REAL-TIME FACIAL EXPRESSION RECOGNITION FOR ENHANCED REMOTE THERAPY

A PREPRINT

Isabella F. Paolucci **Michael Amadi** **Elisa de la Vega Ricardo** **Eric S. Arnold**
School of Data Science, University of Virginia, Charlottesville, VA 22904
ctx8bn@virginia.edu, mxg9xv@virginia.edu, vun4kt@virginia.edu, esa5ch@virginia.edu

July 21, 2025

ABSTRACT

This paper presents a deep learning approach to enhancing remote therapy through real-time facial expression recognition. We develop and evaluate a suite of convolutional neural network (CNN) models—including custom architectures and transfer learning approaches—for classifying emotional states from facial images. Our system is implemented in PyTorch and trained on both FER-2013 and CK+ datasets, which are preprocessed, merged, and augmented to address class imbalance and variability. We compare baseline and SMOTE-enhanced training strategies across multiple architectures, demonstrating that synthetic class balancing significantly improves F1 scores, particularly for underrepresented emotions. ResNet18 emerges as the top-performing model, achieving 59.00% accuracy and 53.23% macro-averaged F1 score on the combined dataset. Error analysis highlights persistent challenges in recognizing minority classes such as "contempt," underscoring the need for targeted improvements in data representation and model generalization. This work focuses on offline evaluation, providing a rigorous foundation for future integration into real-time telehealth systems. Our findings support the potential of CNN-based emotion recognition to improve therapist-client interactions, enhance remote care reliability, and contribute to emotionally intelligent digital health platforms.

Keywords Facial Expression Recognition · Remote Therapy · Deep Learning · Telehealth · Emotion Detection

1 Motivation

Enhancing Therapist-Client Interactions Through Real-Time Emotion Detection.
Effective therapy sessions rely on accurately perceiving a client's emotional state. In remote sessions, subtle cues—such as micro-expressions—can be easily missed due to screen limitations, bandwidth constraints, and reduced visibility Bailenson (2021); Kara and Ersoy (2022). Despite advances in video communication tools, therapists often report difficulty in gauging emotional depth and shifts during telehealth interactions Yellowlees et al. (2008); Edirippulige and Appendices (2009); Norman (2006). This challenge is particularly concerning given the rising demand for remote mental health care and the need for scalable, high-quality therapeutic solutions Zhou et al. (2020); Torous et al. (2020); Torous and Keshavan (2020).

An automated, real-time emotion recognition system offers the potential to continuously capture and analyze facial expressions, providing therapists with an objective emotional map of the session Calvo et al. (2015). Such systems could enhance clinical insight, reduce missed cues, and support more personalized and effective interventions. However, existing emotion recognition frameworks often lack the accuracy, interpretability, or real-time responsiveness required for sensitive therapeutic contexts Li and Deng (2020); Yang et al. (2017).

This work aims to bridge that gap by investigating how state-of-the-art deep learning models can be adapted and evaluated for robust, real-time facial emotion detection in remote therapy settings.

2 Literature Review

2.1 Deep CNN Architectures and Benchmark Performance

Recent advances in facial emotion recognition have been primarily driven by significant developments in deep convolutional neural network (CNN) architectures. For instance, Khaireddin and Chen (2021) reported state-of-the-art results on the FER2013 dataset by utilizing the VGGNet architecture alongside an extensive hyperparameter optimization process. Their method achieved remarkable single-network accuracy, highlighting the robustness and potential of deep learning models to accurately interpret complex human emotional expressions.

2.2 Dataset Generalization and Subtle Emotion Detection

Complementing this research, MOL (2024) demonstrated that CNNs could effectively generalize across diverse datasets such as the CK+ dataset. MOL's study emphasized CNN's capability to detect subtle emotional nuances, reinforcing the framework's applicability in automated emotion detection and broader acceptance within the research community.

2.3 Evaluation Methodologies and Standardization

In exploring methodological frameworks for evaluating facial emotion recognition, Paiva-Silva et al. (2016) highlighted critical inconsistencies and limitations prevalent in many evaluation methods. Their analysis stressed the importance of developing standardized metrics and methodologies to facilitate comparability and reproducibility across different studies, a foundational step toward advancing the reliability and validity of FER systems.

2.4 Edge Computing for Real-Time Applications

Furthermore, Zhang et al. (2019) innovatively integrated CNN-based models with edge computing paradigms to address computational efficiency without sacrificing accuracy. Their approach specifically targeted resource-constrained environments, suggesting practical deployment scenarios where real-time emotion recognition systems are crucial, such as in mobile or embedded applications.

2.5 Transfer Learning and Cross-Dataset Validation

In another recent comparative study, Ma (2024) provided insights into CNN performance across two widely used datasets, FER-2013 and revised CK+, further validating CNN architectures' versatility. This research supported the feasibility and efficacy of transfer learning approaches, whereby CNNs pretrained on large-scale datasets can be effectively adapted for emotion-specific tasks, significantly enhancing their practical application.

2.6 Comprehensive CNN Evaluations and Best Practices

Giannopoulos et al. (2017) contributed substantially by reviewing and validating various deep learning methodologies specifically tailored for the FER-2013 dataset. Their comprehensive assessment confirmed the superiority of CNN models compared to traditional machine learning methods, additionally offering practical guidance on optimal architectures and training strategies, essential for achieving high accuracy in emotion recognition tasks.

2.7 Balancing Speed and Accuracy: Efficient Architectures

Moreover, Dar et al. (2022) introduced the Efficient-SwishNet architecture, which notably balanced computational efficiency with high recognition accuracy. Their method presented significant advancements, particularly suited for real-time applications where speed and accuracy must be optimized simultaneously, thus addressing one of the key challenges in practical FER implementations.

2.8 Summary and Implications

Collectively, these studies underline the ongoing significant progress in facial emotion recognition through advanced CNN frameworks, innovative integration approaches such as edge computing, and rigorous methodological evaluations. These advancements open exciting avenues for further research and practical applications, extending beyond traditional scenarios into complex and resource-constrained environments.

Table 1: FER-2013 Dataset Statistics (80/20 Split)

Expression	Total Count	Train (80%)	Test (20%)
Angry	4,953	3,995	958
Disgust	547	436	111
Fear	5,121	4,097	1,024
Happy	8,989	7,215	1,774
Neutral	6,198	4,965	1,233
Sad	6,077	4,830	1,247
Surprise	4,002	3,171	831

Figure 1: Sample image from CK+ illustrating a labeled emotional expression.

3 Methods

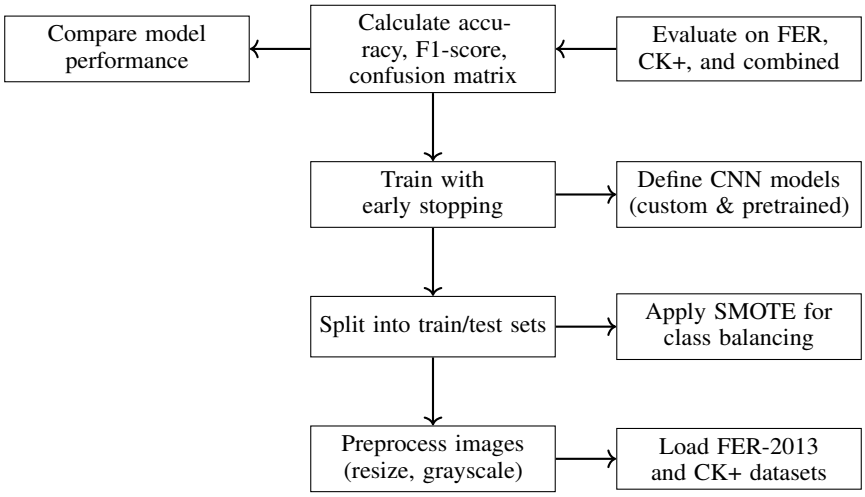

Figure 2: Workflow for the facial expression recognition pipeline.

3.1 Datasets: Open-Source Databases

Reliable, diverse, and well-labeled datasets are fundamental for training robust emotion recognition models. For our project, we utilized the Facial Expression Recognition 2013 dataset (FER-2013), a large-scale dataset with thousands of

labeled images capturing a wide range of facial expressions under varied conditions, and the Cohn-Kanade Plus dataset (CK+), a high-precision dataset with detailed annotations for subtle facial muscle movements. We aimed to combine these datasets, leveraging the broad generalization of FER-2013 with the fine-grained annotation of CK+, resulting in a model that is both robust and precise.

3.2 Methodology

3.2.1 Preprocessing and Exploratory Data Analysis

We first preprocessed and evaluated both the FER-2013 and CK+ datasets independently. Face detection was performed using OpenCV Haar cascade classifiers. Detected faces were resized to 48×48 pixels, converted to grayscale, and normalized to a pixel intensity range of $[0, 1]$.

For Exploratory Data Analysis (EDA), we analyzed FER-2013, CK+, and a combined dataset. FER-2013 includes 28,709 training images, 3,589 validation, and 3,589 test samples. CK+ contributes 920 frames, which we split into training and testing subsets. EDA included class distribution, pixel-intensity statistics (mean, standard deviation), and visual inspection through sample image grids to ensure data quality and class representation.

To address class imbalance—particularly the scarcity of the "contempt" class in CK+—we used the `WeightedRandomSampler` during training to up-weight minority classes, SMOTE (Synthetic Minority Oversampling Technique) on flattened features, and `CrossEntropyLoss` with class weighting. Data augmentation techniques included random horizontal flips, rotations ($\pm 15°$), and scaling transformations ($\pm 10\%$).

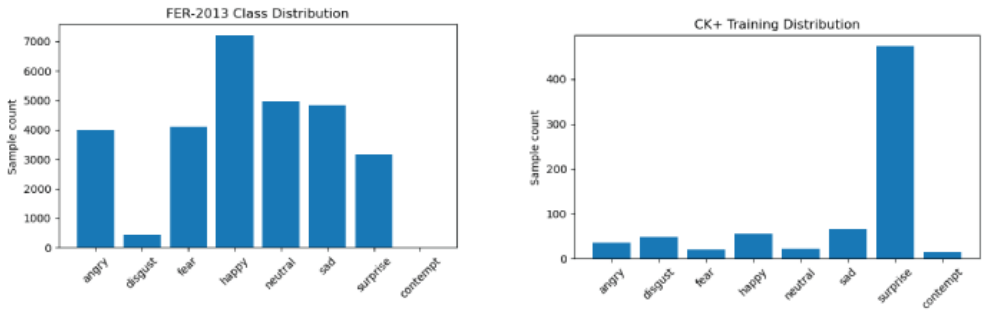

Figure 3: **Class Distributions in FER-2013 and CK+ Datasets.** The top plot shows the class distribution for the FER-2013 dataset, which contains relatively balanced representation across most emotion categories, except for a significant underrepresentation of the "disgust" and "contempt" classes. The bottom plot illustrates the training distribution for the CK+ dataset, which exhibits extreme class imbalance–most notably the overwhelming prevalence of the "surprise" class and the scarcity of "contempt," "fear," and "neutral." These distributions highlight key challenges for model training, particularly in achieving generalization across minority classes and avoiding performance bias.

3.2.2 Model Definitions

We evaluated four types of CNN-based models: custom architectures, transfer learning models, deeper CNNs, and multi-branch networks.

Custom CNN Architectures (MyCNN Series) We developed six variants (CNNv1 through CNNv6). CNNv1 includes two convolutional layers, max-pooling, and fully connected layers. Later versions incrementally increase depth and complexity.

Transfer Learning Architectures We used pretrained ResNet18, VGG16, and DenseNet121 models, each adapted for grayscale input by modifying the first convolutional layer. Final classifier layers were fine-tuned. These models leverage ImageNet-learned features for better generalization.

Deeper CNN (Ma2024CNN) Inspired by Ma (2024), we implemented a five-layer CNN for deeper feature extraction, benchmarking it on FER-2013 and CK+ using the original paper's settings.

Multi-Branch CNN This architecture processes facial images through two parallel streams—texture and shape—that merge before classification. This design tests whether specialized feature paths can improve emotion recognition accuracy.

3.2.3 Methodological Contributions

Our methodology contributes to understanding the effectiveness of architectural designs including custom CNNs, pretrained models, and multi-branch approaches for facial emotion recognition under data constraints.

4 Experiments

4.1 Training Protocol

All models were trained using the AdamW optimizer with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-4}. We used a batch size of 32 and trained each model for up to 10 epochs, employing `EarlyStopping` with a patience parameter of 3 to prevent overfitting. Data augmentation techniques included random horizontal flips, rotations of up to $\pm 15^{\circ}$, and scaling transformations of up to $\pm 10\%$ to improve generalization and robustness.

4.2 Model Architectures

Our evaluation included six primary model types. The *MyCNN Series* consists of simple convolutional neural networks with two convolutional layers (with 32 and 64 filters, respectively), followed by max-pooling, a fully connected layer with 128 units, and a final softmax classification layer. The *EmotionResNet18* model is a ResNet-18 architecture pretrained on ImageNet, adapted to accept 48×48 grayscale input; only the final residual block and classifier were fine-tuned. Similarly, *EmotionVGG16* employs a VGG-16 model with grayscale input modifications and retrained final classifier layers. *EmotionDenseNet121* adapts the DenseNet-121 model for grayscale input, with fine-tuning applied to the dense block and classifier layers.

We also implemented the *Ma2024CNN*, a deeper five-layer CNN based on Ma (2024), which was trained for 50 epochs with a batch size of 512. Finally, our *Multi-Branch CNN* uses a dual-path architecture, separating texture and shape feature extraction before merging and feeding into the final classification layer. This structure is designed to investigate whether parallel feature streams can improve the model's ability to detect subtle emotional cues.

4.3 Evaluation Metrics

We evaluated performance using accuracy and macro-averaged F1 score (both per fold and average). We also report per-class precision, recall, and F1 using `sklearn.classification.report`. Confusion matrices were plotted for each model's final validation fold to aid in visual evaluation.

This standardized evaluation allows direct comparisons between deeper CNNs (e.g., Ma2024CNN), transfer learning models (EmotionResNet18, EmotionVGG16, EmotionDenseNet121), and our own custom designs (MyCNN series).

4.4 Outcomes and Observations

ResNet18 achieved the best overall performance across accuracy and F1-score. Among custom architectures, MyCNNv6 and Ma2024CNN performed well, especially with SMOTE-enhanced training. SMOTE consistently improved minority class performance, particularly in "contempt" and "fear."

Combining FER-2013 and CK+ improved generalization. SMOTE also enhanced robustness across variable image qualities and imbalanced class distributions.

4.5 Error Analysis

Key challenges included class imbalance, overfitting, and model comparison. Despite efforts like `WeightedRandomSampler`, the "contempt" class remained poorly recalled. This reflects the difficulty of learning rare emotional classes with limited samples.

Overfitting was observed in deeper custom models like Ma2024CNN. Techniques such as `EarlyStopping` helped but were insufficient in some cases. Pretrained models consistently outperformed custom CNNs, reinforcing the value of transfer learning for emotion recognition, especially with small datasets.

5 Results

Code and experiments available at: github.com/Datascifer/facemine

The following table summarizes performance across all models on the combined dataset, under both baseline and SMOTE-enhanced training. These results demonstrate the consistent improvements gained from SMOTE and identify the top-performing architectures.

Table 2: Model performance comparison (Combined dataset: Baseline vs. SMOTE)

Model	Accuracy (Baseline)	F1 Score (Baseline)	Accuracy (SMOTE)	F1 Score (SMOTE)
MyCNNv1	42.49%	25.58%	44.57%	35.05%
MyCNNv2	34.19%	16.42%	38.78%	26.73%
MyCNNv3	42.88%	26.44%	47.46%	40.84%
MyCNNv4	42.65%	25.54%	45.57%	39.05%
MyCNNv5	45.56%	27.12%	44.49%	34.33%
MyCNNv6	45.55%	30.53%	52.08%	46.41%
Ma2024CNN	45.56%	31.33%	51.79%	46.23%
DenseNet121	26.59%	15.34%	47.28%	38.90%
VGG16	24.28%	4.88%	16.83%	3.60%
ResNet18	**51.41%**	**40.09%**	**59.00%**	**53.23%**

Model Performance Summary

ResNet18 consistently outperformed all other models, achieving the highest accuracy and F1 score, particularly when SMOTE was applied. With SMOTE, ResNet18 reached 59.00% accuracy and a macro F1 score of 53.23%, making it the top-performing architecture. MyCNNv6 and Ma2024CNN also performed competitively under SMOTE-enhanced conditions, with F1 scores of 46.41% and 46.23% respectively.

VGG16, despite being pretrained, underperformed, likely due to overfitting and insufficient feature transfer in deeper layers. SMOTE showed a marked effect across nearly all models, improving recall of minority classes and overall F1 scores. Training on the combined dataset also enhanced generalization and robustness.

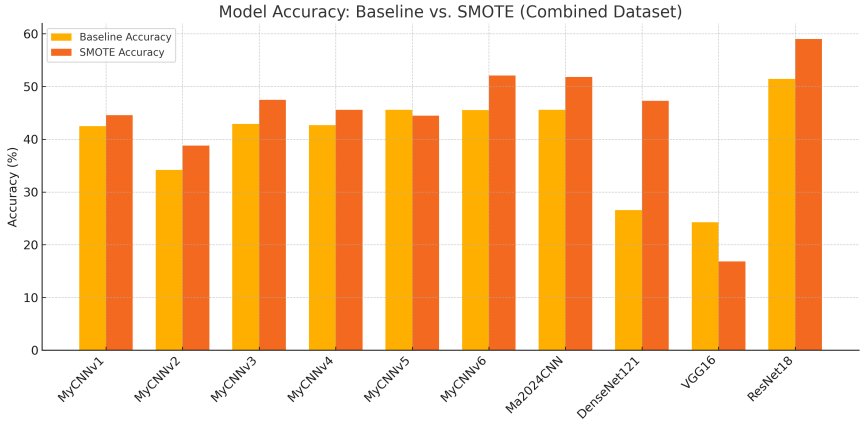

Figure 4: Comparison of model accuracy between baseline and SMOTE-enhanced training on the Combined dataset.

6 Conclusion

6.1 Discussion

The results of this study highlight meaningful progress toward reliable facial expression recognition in remote therapy contexts. By evaluating a range of CNN-based models on FER-2013, CK+, and a merged dataset, we assessed model robustness and limitations under real-world constraints.

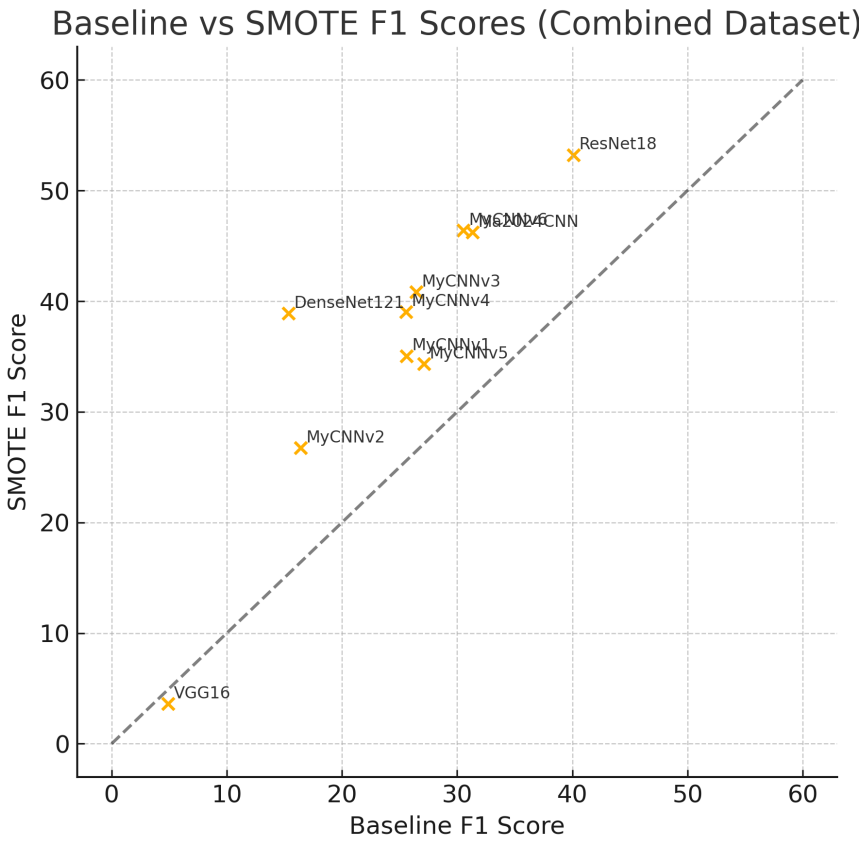

Figure 5: Scatter plot comparing baseline vs. SMOTE-enhanced F1 scores for each model. The diagonal line represents equal performance.

ResNet18 emerged as the most effective model, benefiting from residual connections and pretrained ImageNet features, which allowed for superior generalization. Custom architectures such as MyCNNv6 and Ma2024CNN also delivered strong results when paired with SMOTE, highlighting the benefit of task-specific design and class balancing. SMOTE significantly improved recall for underrepresented classes and contributed to more balanced classification performance overall.

6.2 Conclusion and Future Work

Our findings reinforce the importance of model selection and data preprocessing when designing emotion recognition systems for telehealth. These systems must handle data scarcity, class imbalance, and facial variability—conditions common in real-world applications.

Despite improvements, challenges remain. Minority emotion classes such as “contempt” were consistently misclassified or ignored, even after SMOTE rebalancing. Overfitting was also observed in deeper pretrained models like VGG16 and DenseNet121, which failed to generalize despite data augmentation. These limitations suggest promising directions for future work: Promising directions for future work include architectural improvements to better capture subtle emotional variations; data-centric strategies such as class synthesis and transfer learning from emotion-specific datasets; and live deployment with feedback loops, where real-time facial expression recognition systems can be validated and refined in clinical practice.

Ultimately, this work validates the capability of deep CNNs—especially ResNet18—for emotion recognition in telehealth. By advancing robustness, bias mitigation, and real-time readiness, this research lays a foundation for emotionally aware remote care technologies.

Contributions

Michael Amadi

Engaged in the initial brainstorming and research for Project Milestone I. Wrote and submitted the deliverable for Milestone I. Developed models submitted for Project Milestone I. Participated in the presentation of Milestone I. Developed models submitted for Milestone II. Developed all ten models submitted for Milestone III. Updated the Abstract and Motivation for Project Milestone III, wrote the Literature Review for Milestone III, and compiled the final document in Overleaf. Participated in the Final Presentation.

Eric S. Arnold

Engaged in the initial brainstorming and research for Project Milestone I. Aided in editing Project Milestone II. Wrote the Results and Conclusion sections for Project Milestone III. Participated in the Final Presentation.

Isabella Paolucci

Engaged in the initial brainstorming and research for Project Milestone I. Participated in the presentation of Milestone I. Aided in editing and submitted Project Milestone II. Created the PowerPoint presentation for the Final Project. Wrote the Contributions section of the Final Report, contributed to the Data, Methods, and Experiments section, and helped write and edit the Results and Conclusion sections. Participated in presenting the Final Project.

Elisa de la Vega Ricardo

Engaged in the initial brainstorming and research for Project Milestone I. Created the presentation PowerPoint for Project Milestone I. Developed models for Project Milestone II and Milestone III. Contributed to the Data, Methods, and Experiments section of the Final Report. Presented during the Final Project Presentation.

References

Bailenson, J. N. (2021). Nonverbal overload: A theoretical argument for the causes of zoom fatigue. *Technology, Mind, and Behavior*, 2(1).

Calvo, R. A., D'Mello, S., Gratch, J., and Kappas, A. (2015). *The Oxford Handbook of Affective Computing*. Oxford University Press, Oxford, UK.

Edirippulige, S. and Appendices, T. (2009). Psych-technology: a systematic review of the telepsychiatry literature. In *Psychiatry On Line, The International Forum for Psychiatry*, volume 30. Priory Lodge Education Ltd.

Kara, I. U. and Ersoy, E. G. (2022). A new exhaustion emerged with covid-19 and digitalization: A qualitative study on zoom fatigue. *OPUS Journal of Society Research*, 19(46):365–379.

Li, S. and Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215.

Norman, S. (2006). The use of telemedicine in psychiatry. *Journal of Psychiatric and Mental Health Nursing*, 13(6):771–777.

Torous, J. and Keshavan, M. S. (2020). Covid-19, mobile health and serious mental illness. *Schizophrenia Research*, 218:36.

Torous, J., Myrick, K. J., Rauseo-Ricupero, N., and Firth, J. (2020). Digital mental health and covid-19: using technology today to accelerate the curve on access and quality tomorrow. *JMIR Mental Health*, 7(3):e18848.

Yang, B., Cao, J., Ni, R., and Zhang, Y. (2017). Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. *IEEE Access*, 6:4630–4640.

Yellowlees, P. M., Hilty, D. M., Marks, S. L., Neufeld, J., and Bourgeois, J. A. (2008). A retrospective analysis of a child and adolescent emental health program. *Journal of the American Academy of Child & Adolescent Psychiatry*, 47(1):103–107.

Zhou, X., Snoswell, C. L., Harding, L. E., Bambling, M., Edirippulige, S., Bai, X., and Smith, A. C. (2020). The role of telehealth in reducing the mental health burden from covid-19. *Telemedicine and e-Health*, 26(4):377–379.