Facial Emotion Recognition Using Deep Learning Models: A Clinical Perspective for Monitoring PTSD-Related Emotional Responses

Abstract
Post-Traumatic Stress Disorder (PTSD) is characterized by heightened emotional responses, such as fear and anger, which are often triggered by reminders of traumatic events. This study explores the use of facial emotion recognition (FER) models to monitor these emotional responses in clinical settings, providing objective insights into patients' emotional regulation during therapy. Using a dataset of facial images labeled with seven emotions—anger, contempt, disgust, fear, happiness, sadness, and surprise—we developed and compared a Convolutional Neural Network (CNN) and ResNet50 model. The CNN outperformed ResNet50, achieving higher test accuracy and more consistent performance across demographic subgroups, including age and gender. However, subgroup analysis revealed lower accuracy for the RU country group, highlighting the potential influence of cultural factors on emotion recognition. Model examination techniques, such as saliency maps and sensitivity analysis, confirmed that the CNN focused on meaningful facial features, ensuring interpretability and reliability. While the study demonstrates the potential of FER systems to support PTSD therapy, challenges such as dataset size and demographic diversity remain, necessitating future work to enhance robustness and generalizability.

Introduction
Post-Traumatic Stress Disorder (PTSD) is a psychiatric condition characterized by heightened emotional reactivity to triggers that remind individuals of traumatic events. Emotional responses such as fear, anger, and sadness are often amplified in PTSD, making it difficult for affected individuals to regulate their emotional states. Facial expressions serve as a critical channel for understanding emotional experiences, and abnormalities in recognizing or responding to emotions are well-documented in PTSD populations (Gapen, 2009). Facial Emotion Recognition (FER) models provide an opportunity to observe and quantify these emotional reactions in real time, potentially serving as an innovative tool in clinical therapy to monitor emotional regulation and therapeutic progress.

Recent advancements in machine learning and deep learning, particularly through Convolutional Neural Networks (CNNs) and transfer learning with models like ResNet, have enabled significant progress in the automated recognition of facial expressions. These technologies analyze subtle changes in facial features to detect emotions such as anger, contempt, happiness, and fear. By implementing FER models in clinical settings, practitioners can gain insights into patients' emotional states without relying solely on self-reports, which may be incomplete or inaccurate due to the patient's distress.

Several studies emphasize the role of the amygdala—a brain region associated with emotional processing and threat detection—in PTSD. Hyperactivity of the amygdala has been linked to heightened sensitivity to emotional stimuli, particularly fear and anger (Chappidi Suneetha, 2022). Individuals with PTSD may exhibit biases in facial emotion recognition, often perceiving neutral or ambiguous faces as threatening (Gapen, 2009). This altered perception can contribute to interpersonal difficulties, which are a hallmark of PTSD. Machine learning-based emotion recognition systems could help identify these biases, facilitating targeted therapeutic interventions.

In this project, we leverage facial emotion recognition models to explore emotional responses in individuals across distinct demographic subgroups, particularly focusing on country-based groups.

The dataset, sourced from Kaggle, consists of images capturing individuals exhibiting seven primary emotions—anger, contempt, disgust, fear, happiness, sadness, and surprise. Each image is labeled with metadata such as set ID, gender, age, and country. By employing both CNN and ResNet architectures, we aim to evaluate the general and subgroup-specific performance of FER models. The inclusion of subgroup analyses (e.g., age, gender, and country) is essential to assess model robustness across diverse populations and investigate potential cultural or ethnic influences on emotion expression and recognition.

In addition to overall performance evaluation, we employ model examination techniques such as saliency maps and sensitivity analysis to improve interpretability. These methods allow us to visualize critical regions of the face that contribute to emotion predictions, ensuring that the models are reliable and applicable in real-world settings (Savchenko, Savchenko, & Makarov, 2022). While the CNN model demonstrated higher intergroup performance compared to ResNet, subgroup analyses revealed performance disparities, particularly in the "RU" country group, suggesting potential cultural or demographic biases.

By combining advanced machine learning techniques with subgroup analysis and interpretability methods, this study highlights the potential of FER models to support clinical settings, particularly for PTSD therapy. FER systems could provide objective measures of emotional changes during therapeutic interventions, enabling clinicians to monitor patients' progress and adapt treatments to address emotional dysregulation effectively. Future work should explore larger, ethnically diverse datasets to enhance the robustness and applicability of these models in real-world clinical environments.

Methods

Dataset Description

This study utilized a dataset obtained from Kaggle containing facial images depicting seven distinct emotions: anger, contempt, disgust, fear, happiness, sadness, and surprise. Each image is labeled accordingly, making the dataset well-suited for training and evaluating machine learning models for emotion recognition.

The dataset is organized into two main components. First, there is a directory structure (i.e., DATA_DIR) that includes 19 folders, each corresponding to a unique participant (set_id). Within these folders, images illustrate various emotional expressions. Second, there is a metadata file (emotions_df in CSV format) that provides additional demographic details. These details include the participant's unique identifier (set_id), gender, age, and country of origin. Together, the labeled images and demographic metadata enable subgroup analyses to investigate potential performance variations across different countries, genders, and age groups.

Data Preprocessing

The preprocessing of the dataset involved several crucial steps to prepare it for analysis and model training. First, the images were standardized to ensure uniformity in size and format. For the CNN model, images were resized to 48x48 pixels, which was sufficient to capture essential facial features while reducing computational complexity. In contrast, the ResNet50 model required images to be resized to 224x224 pixels, aligning with the input size required for the pre-trained architecture. All pixel values were normalized to the range [0, 1], a transformation that stabilized the training process and improved computational efficiency. Emotion labels were extracted directly from the filenames of the images and cleaned to ensure consistency across the dataset. These textual labels were then encoded into numerical indices using a label encoder, making them suitable for input to machine learning models. To address potential overfitting and enhance the model's ability to generalize, data augmentation was applied exclusively to the training set. Augmentation techniques included random rotations, horizontal and vertical shifts, zooming, shearing, and horizontal flipping, creating diverse variations of the original images without altering their inherent emotional content. The test set remained untouched during augmentation to preserve its integrity for unbiased evaluation. This preprocessing pipeline ensured the dataset was well-structured, standardized, and robust, ready for effective training of the CNN and ResNet models.

To ensure the independence of the training and test sets, a stratified sampling approach was employed during dataset splitting. This method maintained the proportion of each demographic group (such as country) and emotion class across both subsets, ensuring a balanced representation in the training and test sets. Additionally, special care was taken to prevent data leakage by ensuring that images from the same participant, identified by their set_id, were not split between the training and test sets. This precaution ensured that the model could not inadvertently learn individual-specific features, which would have compromised the generalizability of the results. Furthermore, data augmentation techniques, such as random rotations, shifts, zooming, and flipping, were applied exclusively to the training set to enhance its diversity and prevent overfitting. The test set was left untouched during augmentation to preserve its integrity as an unbiased benchmark for evaluating the models' performance. These measures collectively ensured that the test set provided a reliable and independent assessment of the models' ability to generalize to unseen participants and emotional expressions. The training set was used for model development and optimization, while the test set served as an independent benchmark for evaluating the models' generalization ability. The training set comprised a diverse range of participants, ensuring adequate representation of the key demographic groups: RU (Russia) and Other (a combination of India and the Philippines), as well as both age groups (under 31 and beyond 31 years old) and gender groups (male and female). This stratification process minimized bias and ensured that the cohorts reflected real-world variations commonly observed in clinical or practical emotion recognition tasks.

Furthermore, the representativeness of the cohorts was assessed by analyzing their distribution of demographic attributes and emotion labels. The emotion class distributions in both the training and test sets closely mirrored those in the full dataset, indicating that the stratified sampling effectively maintained the overall structure of the data. Such representativeness is critical for real-world clinical settings, where emotion recognition systems are expected to generalize across diverse populations and varying facial expressions. By ensuring that the training and test sets retained the diversity and balance of the original dataset, we enhanced the reliability of the models' evaluation and their applicability to practical use cases.

To evaluate the robustness and fairness of the models, a comprehensive subgroup analysis was performed for both the CNN and ResNet50 models. The dataset was divided into key demographic subgroups to assess variations in model performance across different populations. Specifically, participants were categorized by country into two groups: RU (Russia) and Other (a combination of India and the Philippines), by age into Under_31 (ages below 31) and Beyond_31 (ages 31 and above), and by gender into Male and Female. For each subgroup, the models' predictions were evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. These metrics offered a detailed understanding of how well the models generalized across diverse subgroups and whether there were any biases in their predictions. By applying the same evaluation process to both CNN and ResNet50, the analysis provided a basis for comparing their performance consistency and identifying any disparities in subgroup predictions. This approach ensured a thorough investigation into the models' strengths and limitations, particularly in addressing the challenges posed by demographic diversity.

Model Development

Two deep learning models, Convolutional Neural Network (CNN) and ResNet50, were developed to classify facial images into seven distinct emotions. The CNN model was designed for computational efficiency and simplicity, serving as a foundational baseline for the study. It consisted of three convolutional layers that extracted hierarchical features from the images, identifying patterns such as edges, textures, and higher-level facial structures. Each convolutional layer was followed by a max-pooling layer to reduce the spatial dimensions of the feature maps, retaining critical features while minimizing computational overhead. The output of the convolutional layers was flattened and passed through fully connected dense layers to combine the extracted features. A softmax activation layer at the end of the network produced probabilities for the seven emotion classes. The CNN operated on images resized to 48x48 pixels, a size that balanced computational efficiency with sufficient resolution for facial emotion recognition.

The ResNet50 model utilized transfer learning by leveraging pre-trained weights from the ImageNet dataset. ResNet50 is specifically designed to extract rich and complex features from high-resolution images using residual connections, which helps mitigate the vanishing gradient problem in deeper networks. The base layers of ResNet50 were frozen to retain the pre-trained feature representations, while a custom classification head was added to adapt the model for the emotion recognition task. This classification head included a global average pooling layer to aggregate spatial information across the feature maps, followed by dense layers with ReLU activations to transform the features into meaningful representations. A dropout layer was added to reduce overfitting, and the final softmax activation layer output probabilities for the seven emotion classes. ResNet50 was trained on images resized to 224x224 pixels, aligning with the input size required for its pre-trained architecture.

Both of the models were trained using the Adam optimizer with the categorical cross-entropy loss function, a standard choice for multi-class classification tasks. Early stopping was also employed to both of the models to prevent overfitting by halting training when validation accuracy plateaued. Additionally, a learning rate reduction strategy dynamically adjusted the learning rate during training to optimize convergence. This systematic approach to training ensured that both models were well-prepared for the subsequent evaluation and analysis stages.

To further examine and interpret the models, two techniques were applied. First, saliency maps using Grad-CAM and Integrated Gradients were generated to visualize the regions of input images that were most influential in the model's predictions. These visualizations provided insight into whether the models relied on meaningful facial features for classification. Second, sensitivity analysis was performed using an occlusion-based approach, where specific regions of an image were systematically masked to measure the impact on the model's output. This technique highlighted the importance of specific facial regions in the models' decision-making process.

Code Availability

https://github.com/Quinnie2959/Project3/blob/7af3fbec0ff279e7255637fa0bfc9e99fbedab85/facial_emotion_recognition.ipynb

Results

Model Performance

1. General Performance

Both the CNN and ResNet50 models demonstrated strong performance on the dataset, as evidenced by their training and validation accuracy and loss plots (Figures 1 and 2). These plots indicate that both models successfully learned from the training data and generalized reasonably well to the validation set without significant overfitting.

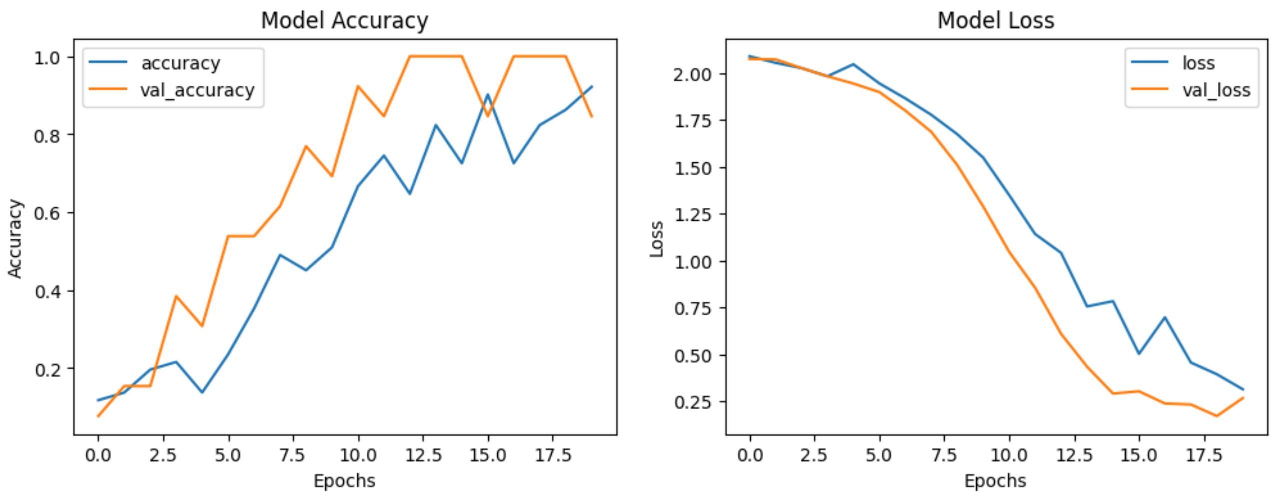

Figure 1. Model Accuracy and Loss of the CNN Model

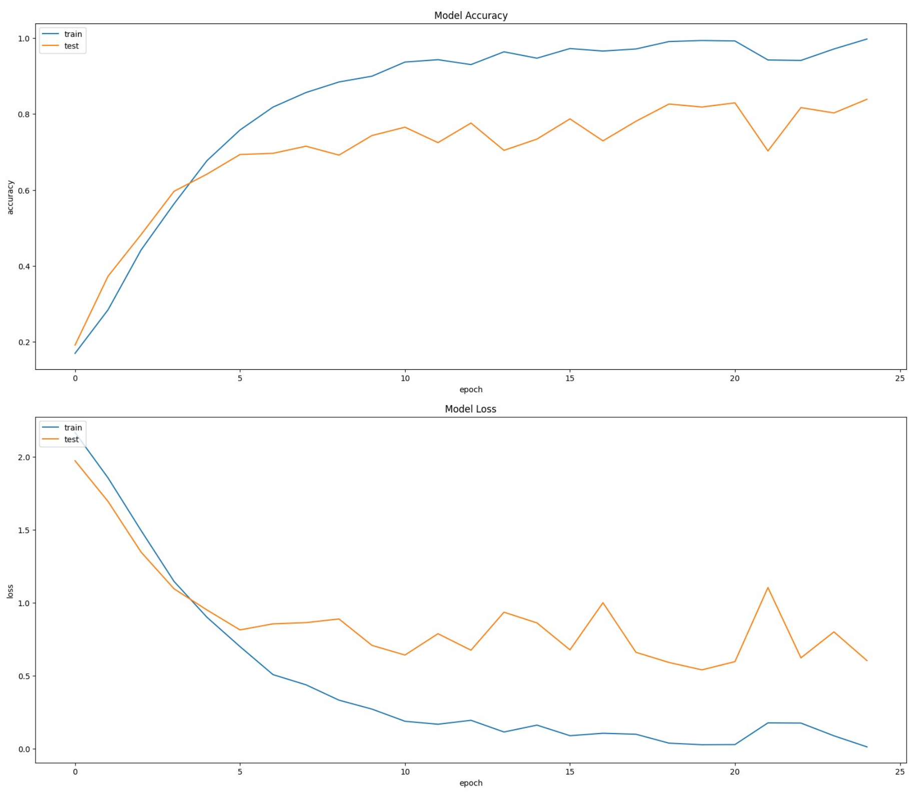

Figure 2. Model Accuracy and Loss of the ResNet50 Model

The overall test accuracies for the models provide further insights into their comparative performance. The CNN achieved a higher test accuracy of 84.62%, compared to ResNet50's accuracy of 83.88%, highlighting its superior ability to generalize to unseen data in this specific task. Additionally, the loss metrics reinforce this conclusion, with CNN exhibiting a lower test loss of 0.2672 than the 0.6047 of ResNet50, indicating a better fit to the dataset.

To illustrate the effectiveness of the CNN model in emotion recognition, Figure 3 presents a sample of its predictions alongside the corresponding true labels. This visualization demonstrates the model's ability to correctly identify diverse emotions across a variety of facial expressions, further validating its performance.

Figure 3. Predicted Emotions with the Actual Emotion Labels of theCNN Model

2. Subgroup Analysis

Subgroup analysis revealed important differences in the performance of the models across various demographic groups, particularly when considering age, gender, and country.

For the CNN, accuracy remained consistent across age groups (Under_31 and Beyond_31) and gender groups (Male and Female), demonstrating its robustness in generalizing across these demographics. However, a slight discrepancy was observed in the country groups, where the accuracy for the RU group, which is 67.86%, was marginally lower than for the Other group. This prompted further evaluation of the ResNet50 model to determine whether it could better capture variations in emotion recognition between these groups.

While ResNet50 exhibited excellent training accuracy that was near-perfect performance, its test accuracy was notably lower, particularly for the RU group. This discrepancy suggests that ResNet50 may have overfitted to the training data, failing to generalize effectively to the test set, especially for specific subgroups. The inconsistency in its predictions across subgroups highlights a potential limitation of the model in handling demographic diversity, making the CNN a more reliable choice for this dataset.

Clinical Utility Evaluation

The Positive Predictive Value (PPV) was selected as the primary metric to evaluate the clinical utility of the emotion recognition model, as it measures the reliability of the model's predictions for specific emotions. PPV reflects the proportion of positive predictions that are correct, which is particularly relevant for clinical applications like PTSD therapy, where accurate detection of heightened emotional responses such as fear and anger is crucial for monitoring patient progress.

The results demonstrate that the model achieved high PPV values (1.00) for most emotions, including Anger, Contempt, Disgust, Fear, Happy, and Surprised, indicating that when the model predicts these emotions, it does so with perfect reliability. However, the PPV for Sad was lower at 0.33, and for Neutral, the PPV was 0.00, suggesting that the model struggles to reliably identify these specific emotional states. These discrepancies may stem from the subtle nature of sadness and neutrality compared to more exaggerated expressions like anger or surprise, which are easier for the model to recognize.

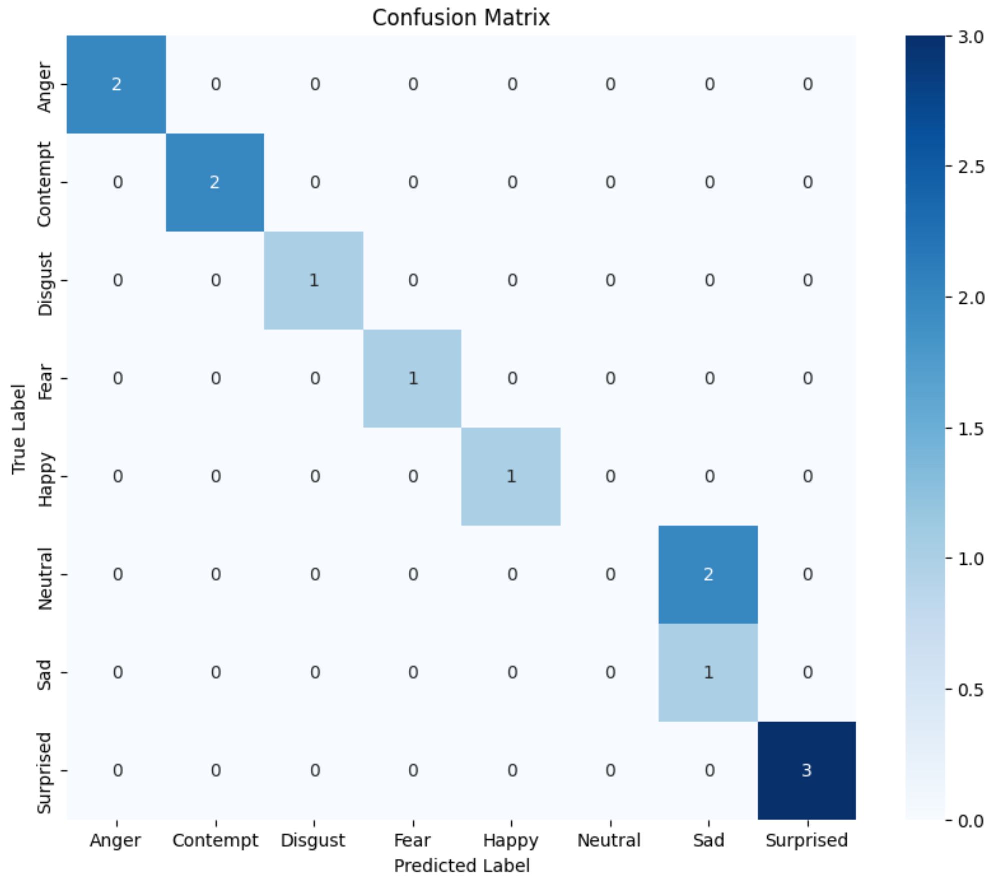

Figure 4. Confusion Matrix of the CNN Model for PPV Calculation

By prioritizing PPV, we ensure that the model's predictions can be trusted when identifying emotions that are particularly relevant to clinical outcomes. For instance, in PTSD therapy, the accurate detection of fear and anger—both of which achieved a PPV of 1.00—provides clinicians with reliable tools to monitor emotional dysregulation in patients. The results highlight the model's potential utility in clinical settings while also identifying areas for improvement, such as better recognition of subtle expressions like sadness and neutrality, through enhanced training or dataset augmentation.

Model Examination Techniques

1. Saliency Maps

Saliency maps were generated to highlight the regions of the input images that the CNN model found most influential in predicting the emotions. These maps visually emphasize the specific areas of the face where the model focused when making its predictions. In the saliency maps as shown in Figure 4, the model primarily attends to facial features such as the eyes, nose, and mouth. This is particularly evident in correctly classified cases, where the saliency regions align well with human intuition regarding facial expressions. For example, in Figure 5 (true label: 7, predicted: 7), the model accurately identified areas around the subject's mouth and eyes, which are crucial for recognizing the "surprise" expression.

The relevance of these results lies in their direct connection to model performance: saliency maps confirm that the CNN model relies on meaningful features in the input images to make its predictions, rather than irrelevant or background areas. This strengthens confidence in the model's interpretability and reliability.

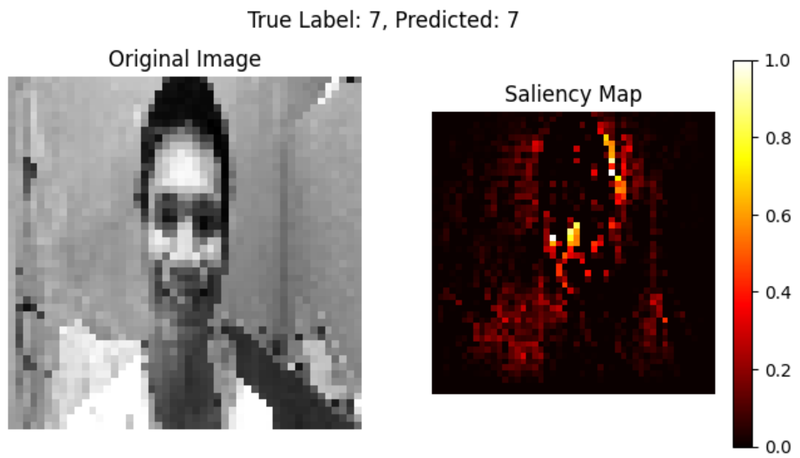

Figure 5. Saliency Map Comparing to the Original Image of a Person

2. Sensitivity Maps

Sensitivity maps, created using occlusion-based analysis, provide another layer of interpretability by systematically masking regions of the input image and observing the changes in model predictions. The resulting map of the same person as the above one was shown in Figure 6, revealing the importance of specific regions for the CNN's decision-making process. Bright regions in the sensitivity map indicated areas where occlusion led to significant changes in the model's confidence, highlighting their importance for emotion classification. For example, in Figure 6, the regions around the subject's forehead and mouth strongly influenced the model's prediction. This was also corresponding to the saliency map shown in Figure 5.

The sensitivity maps complement the saliency maps by further validating that the CNN relies on critical facial regions for classification. However, compared to saliency maps, sensitivity maps tend to be coarser, as occlusion operates on larger patches of the image. Despite this, they provide valuable insights into model robustness, showing how predictions are affected when important facial regions are altered.

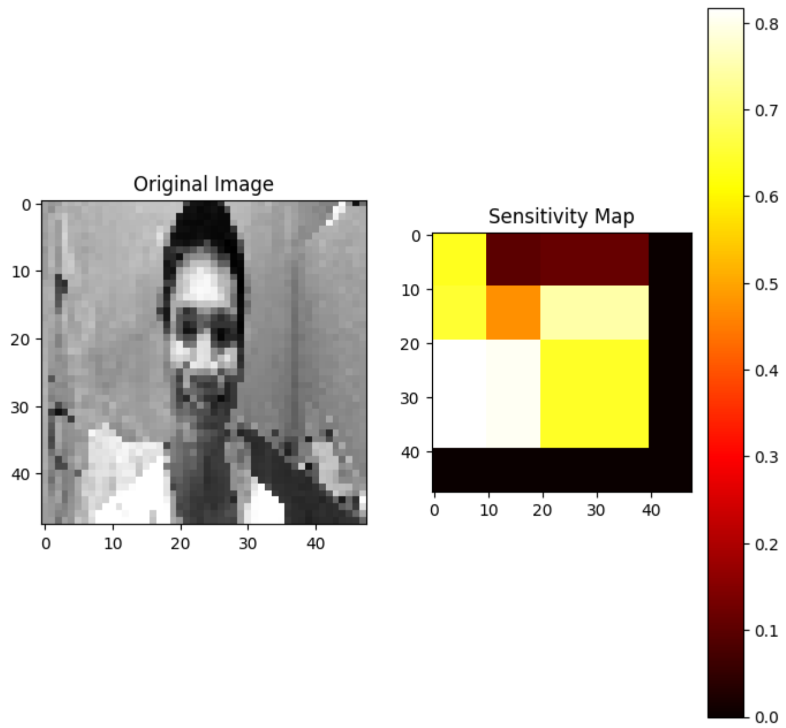

Figure 6. Sensitivity Map Comparing to the Original Image of the Same Person

Model Interpretability and Reliability Analysis

To better understand the CNN model's decision-making process and evaluate its robustness, saliency maps and sensitivity maps were applied as model examination techniques. These methods visualize the regions of the input images that contribute most to the model's predictions, providing critical insights into its interpretability and performance.

The relevance of these maps to model performance is clear: both saliency and sensitivity maps show that the CNN consistently focuses on key facial features such as the eyes, mouth, and forehead—regions that are intuitively important for recognizing emotions. For example, in correctly predicted cases, bright regions on the maps highlight areas around the subject's expressions, validating the model's ability to rely on meaningful patterns rather than irrelevant background information. This alignment with human expectations supports the model's strong overall performance.

The interpretability offered by these methods is particularly significant at the case level. By visualizing the decision-making process, these maps help explain individual predictions, building transparency and trust in the model. This is especially critical in real-world applications where understanding why a model makes a certain prediction can inform practical decision-making. If such interpretability were lacking, it would be difficult to ensure the model's fairness and reliability.

Furthermore, the examination techniques provide valuable insights into the reliability and robustness of the CNN under data shifts, such as demographic variations. In subgroup analyses, the CNN showed slight performance differences in the RU group compared to Other, suggesting that specific variations in facial expressions or image quality might exist across these groups. However, the saliency and sensitivity maps confirm that the model continues to rely on relevant facial features, reinforcing its robustness. These visualizations highlight areas for potential improvement, ensuring the model can handle diverse and evolving real-world datasets more effectively.

Conclusion

This study explored the potential of deep learning models, specifically a custom Convolutional Neural Network (CNN) and ResNet50, for recognizing facial emotions as a step toward understanding emotional responses in clinical settings, particularly for individuals with Post-Traumatic Stress Disorder (PTSD). Individuals with PTSD often exhibit heightened emotional responses such as fear and anger when exposed to traumatic triggers, and monitoring these responses can provide valuable insights into their emotional regulation and therapeutic progress.

Our results demonstrated that the CNN model outperformed ResNet50, achieving better overall accuracy and more consistent performance across demographic subgroups, including gender and age. However, subgroup analysis revealed a slight decrease in accuracy for the RU country group. This discrepancy raises questions about the impact of cultural and ethnic factors on facial emotion expression and recognition. Variations in how emotions are displayed or perceived across different cultural contexts may introduce biases into the model, emphasizing the importance of using diverse and representative datasets to mitigate such challenges.

Although the ResNet50 model exhibited strong training accuracy, its poor generalization to the test set, particularly in subgroup analyses, highlights the issue of overfitting. Future improvements, such as fine-tuning deeper layers, applying stronger regularization, or using smaller architectures like MobileNet, could enhance ResNet's performance and its ability to generalize across diverse populations. Additionally, targeted data augmentation strategies tailored to underrepresented subgroups may help address imbalance and improve robustness.

The saliency maps and sensitivity analyses applied to the CNN model provided critical insights into its interpretability, confirming that the model consistently focused on relevant facial regions (e.g., eyes, mouth) when predicting emotions. These methods enhance trust in the model's decision-making process and demonstrate its potential for clinical applications, where explainability is essential. By visualizing the model's behavior, we can ensure that it reliably identifies emotional cues, which is particularly important when monitoring emotional changes in individuals with PTSD during therapy sessions.

One of the key limitations of this study lies in the size and diversity of the dataset. While the models performed reasonably well, the small dataset limited their ability to learn robust and generalizable features. Furthermore, the scarcity of publicly available datasets that include both facial images and ethnicity poses a significant challenge. For clinical applications, larger and more diverse datasets are essential to ensure that emotion recognition systems are fair, unbiased, and applicable across cultural and demographic groups.

In conclusion, this study highlights the potential of facial emotion recognition models as tools for monitoring emotional responses in individuals with PTSD. The models, particularly the CNN, provide reliable performance and interpretability, enabling clinicians to observe subtle changes in emotional expressions that may otherwise go unnoticed. While challenges such as dataset limitations and subgroup performance disparities remain, future efforts to improve model robustness and expand datasets will be critical. Emotion recognition systems hold promise as valuable aids in clinical therapy, offering objective measures to track emotional regulation and therapeutic progress, ultimately contributing to better mental health outcomes for individuals with PTSD.

Checklist for supervised clinical ML study

Before paper submission			
Study design (Part 1)	**Completed: page number**		**Notes if not completed**
The clinical problem in which the model will be employed is clearly detailed in the paper.	☐	1	
The research question is clearly stated.	☐	1-2	
The characteristics of the cohorts (training and test sets) are detailed in the text.	☐	3	
The cohorts (training and test sets) are shown to be representative of real-world clinical settings.	☐	3	
The state-of-the-art solution used as a baseline for comparison has been identified and detailed.	☐	1	
Data and optimization (Parts 2, 3)	**Completed: page number**		**Notes if not completed**
The origin of the data is described and the original format is detailed in the paper.	☐	2	
Transformations of the data before it is applied to the proposed model are described.	☐	2-3	

The independence between training and test sets has been proven in the paper.	☐	3	
Details on the models that were evaluated and the code developed to select the best model are provided.	☐	4	
Is the input data type structured or unstructured?	☐ Structured		☐ Unstructured
Model performance (Part 4)	**Completed: page number**		**Notes if not completed**
The primary metric selected to evaluate algorithm performance (eg: AUC, F-score, etc) including the justification for selection, has been clearly stated.	☐	5-6	
The primary metric selected to evaluate the clinical utility of the model (eg PPV, NNT, etc) including the justification for selection, has been clearly stated.	☐	7	
The performance comparison between baseline and proposed model is presented with the appropriate statistical significance.	☐	5-6	
Model Examination (Parts 5)	**Completed: page number**		**Notes if not completed**
Examination Technique 1[a]	☐	8	
Examination Technique 2[a]	☐	9	
A discussion of the relevance of the examination results with respect to model/algorithm performance is presented.	☐	9	

A discussion of the feasibility and significance of model interpretability at the case level if examination methods are uninterpretable is presented.	☐	9	
A discussion of the reliability and robustness of the model as the underlying data distribution shifts is included.	☐	10	
*Common examination approaches based on study type: * For studies involving exclusively structured data coefficients and sensitivity analysis are often appropriate * For studies involving unstructured data in the domains of image analysis or NLP: saliency maps (or equivalents) and sensitivity analysis are often appropriate			
Reproducibility (Part 6): choose appropriate tier of transparency			**Notes**
Tier 1: complete sharing of the code		☐	
Tier 2: allow a third party to evaluate the code for accuracy/fairness; share the results of this evaluation		☐	
Tier 3: release of a virtual machine (binary) for running the code on new data without sharing its details		☐	
Tier 4: no sharing		☐	

PPV: Positive Predictive Value

NNT: Numbers Needed to Treat

[a] Common examination approaches based on study type: for studies involving exclusively structured data, coefficients and sensitivity analysis are often appropriate; for studies involving unstructured data in the domains of image analysis or natural language processing, saliency maps (or equivalents) and sensitivity analyses are often appropriate. Select 2 from this list or chose an appropriate technique, document each technique used on the appropriate line above.

References

Chappidi Suneetha, D. R. A. (2022). A Survey Of Machine Learning Techniques On Speech Based Emotion Recognition And Post Traumatic Stress Disorder Detection. *Neuroquantology*, *20*(14), 69-79.

Gapen, M. A. (2009). *Facial emotion recognition difficulties in individuals with PTSD symptoms* (Doctoral dissertation, Emory University).

Kaggle. *Facial Emotion Recognition Dataset.* Retrieved from https://www.kaggle.com/datasets/tapakah68/facial-emotion-recognition/data

Savchenko, A. V., Savchenko, L. V., & Makarov, I. (2022). Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, *13*(4), 2132-2143.