), showcasing its versatility in identifying a wide range of emotions.

Table 3. Accuracy and F1 Score of Different SER Models on Various Combined Datasets (R for RAVEDSs, S for SAVEE, T for TESS, C for CREMA-D).

SER Model	R	S	T	C	Accuracy	F1 Score
MDP[55]	✓	✓	✓	✓	53.25%	0.53
SB[56]	✓	✓	✓	✓	59.35%	0.59
SEDC[56]	✓	✓	✓	✓	61.59%	0.62
Proposed Model	✓	✓	✓	✓	81.10%	0.81

Combined datasets were used for validation and testing to ensure a robust evaluation and generalizability of results on diverse data sets without limiting to a single data set. The proposed model outperformed current state-of-the-art models in terms of accuracy and F1 score, confirming its suitability and reliability in vocal tone emotion recognition across diverse datasets. Future evaluations should include context specific scenarios and larger datasets to validate the effectiveness and applicability of the model.

Sri Lankan Specific Emotion Recognition with CNN-LSTM Model Using Primary Data Set: A Sri Lankan dataset was created, and the pretrained CNN-LSTM model was customized to suit the Sri Lankan specific emotion recognition better. Modifications included removing layers, adding new layers, and increasing the dropout rate. These changes and transfer learning significantly improved the model's emotion prediction performance for the Sri Lankan context.

The Fine-tuned model achieved an accuracy of 73.26%, precision of 0.758, recall of 0.7326, and F1 score of 0.7376 on the primary dataset. Model showed a strong prediction for anger (precision: 0.85) and neutral (precision: 0.86) emotions and good performance for happiness (precision: 0.70) and sadness (precision: 0.77). Weaker predictions for fear (precision: 0.55) indicate room for improvement in future work. Considering the absence of Sri Lankan specific emotion prediction based on vocal tones and the unavailability of a publicly accessible dataset specific to Sri Lanka, the model's performance can be deemed acceptable.

4.1.3 Assessing Emotion Recognition in Vocal Text

T5 Model with Combined Data Sets (Emotion, Emotion-Stimulus): The fine-tuned T5 model achieved promising results in text-based emotion analysis, with an accuracy of 86%, a precision of 0.79, a recall of 0.83, and an F1-score of 0.80. A comparison with state-of-the-art models is shown in Table 4.

The proposed approach demonstrates acceptable accuracy compared to previous state-of-the-art models. Even though it may not be the most accurate, its performance is considered satisfactory for this study. Since the text (English) data analysis is not specific to the Sri

Lankan context, the model was trained only on a combined secondary dataset. The model's overall performance is enhanced by selecting a diverse dataset that includes a wide range of data.

Table 4. Comparison of T5-Based Text Emotion Analysis with State-of-the-Art Models.

Model	Dataset	Accuracy	Reference
T5 (fine-tuning)	Amazon review	0.84	Model, 2021[57]
T5 (fine-tuning)	IMDB	0.86	Model, 2021[58]
T5 (fine-tuning)	Combined Dataset	0.86	This study
XLNet (fine-tuning)	-	0.89	Model, 2019[59]

4.2. Model Performances in Real Context

In this section, we practically evaluate each modality's performance (facial expression, vocal tone, vocal text) in the Sri Lankan context. A total of 26 participants engaged with the mobile application, where they provided feedback on their predicted emotions based on videos they had previously uploaded. The accuracy of the predictions was assessed by evaluating the number of correct predictions after 10 iterations, taking into account 260 responses.

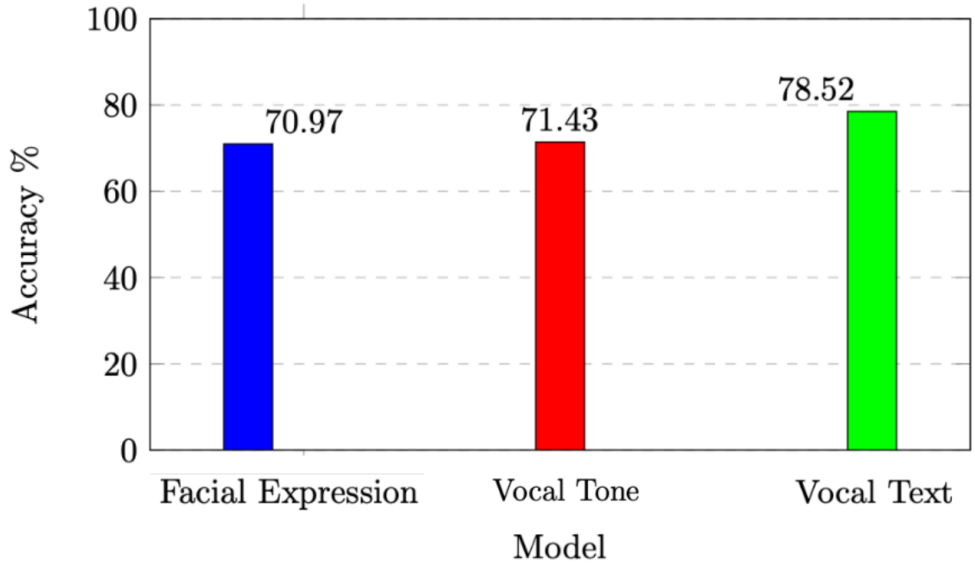

Figure 4. Model Accuracy Comparison in Real Context

The Figure 4 clearly demonstrates that the Vocal Text model exhibits the highest accuracy rate (78.52%), outperforming both the Facial Expression (70.97%) and Vocal Tone (71.43%) models. This indicates the practical usability of all three models in the Sri Lankan context, as they achieve an accuracy of over 70%.

4.3. Personalized Modality Identification for Individuals

This study aims to evaluate the effectiveness of having personalized multimodal emotion prediction through experiments. Participants were asked to provide feedback on predicting emotion for the first ten iterations. This feedback is used to adjust the weights in the weighted function described in the 3.6 section. Fig. 6 displays weight adjustments based on predicted emotions compared to the user given feedback to identify the most effective emotion prediction modality for the participant with the identification "Patients ID-2,".

Figure 5 indicates that "PatientsID-2" achieved the highest weight in the Vocal Tone modality (77.3) for the first 10 prediction iterations, while facial expression and vocal text received lower weights (17.5 and 23.5, respectively). This participant will prioritize the Vocal Tone model for future emotion predictions as it's the most accurate modality among the other two modalities (Facial Expression and vocal Text). This personalized approach is applied to all 26 participants to ensure accurate and individualized emotion prediction experiences.

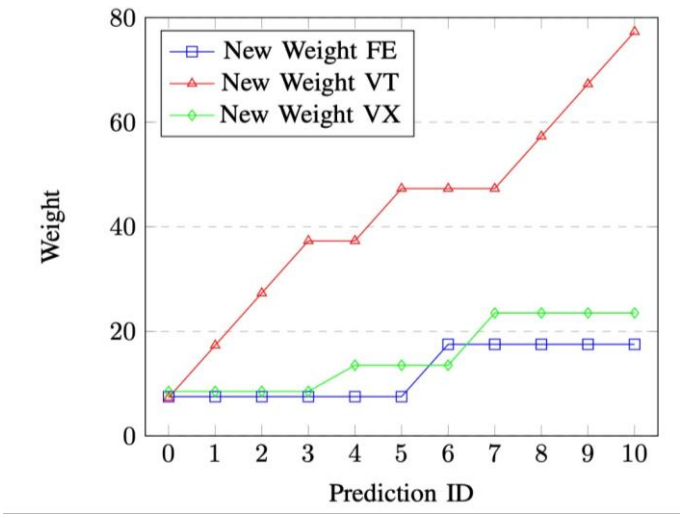

Figure 5. Vocal Tone Weighted Distribution [”Patients Id-2”]

4.4. Controlled Experiment

A controlled experiment compared the study and control groups receiving activities based on predicted emotion and random interventions. Statistical analysis was performed to determine if the RL enabled assistant's personalized activity suggestion effectively reduced anxiety and depression symptoms.

4.4.1. Depression and Anxiety Variation

Table 5 shows changes in depression and anxiety levels of both study and control groups before and after the intervention which uses the given multi modal personalized emotion prediction and emotional aware activity suggestion mobile application for a duration of 2 weeks. The symbols (+) and (-) indicate the change in participant numbers in the study and control groups before and after the intervention.

Depression: In the study group, normal depression symptoms increased by 23.08% (from 7.69% to 30.77%), while severe and extremely severe symptoms were reduced by 15.38% (from 23.08% to 7.69%). In the control group, normal symptoms decreased by 23.08% (from 30.77% to 7.69%), with a slight reduction in severe symptoms by 7.69% (from 15.38% to 7.69%) and no change in extremely severe symptoms (remaining at 15.38%).

Table 5. Change in Depression and Anxiety Symptoms Levels Pre-intervention and Post-intervention.

Level	Depression		Anxiety	
	Change in Study Group	Change in Control Group	Change in Study Group	Change in Control Group
Normal	23.08%	-23.08%	7.69%	-7.69%
Mild	15.38%	15.38%	23.08%	7.69%
Moderate	-7.69%	15.38%	38.46%	-23.08%
Severe	-15.38%	-7.69%	-15.38%	15.38%
Extremely Severe	-15.38%	0.00%	-53.85%	7.69%
Normal	23.08%	-23.08%	7.69%	-7.69%

Anxiety: In the study group, normal anxiety symptoms increased by 7.69% (from 7.69% to 15.38%), while severe symptoms were reduced by 15.38% (from 15.38% to 0.00%) and extremely severe symptoms by 53.85% (from 53.85% to 0.00%). In the control group, normal symptoms decreased by 7.69% (from 7.69% to 0.00%), with an increase in severe symptoms by 15.38% (from 23.08% to 38.46%) and extremely severe symptoms by 7.69% (from 7.69% to 15.38%). These results indicate that the RL-enabled agent's emotionally-aware activity suggestions had a positive impact on reducing depression and anxiety symptoms in the study group compared to the control group.

4.5. Impact on Mental Well Being

This section examines the impact of personalized activity suggestions on mental well being using pre and post-intervention DASS-21 questionnaire scores. The focus is on average anxiety and depression scores for each group to draw broader conclusions about the intervention's overall impact.

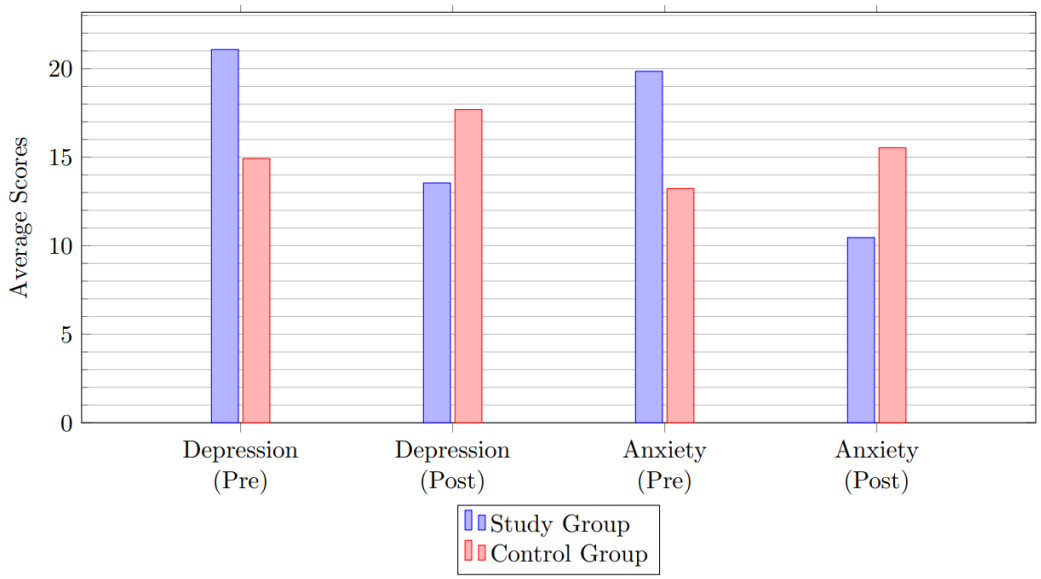

Figure 6. Average DASS-21 Scores

Figure 6 shows DASS-21 scores for the study and control groups before and after the intervention. The study group experienced a decrease in average depression score from 21.08 to 13.54 and a decrease in average anxiety score from 19.85 to 10.46. In contrast, the control group displayed an increase in average depression score from 14.92 to 17.69 and an increase in average anxiety score from 13.23 to 15.53.

5. Conclusions

This study aimed to devise a precise, individualized emotion prediction mechanism capable of accurately forecasting each individual's emotions across three modalities: facial expression, vocal tone, and vocal text. The study successfully achieved this by developing and validating three models. During the initial training stage using secondary datasets, the models demonstrated 76.21%, 81.1%, and 86% accuracy. After fine-tuning, the facial expression and vocal tone models displayed 72% and 73% accuracy, respectively. This is discussed in section 4.1, where each model is examined in dedicated subsections, comparing similar approaches and their outcomes to support the objective of developing a precise emotion prediction mechanism, particularly in the Sri Lankan context. Furthermore in the experimental stage, as shown in Figure 5, all models achieved over 70% accuracy in practical usage. These results underscore the effectiveness of these models in accurately forecasting emotions across the three modalities.

A personalized emotion prediction mechanism across three modalities was developed to enhance emotionally-aware predictive models at a broader level, as discussed in section 4.3. As highlighted in that section, the absence of vocal tone as an option for emotion prediction in "Patient ID 2" potentially reduced the effectiveness of the intervention. This underscores the importance of identifying personalized emotion prediction mechanisms tailored to individual users. In conclusion, the study emphasizes the critical need for developing a personalized emotion prediction mechanism that can accurately predict emotions across multiple modalities.

The second objective of this study was to develop an RL enabled, emotionally-aware activity suggestion agent to enhance the mental well-being of individuals displaying symptoms of anxiety and depression. The controlled experiment demonstrated the effectiveness of the RL agent, with the study group showing significant improvements in both depression and anxiety scores compared to the control group (see section 4.4). Specifically, the study group

experienced a 23.08% increase in normal depression symptoms and a 15.38% reduction in severe and extremely severe symptoms, while the control group displayed a 23.08% decrease in normal depression symptoms with no reduction in extremely severe symptoms. For anxiety, the study group had a 7.69% increase in normal symptoms and a 53.85% reduction in extreme severe symptoms, in contrast to the control group, which experienced an increase in both severe and extreme severe symptoms.

Further analysis using pre and post-intervention DASS-21 questionnaire scores (see section 4.5) revealed that the study group's average depression score decreased by 7.54 points and anxiety score by 9.39 points. In contrast, the control group showed increases in depression and anxiety scores by 2.77 and 2.3 points, respectively. These results highlight the effectiveness of the RL agent powered by emotionally aware interventions in significantly enhancing mental well being, effectively addressing the main objectives of the research and demonstrating the potential of personalized, emotion driven approaches in mental health care.

However, certain limitations should be acknowledged. The assumption that participants represent the broader young population in Sri Lanka may affect the generalizability of the findings. The focus on English-speaking participants could limit inclusivity, and reliance on personal devices may introduce variability in data quality. Additionally, the short study duration may not fully capture long term mental health trends. Future research should refine emotion prediction models, expand sample sizes, and extend study durations to address these limitations. Exploring advanced techniques for emotion predictions and more sophisticated RL algorithms like Deep Q-Network and TD3 could further enhance the effectiveness of these interventions. Despite these challenges, the study lays a solid foundation for future personalized mental health care advancements.

Author Contributions: Conceptualization: Amod Pathirana, Dharshana Kasthurirathna, Dumidu Kasun Rajakaruna, and Ajantha Atukorale; Methodology: Amod Pathirana and Dumidu Kasun Rajakaruna; Software: Dumidu Kasun Rajakaruna and Amod Pathirana; Deep Learning: Amod Pathirana and Dumidu Kasun Rajakaruna; Resources: University of Colombo School of Computing (UCSC) and Ajantha Atukorale; Data curation: Amod Pathirana, Dumidu Kasun Rajakaruna, Dharshana Kasthurirathna and Maheshi Yatiipansalawa; Writing original draft preparation: Amod Pathirana, Dumidu Kasun Rajakaruna and Maheshi Yatiipansalawa; Writing review and editing: Amod Pathirana; Validation: Amod Pathirana, Dumidu Kasun Rajakaruna, Dharshana Kasthurirathna, Maheshi Yatiipansalawa, and Ajantha Atukorale; Formal analysis: Rekha Aththidiye, Pandithakoralage, Pathiraja, Dharshana Kasthurirathna, Ajantha Atukorale, Amod Pathirana, Dumidu Kasun Rajakaruna; Investigation: Amod Pathirana, Dumidu Kasun Rajakaruna, Dharshana Kasthurirathna, and Ajantha Atukorale; Supervision: Dharshana Kasthurirathna, Ajantha Atukorale, Rekha Aththidiye, Pandithakoralage and Pathiraja; Funding acquisition: None.

Funding: This research received no external funding.

Data Availability Statement:

- **Publicly Available:** Fine-tuned Facial Expression: View Model; Fine-tuned Vocal Tone Models:View Model; Mobile Application: View Mobile App; Mobile App User Manual: View User Manual; Backend: View Backend; RL Model: View RL Model; FER2013: View FER2013 Dataset, RAVDESS: View RAVDESS, SAVEE: View SAVEE; TESSToronto: View TESS-Toronto; CREMA-D: View CREMA-D; Emotion-stimulus: View Emotion-stimulus; Emotions: View Emotion Dataset; DASS-21 Questionnaire: View DASS-21 Questionnaire; Information Sheet: View Information Sheet; Consent Form: View Consent Form; DASS-21 Questionnaire via Google Forms Results (Public View): View Google Forms Results; CBT Activities: View CBT Activities.
- **Restricted Access:** Primary Data Set (Facial Expression): View Facial Expression Data Set, Primary Data Set (Vocal Tone): View Vocal Tone Data Set
- **No Data Available**: In alignment with ethical guidelines and privacy regulations, the user-provided videos utilized for emotion prediction are not stored in our databases. These videos are securely deleted immediately after the completion of the emotion analysis, ensuring the protection of user privacy and data integrity.

Acknowledgments: The research team wishes to express deep gratitude to key contributors to this study. We appreciate our external advisors, Drs. Aththidiye, Pandithakoralage, and Ms. Pathiraja, for their input related to our study topic. Thanks go to our faculty members for their useful feedback, and to course lecturers for providing essential academic knowledge. Lastly, we recognize the support of our parents, friends, and fellow students. To all, we extend sincere thanks.

Conflicts of Interest: The authors declare no conflict of interest. This university-led study, conducted without external funding or institutional influence, involved voluntary participants from the university student population, selected and grouped anonymously. The research included collaboration with two independent mental healthcare professionals to ensure unbiased judgments. No personal relationships or associations influenced the study, and there are no known relationships with potential reviewers. All aspects were carried out impartially, ensuring the integrity of the findings. Note : The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results".

References

[1] World Health Organization, "Mental disorders." 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[2] National Institute of Mental Health (NIMH), "Any Anxiety Disorder." [Online]. Available: https://www.nimh.nih.gov/health/statistics/any-anxiety-disorder

[3] Kaiser Family Foundation (KFF), "Adults Reporting Symptoms of Anxiety or Depressive Disorder During the COVID-19 Pandemic by Sex." [Online]. Available: https://www.kff.org/other/state-indicator/adults-reporting-symptoms-of-anxiety-or-depressive-disorder-during-the-covid-19-pandemic-by-sex/

[4] D. F. Santomauro *et al.*, "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic," *Lancet*, vol. 398, no. 10312, pp. 1700–1712, Nov. 2021, doi: 10.1016/S0140-6736(21)02143-7.

[5] A. Ojagbemi and O. Gureje, "Mental health in low- and middle-income countries," in *Oxford Textbook of Social Psychiatry*, 1st ed., D. Bhugra, D. Moussaoui, and T. J. Craig, Eds. Oxford: Oxford University PressOxford, 2022, pp. 699–712. doi: 10.1093/med/9780198861478.003.0072.

[6] World Health Organization, "Depressive disorder (depression)." 2023.

[7] C. D. Mathers and D. Loncar, "Projections of Global Mortality and Burden of Disease from 2002 to 2030," *PLoS Med.*, vol. 3, no. 11, p. e442, Nov. 2006, doi: 10.1371/journal.pmed.0030442.

[8] Psychiatry.org, "Warning Signs of Mental Illness." [Online]. Available: https://www.psychiatry.org/patients-families/warning-signs-of-mental-illness

[9] P. R. Muskin, "What are anxiety disorders?" 2021. [Online]. Available: https://www.psychiatry.org/patients-families/anxiety-disorders/what-are-anxiety-disorders

[10] Mind, "Anxiety Signs and Symptoms." 2024. [Online]. Available: https://www.mind.org.uk/information-support/types-of-mental-health-problems/anxiety-and-panic-attacks/symptoms/

[11] J. Truschel, "Depression definition and DSM-5 diagnostic criteria - psycom." 2022. [Online]. Available: https://www.psycom.net/depression/major-depressive-disorder/dsm-5-depression-criteria

[12] S. Yoon, V. Dang, J. Mertz, and J. Rottenberg, "Are attitudes towards emotions associated with depression? A conceptual and meta-analytic review," *J. Affect. Disord.*, vol. 232, pp. 329–340, 2018, doi: 10.1016/j.jad.2018.01.009.

[13] P. Ekman, "Universal facial expressions of emotion," *Calif. Ment. Heal. Res. Dig.*, vol. 8, no. 4, pp. 151–158, 1970.

[14] J. F. Cohn *et al.*, "Detecting depression from facial actions and vocal prosody," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, Sep. 2009, pp. 1–7. doi: 10.1109/ACII.2009.5349358.

[15] U. Smrke, I. Mlakar, S. Lin, B. Musil, and N. Plohl, "Language, Speech, and Facial Expression Features for Artificial Intelligence–Based Detection of Cancer Survivors' Depression: Scoping Meta-Review," *JMIR Ment. Heal.*, vol. 8, no. 12, p. e30439, Dec. 2021, doi: 10.2196/30439.

[16] A. K. Silberbogen, E. Ulloa, D. L. Mori, and K. Brown, "A telehealth intervention for veterans on antiviral treatment for the hepatitis C virus.," *Psychol. Serv.*, vol. 9, no. 2, pp. 163–173, May 2012, doi: 10.1037/a0026821.

[17] E. A. Elliott and A. M. Jacobs, "Facial expressions, emotions, and sign languages," *Front. Psychol.*, vol. 4, p. 115, 2013, doi: 10.3389/fpsyg.2013.00115.

[18] P. Dursun, M. Emül, F. Gençöz, and others, "A Review of the Literature on Emotional Facial Expression and Its Nature," *Yeni Symp.*, vol. 48, no. 3, 2010.

[19] Z. Ahanin, M. A. Ismail, N. S. S. Singh, and A. AL-Ashmori, "Hybrid Feature Extraction for Multi-Label Emotion Classification in English Text Messages," *Sustainability*, vol. 15, no. 16, p. 12539, Aug. 2023, doi: 10.3390/su151612539.

[20] R. Zhang, C. Jia, and J. Wang, "Text emotion classification system based on multifractal methods," *Chaos, Solitons & Fractals*, vol. 156, p. 111867, Mar. 2022, doi: 10.1016/j.chaos.2022.111867.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *arXiv*, Sep. 2014, pp. 1–14. [Online]. Available: https://arxiv.org/abs/1409.1556

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, vol. 2016-Decem, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[23] M. A. H. Akhand, S. Roy, N. Siddique, M. A. S. Kamal, and T. Shimamura, "Facial Emotion Recognition Using Transfer Learning in the Deep CNN," *Electronics*, vol. 10, no. 9, p. 1036, Apr. 2021, doi: 10.3390/electronics10091036.

[24] B. Council, "Facial Expressions, Cultural Difference, Empathy." 2024.

[25] R. Jack, "Perception of Facial Expressions differs across Cultures," *American Psychological Association*, 2011.

[26] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression.," *J. Pers. Soc. Psychol.*, vol. 70, no. 3, pp. 614–636, 1996, doi: 10.1037/0022-3514.70.3.614.

[27] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011, doi: 10.1016/j.patcog.2010.09.020.

[28] E. van der Westhuizen and T. R. Niesler, "Synthesised bigrams using word embeddings for code-switched ASR of four South African language pairs," *Comput. Speech Lang.*, vol. 54, pp. 151–175, Mar. 2019, doi: 10.1016/j.csl.2018.10.002.

[29] C. Doğdu, T. Kessler, D. Schneider, M. Shadaydeh, and S. R. Schweinberger, "A Comparison of Machine Learning Algorithms and Feature Sets for Automatic Vocal Emotion Recognition in Speech," *Sensors*, vol. 22, no. 19, p. 7561, Oct. 2022, doi: 10.3390/s22197561.

[30] H. A. Abdulmohsin, H. B. Abdul wahab, and A. M. J. Abdul hossen, "A new proposed statistical feature extraction method in speech emotion recognition," *Comput. Electr. Eng.*, vol. 93, p. 107172, Jul. 2021, doi: 10.1016/j.compeleceng.2021.107172.

[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv*. Jan. 16, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[32] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[33] V. K. Jain, S. Kumar, and S. L. Fernandes, "Extraction of emotions from multilingual text using intelligent text processing and computational linguistics," *J. Comput. Sci.*, vol. 21, pp. 316–326, Jul. 2017, doi: 10.1016/j.jocs.2017.01.010.

[34] J. Guo, "Deep learning approach to text analysis for human emotion detection from big data," *J. Intell. Syst.*, vol. 31, no. 1, pp. 113–126, Jan. 2022, doi: 10.1515/jisys-2022-0001.

[35] L. Cai, Y. Hu, J. Dong, and S. Zhou, "Audio-Textual Emotion Recognition Based on Improved Neural Networks," *Math. Probl. Eng.*, vol. 2019, no. 1, pp. 1–9, Jan. 2019, doi: 10.1155/2019/2593036.

[36] D. Matsumoto and H. S. Hwang, "Culture and Emotion: The Integration of Biological and Cultural Contributions," *J. Cross. Cult. Psychol.*, vol. 43, no. 1, pp. 91–118, Jan. 2012, doi: 10.1177/0022022111420147.

[37] D. Keltner and J. Haidt, "Social functions of emotions." The Guilford Press, 2001.

[38] S. An, L.-J. Ji, M. Marks, and Z. Zhang, "Two Sides of Emotion: Exploring Positivity and Negativity in Six Basic Emotions across Cultures," *Front. Psychol.*, vol. 8, p. 610, Apr. 2017, doi: 10.3389/fpsyg.2017.00610.

[39] Psychology Tools, "Behavioral Activation Activity Diary." 2024.

[40] B. W. Dunlop, K. Scheinberg, and A. L. Dunlop, "Ten ways to improve the treatment of depression and anxiety in adults.," *Ment. Health Fam. Med.*, vol. 10, no. 3, pp. 175–81, Sep. 2013, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24427185

[41] J. S. Beck, *Cognitive behavior therapy: Basics and beyond*. New York, NY, USA: The Guilford Press, 2021.

[42] C. E. Ackerman, "CBT techniques: 25 cognitive behavioral therapy worksheets." 2023.

[43] S. Asokan, P. Geetha Priya, Sn. Natchiyar, and M. Elamathe, "Effectiveness of distraction techniques in the management of anxious children – A randomized controlled pilot trial," *J. Indian Soc. Pedod. Prev. Dent.*, vol. 38, no. 4, p. 407, 2020, doi: 10.4103/JISPPD.JISPPD_435_20.

[44] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[45] M. Banoula, "What is Q-learning: Everything you need to know." 2023.

[46] A. Violante, "Simple reinforcement learning: Q-learning." 2019.

[47] M. Milne-Ives, E. Selby, B. Inkster, C. Lam, and E. Meinert, "Artificial intelligence and machine learning in mobile apps for mental health: A scoping review," *PLOS Digit. Heal.*, vol. 1, no. 8, p. e0000079, Aug. 2022, doi: 10.1371/journal.pdig.0000079.

[48] eMoods, "Track your moods, improve your wellbeing." Yottaram, LLC.

[49] M. T. H. Le, T. D. Tran, S. Holton, H. T. Nguyen, R. Wolfe, and J. Fisher, "Reliability, convergent validity and factor structure of the DASS-21 in a sample of Vietnamese adolescents," *PLoS One*, vol. 12, no. 7, p. e0180557, Jul. 2017, doi: 10.1371/journal.pone.0180557.

[50] J. D. Henry and J. R. Crawford, "The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample," *Br. J. Clin. Psychol.*, vol. 44, no. 2, pp. 227–239, Jun. 2005, doi: 10.1348/014466505X29657.

[51] S. T. Arokkiya Mary and L. Jabasheela, "An Evaluation of Classification Techniques for Depression, Anxiety and Stress Assessment," in *Proceedings of the International Conference for Phoenixes on Emerging Current Trends in Engineering and Management (PECTEAM 2018)*, 2018, pp. 64–69. doi: 10.2991/pecteam-18.2018.13.

[52] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 1–67, 2020.

[53] P. N. R. Bodavarapu and P. V. V. S. Srinivas, "Facial expression recognition for low resolution images using convolutional neural networks and denoising techniques," *Indian J. Sci. Technol.*, vol. 14, no. 12, pp. 971–983, 2021, doi: 10.17485/IJST/v14i12.14.

[54] G. P. Kusuma, Jonathan, and A. P. Lim, "Emotion Recognition on FER-2013 Face Images Using Fine-Tuned VGG-16," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 5, no. 6, pp. 315–322, 2020, doi: 10.25046/aj050638.

[55] M. G. de Pinto, M. Polignano, P. Lops, and G. Semeraro, "Emotions Understanding Model from Spoken Language using Deep Neural Networks and Mel-Frequency Cepstral Coefficients," in *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, May 2020, pp. 1–5. doi: 10.1109/EAIS48028.2020.9122698.

[56] R. M. B. Novais, "A framework for emotion and sentiment predicting supported in ensembles," Universidade do Algarve, 2022.

[57] M. Ciolino, J. Kalin, and D. Noever, "Fortify Machine Learning Production Systems: Detect and Classify Adversarial Attacks," *arXiv*. Feb. 18, 2021. [Online]. Available: http://arxiv.org/abs/2102.09695

[58] P.-C. Tu and H.-K. Pao, "A Dropout Style Model Augmentation for Cross Domain Few-Shot Learning," in *2021 IEEE International Conference on Big Data (Big Data)*, Dec. 2021, pp. 1138–1147. doi: 10.1109/BigData52589.2021.9671673.

[59] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 5753–5763.