

## Article

# Human-Centered and Quantitative Explainability Evaluation of Facial Emotion Recognition for Trustworthy Mental Health Monitoring

Dina Shehada <sup>1,2,\*</sup> , Hissam Tawfik <sup>1,3,\*</sup> , Ahmed Bouridane <sup>4</sup>  and Abir Hussain <sup>1</sup> 

<sup>1</sup> Department of Electrical Engineering, University of Sharjah, Sharjah P.O. Box 27272, United Arab Emirates; abir.hussain@sharjah.ac.ae

<sup>2</sup> Department of Engineering and Information Technology, University of Dubai, Dubai P.O. Box 14143, United Arab Emirates

<sup>3</sup> School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds LS1 3HE, UK

<sup>4</sup> College of Computing and Informatics, University of Sharjah, Sharjah P.O. Box 27272, United Arab Emirates; abouridane@sharjah.ac.ae

\* Correspondence: dshehada@ud.ac.ae (D.S.); htawfik@sharjah.ac.ae (H.T.)

## Abstract

Facial emotion recognition (FER) systems can serve as a valuable non-invasive tool for assessing emotional states linked to mental health. However, two main issues hinder their adoption in clinical settings, including privacy concerns inherent to centralized data processing and the lack of transparent decision-making processes. This paper proposes a privacy-preserving and explainable FER framework that implements a collaborative distributed training approach for lightweight convolutional neural network (CNN) architectures across three heterogeneous datasets: RAF-DB, ExpW, and FER2013. SHapley Additive exPlanations (SHAP) guided the optimization of CNN filter configurations, prioritizing high accuracy, cross-dataset generalization, and interpretable, trustworthy explanations. A multidimensional explainability evaluation framework is developed that combines perturbation-based faithfulness and feature localization metrics into a Global Explanation Quality Score (GEQS) for quantitative assessment of explanation quality. A qualitative user study was conducted to assess alignment between human perception and quantitative explainability metrics. Guided by explainability-driven evaluation, a lightweight CNN achieves 74.3% mean accuracy, highlighting its effectiveness for cross-dataset generalization. Results demonstrate alignment between quantitative GEQS and human evaluation, with both identifying the same model architecture. However, qualitative analysis shows that highlighting emotion-relevant facial features does not always ensure user trust. The practical viability of the proposed FER system in resource-constrained clinical environments is demonstrated through implementation on Raspberry Pi 4 with integrated SHAP explainability, achieving 60 ms inference time.



Academic Editor: Paolo Bellavista

Received: 14 January 2026

Revised: 20 February 2026

Accepted: 21 February 2026

Published: 26 February 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

**Keywords:** explainable AI; SHapley Additive exPlanations; quantitative evaluation; human-based evaluation; facial emotion recognition

## 1. Introduction

Facial emotion recognition (FER) systems demonstrate strong potential in assistive healthcare technologies, including mental health monitoring and psychological assessment [1]. These systems provide non-invasive solutions for continuous monitoring of emotional states associated with stress, depression, emotional dysregulation, and overall psychological well-being through facial expression analysis. Facial expressions

represent a natural and passive modality for assessing emotional and stress-related states, enabling assistive technologies that support early detection and timely intervention. However, deploying FER systems in real-world settings introduces two critical challenges. First, conventional FER approaches rely on centralized data collection and training, which require aggregating large volumes of sensitive facial data. Such practices raise serious privacy concerns, as facial data are highly personal and subject to strict ethical and regulatory constraints in healthcare environments [2]. Second, most state-of-the-art FER models operate as black-box systems, offering limited interpretability. A trustworthy system should be explainable, providing faithful reasons for its predictions, and interpretable, making its decisions easy to understand [3]. The adoption of FER systems in clinical settings therefore remains limited due to these fundamental issues. In healthcare applications, where system outputs may directly influence diagnosis, monitoring, or therapeutic decisions, insufficient explainability undermines trust, accountability, and safe deployment. Consequently, FER systems designed for mental health applications must simultaneously ensure strong privacy preservation and reliable explainability to be viable for real-world clinical use.

Federated learning (FL) was introduced as a distributed collaborative training approach that does not require centralized data collection, thereby preserving user privacy by enabling model training without sharing raw data [4]. FL has been explored in a limited number of facial emotion recognition (FER) studies, with most existing approaches primarily focusing on predictive performance, while explainability within these frameworks remains largely under-explored. Moreover, FER systems in federated settings are challenged by heterogeneous and non-independent and identically distributed (non-IID) data distributions across clients [5], where local data differ substantially in distributions and demographic characteristics and capture conditions inherent to distributed learning settings [6]. Such variability hinders stable model convergence, reduces generalization capability, and undermines the reliability of the learned representations.

Explainable artificial intelligence (XAI) techniques have been increasingly applied to FER models to visualize salient facial regions and interpret predictions. However, most prior work treats explainability as a post hoc visualization tool relying primarily on qualitative inspection [7]. Such approaches provide limited insight into explanation fidelity, spatial focus, and cross-dataset generalization. Furthermore, explainability is rarely leveraged as a criterion for architecture optimization and model design, especially in federated and resource-constrained settings.

To address these gaps, this work proposes and systematically evaluates a privacy-preserving and explainability-guided FER framework for continuous emotional state monitoring in mental health applications. Rather than optimizing accuracy alone, explainability is integrated directly into the model selection process using SHapley Additive exPlanations (SHAP) [8]. Lightweight convolutional neural network (CNN) architectures are trained under a federated learning setup using heterogeneous real-world datasets, enabling cross-dataset generalization while maintaining strict client-side privacy. A multidimensional explainability evaluation framework is introduced to systematically assess explanation quality using both quantitative metrics and qualitative human-centered assessment. The main contributions of this work are summarized as follows:

- This paper introduces the Global Explanation Quality Score (GEQS), a multidimensional explainability evaluation framework that formalizes explanation quality through the unified integration of perturbation-based faithfulness and feature localization metrics. The framework is validated through human-centered evaluation, showing strong agreement between quantitative scores and human judgment of explanations. By using explainability to guide model design, GEQS enables guided optimization of CNN archi-

tures, prioritizing models that achieve strong predictive performance while providing trustworthy explanations aligned with human perception.

- We develop a privacy-preserving federated learning approach for lightweight CNN models across heterogeneous FER datasets (RAF-DB, ExpW, and FER2013). Experimental results show strong cross-dataset performance, achieving a mean cross-dataset accuracy of 74.3% and improvements of 10.6% on FER2013 and 6.5% on RAF-DB over competitive baselines, while preserving data privacy through decentralized training.
- We validate the practical viability of the proposed federated explainable FER framework for clinical mental health applications through deployment on a Raspberry Pi 4 edge device. The system achieves real-time inference of 60 ms, while on-device SHAP explanations demonstrate the feasibility of integrating explainable FER into edge-based intelligent healthcare systems.

The remainder of this paper is organized as follows. Section 2 reviews related works and highlights the main gaps. Section 3 presents the proposed methodology. Section 4 presents the experimental results. Section 5 provides a discussion of the results. Section 6 presents a comparison with cross-dataset FER systems. Section 7 demonstrates the deployment viability of the proposed system on edge hardware.

Finally, Section 8 concludes the paper and highlights the future work.

## 2. Related Work

This section reviews key advancements in facial emotion recognition (FER) systems, including conventional deep learning-based approaches, application-specific frameworks, and recent federated learning techniques for privacy-preserving FER.

Recent deep learning approaches in FER have primarily focused on improving recognition accuracy through enhanced feature representation and attention mechanisms. For example, Minaee et al. [9] proposed an attention-based CNN using Spatial Transformer Networks to focus on salient facial regions such as the eyes and mouth, achieving accuracies of 70.0% on FER2013, 98.0% on CK+, 99.3% on FERG, and 92.8% on JAFFE. However, since the model was trained and evaluated separately on each dataset, its ability to generalize across domains remained limited.

Several works have explored complex CNN architectures to improve FER accuracy in specific applications. Hou et al. [10] developed a student engagement assessment system combining Multi-Task Cascaded Convolution Neural Networks (MTCNN) for face detection with dual feature extraction using VGG16 and ECANet, achieving 67.4% accuracy on FER2013 and 99.2% on CK+. A Multiple Branch Cross-Connected CNN (MBCC-CNN) was proposed in [11]. However, the high computational demands of these architectures present a challenge on edge devices. Lightweight FER models such as the DenseNet proposed by Zhou et al. [12] present a compact architecture along with Histogram of Oriented Gradients (HOG) features. This combination achieved 67% accuracy on FER2013. Kim et al. [13] applied Log-Level Threshold Quantization (LLTQ) to reduce computational cost and obtained 86.5% accuracy on FER+, although subsequent studies reported that energy consumption remained relatively high [14].

Ensemble and dual-network strategies have also been explored. Saurav et al. [15] proposed a dual CNN model achieving accuracies of 72.8% on FER2013 and 98.5% on CK+. The ensemble-based EmNet [16] improved performance across multiple datasets but at the cost of increased model size and complexity. Zhao et al. [17] focused on model compression for lightweight emotion recognition; however, the use of a non-public dataset limited reproducibility and comparative evaluation.

Cross-dataset learning and fine-grained FER have been investigated to improve robustness. Huang et al. [18] combined CNNs with squeeze-and-excitation residual networks,

achieving 83.4% accuracy on RAF-DB after fine-tuning on AffectNet. Dias et al. [19] explored cross-dataset training using VGG-Face and a random patch strategy, highlighting persistent dataset bias and challenges associated with posed expressions. Liang et al. [20] proposed MSAU-Net for fine-grained FER, achieving improved accuracy at the expense of increased computational cost.

Recent studies have also examined real-time and ensemble-based FER systems. Reghunathan et al. [21] compared several transfer learning architectures, achieving a maximum accuracy of 73.6% on FER2013 but with high computational overhead. Gupta et al. [22] developed a learner engagement detection system using ResNet-50, achieving 73.4% accuracy on FER2013 and 89.56% on CK+. Despite their strong performance, these models continue to face challenges related to efficiency, dataset diversity, and real-world generalization.

FER has also been increasingly applied in assistive and healthcare contexts to support mental health assessment and inclusivity. Ma et al. [23] proposed a depression detection framework based on facial regions of interest, Discrete Wavelet Transform features, and Bayesian Networks, achieving 81.7% accuracy despite limited data availability. Shangguan et al. [24] introduced a dual-stream Multiple Instance Learning framework for depression detection from facial videos, achieving 74.7% accuracy on a small dataset. Similarly, Gue et al. [25] retrained EfficientNet for binary depression classification using emotion groupings, but performance was constrained by subjective emotion mapping and dataset size.

Ramis et al. [26] evaluated FER models for individuals with intellectual disabilities using AlexNet and VGG19 across multiple datasets. While accuracies of 86–89% were achieved on typical populations, performance dropped significantly on the MuDERI dataset, highlighting inclusivity challenges. Using LIME and RISE, the study revealed that existing models struggle to capture subtle expressive variations.

For children's FER in online learning environments, Rathod et al. [27] introduced a small-scale dataset and achieved 89.31% accuracy using ResNet152V2. Although Score-Grad provided more interpretable visualizations, generalization was limited by dataset size. Ruangdit et al. [28] proposed a multimodal FER system combining facial and speech modalities, achieving 85.8% facial recognition accuracy, but lacking thorough interpretability and clinical validation.

Finally, Hettiarachchi et al. [29] examined gender effects in multimodal emotion recognition and showed that gender-specific models often outperform gender-neutral ones, particularly in speech-based FER.

### *Limitations of Existing FER Approaches*

While the field of FER has made significant progress, its integration with FL remains largely underexplored. Recent studies have investigated explainable FER systems from different perspectives [30–40]. However, these works primarily focus on applying XAI techniques to visualize or interpret FER model predictions, without providing a rigorous evaluation of explanation quality. Moreover, none of these approaches consider federated learning settings, where data privacy, client heterogeneity, and decentralized training introduce additional challenges for explainability.

Many existing FER studies, particularly those targeting mental health applications, either do not incorporate explainability at all or use XAI only as a basic visualization tool, which limits the transparency and trustworthiness of their decisions [23–25]. Even when explainability techniques were applied, explanation quality was usually assessed visually rather than with standardized metrics [26,27,30–34,36,37]. In parallel, most FER systems rely on computationally expensive deep learning architectures that are unsuitable for resource-constrained or edge environments [25–28,33–37,39,41–43]. Lightweight model designs that optimize accuracy, efficiency, and explainability remain largely unexplored.

Furthermore, many FER approaches show limited generalization to real-world and clinical settings, as they are often evaluated on single datasets or controlled, posed emotion data that fail to capture real-world variability [30–33,38,40]. Although federated learning has been explored in a few FER studies, these setups are typically limited in scale, involving a small number of clients or acted datasets, which reduces their practical relevance [42,44,45]. As a result, truly privacy-preserving, cross-dataset FER systems with reliable explainability and real-world deployability remain an open challenge.

These limitations hinder the development of reliable and deployable FER systems for mental health applications. To address them, this work presents a lightweight FER framework that integrates federated learning with a rigorous quantitative evaluation of explainability. The proposed approach is computationally efficient, interpretable, and privacy-preserving, enabling reliable deployment in real-world mental health monitoring scenarios. The detailed methodology is described in the following section.

### 3. Proposed Methodology

#### 3.1. FER for Continuous Monitoring of Emotional State in Healthcare Applications

FER systems are used as non-invasive tools for the continuous monitoring of emotional states in healthcare applications, supporting pain assessment, recovery tracking, and mental health monitoring [46,47]. Such systems can support early detection and timely intervention by tracking symptoms related to stress, depression, emotional dysregulation, and overall psychological well-being. According to research in clinical psychology anger, happy, sad, and neutral emotions are markers of emotional dysregulation and mood swings, capturing both positive and negative states [48,49]. The neutral class serves as an important reference state for typical emotional expression [50,51], which is particularly informative for identifying anhedonia [52].

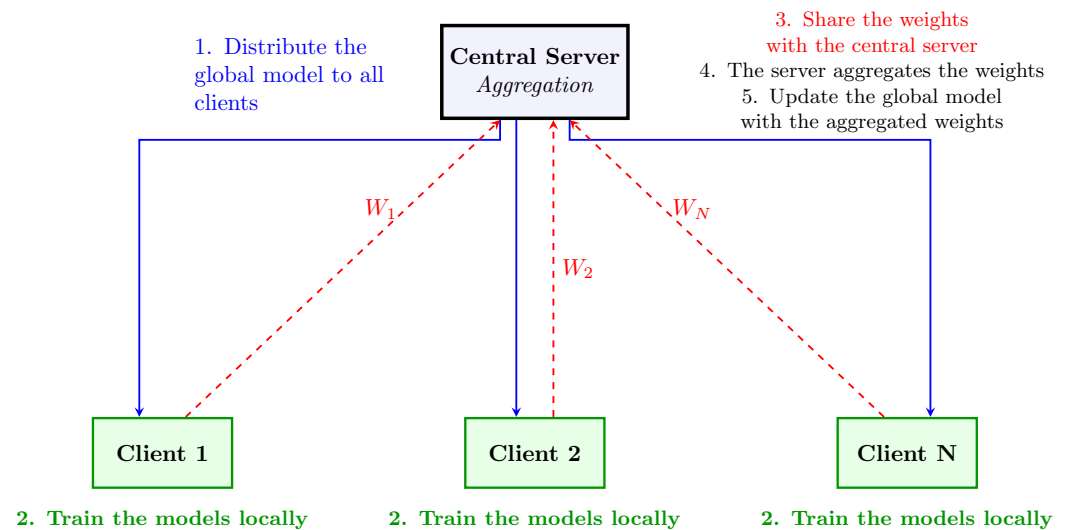
While sadness is a primary symptom of depression and is linked to increased psychological vulnerability, reduced experience of happiness is a fundamental characteristic of anxiety and depressive disorders, reflecting blunted responses to positive events [53,54], which has been associated with reduced psychological well-being [55]. Additionally, emotional dysregulation, stress, and poor mental health have also been linked to anger [56,57]. Furthermore, the same label space has been used in several prior FER systems designed for mental health and assistive applications [28,34].

#### 3.2. Collaborative Learning on Diverse Distributed Data for Generalized Models

To support privacy-preserving deployment while leveraging diverse user data, we adopt a federated learning setup for training the FER model. Multiple distributed clients participate in collaborative model training under the coordination of a central server. Each client retains its local facial data, trains a lightweight CNN model on-device, and transmits only its model updates to the server. The server aggregates the updates to train the global model, which is redistributed to all clients. This iterative training process is repeated over multiple communication rounds, enabling learning from diverse, non-centralized data sources while preserving user privacy and promoting improved generalization across populations. An overview of the proposed FL-based FER framework is shown in Figure 1.

A heterogeneous training environment was utilized to mimic real-world diversity; therefore, three widely used FER datasets capturing natural expressions in the wild and providing rich demographic and environmental diversity were employed for training and testing the framework. The RAF-DB dataset [58] contains 12,488 web-collected images; 9983 were used for training and 2505 for testing, spanning a wide range of ages, ethnic backgrounds, genders, head poses, and illumination conditions. The FER2013 dataset [59] provides 16,482 grayscale images captured under challenging real-world conditions, includ-

ing occlusions and non-frontal views, with 13,269 samples used for training and 3213 for testing. The ExpW dataset [60] further contributes large-scale variability with 79,650 images containing both posed and spontaneous expressions, of which 55,739 images were used for training and 23,911 for testing.



**Figure 1.** Overview of the proposed federated learning-based privacy-preserving FER system, demonstrating the iterative steps of global model distribution, local training, and secure aggregation.

As shown in Table 1, the class distributions vary across datasets. FER2013 exhibits a relatively balanced distribution across the four emotions, whereas ExpW and RAF-DB show moderate class imbalance, particularly with a higher proportion of happy and neutral samples compared to anger and sadness.

**Table 1.** Per-class dataset distributions for RAF-DB, FER2013, and ExpW.

| Dataset      | Class   | Train  | Test   | Total   |
|--------------|---------|--------|--------|---------|
| FER2013      | Anger   | 3365   | 776    | 4141    |
|              | Happy   | 3336   | 833    | 4169    |
|              | Sad     | 3192   | 798    | 3990    |
|              | Neutral | 3376   | 806    | 4182    |
| ExpW         | Anger   | 2576   | 1095   | 3671    |
|              | Happy   | 21,569 | 8968   | 30,537  |
|              | Sad     | 7343   | 3216   | 10,559  |
|              | Neutral | 24,251 | 10,632 | 34,883  |
| RAF-DB       | Anger   | 705    | 162    | 867     |
|              | Happy   | 4772   | 1185   | 5957    |
|              | Sad     | 1982   | 478    | 2460    |
|              | Neutral | 2524   | 680    | 3204    |
| <b>Total</b> |         | 78,991 | 29,629 | 108,620 |

Training samples from FER2013, RAF-DB, and ExpW were merged and shuffled to form a heterogeneous dataset that was distributed across three clients. Each client was assigned an equal number of unique, non-overlapping samples from each emotion class. While label distributions are nearly uniform across clients, heterogeneity arises from the integration of multiple datasets exhibiting variations in data, reflecting cross-domain variability. No class rebalancing was applied, preserving the natural imbalance in the merged distribution to reflect realistic federated scenarios.

Before training, all facial images were preprocessed to ensure consistency across datasets. Images were converted to grayscale when necessary and resized to  $48 \times 48$  pixels, reducing computational overhead for edge deployment. Pixel values were normalized to  $[0, 1]$ .

Each client trained its local model for 15 epochs per communication round using the Adam optimizer (batch size = 64, learning rate = 0.001) for 20 rounds until convergence was reached. Model updates were aggregated using the Federated Averaging (FedAvg) aggregation strategy [4], where client contributions were weighted according to their sample distributions. The aggregation process is formalized as follows [4]:

$$w^t = \sum_{k=1}^K \frac{n_k}{n} w_k^t \quad (1)$$

where  $K$  denotes the number of clients,  $n_k$  represents the number of samples of client  $k$ ,  $w_k^t$  are the local model parameters from client  $k$  after local training, and  $w^t$  represents the aggregated global model parameters at round  $t$ . The complete federated training configuration details, including data partitioning strategy, aggregation parameters, and random seeds, are provided in Appendix A.3 (Table A2).

### 3.3. Explainability-Guided Selection of Lightweight CNN Architectures with SHAP

In this work, model optimization and selection are driven by explainability in addition to accuracy. Multiple lightweight CNN architectures with different feature extraction filter configurations were evaluated in terms of their prediction accuracy. Additionally, SHAP explainability method was used to evaluate how well each model concentrates on facial regions that are relevant to the corresponding emotions.

Three-layer models with filters (16-32-64), (32-64-128), (64-128-256), (128-256-512), and (128-128-256), and four-layer models with (16-32-64-128), (32-64-128-256), and (64-128-256-512) were evaluated. Parameter counts ranged from 164 K to 3.6 M. These configurations enable direct comparison with well-known efficient architectures suitable for edge devices such as MobileNetV2 (3.5 M parameters,  $\approx 14$  MB) [61], SqueezeNet (1.24 M parameters), and ShuffleNet (1.0–2.3 M parameters) [62]. Our prior work revealed that Grad-CAM++ exhibits a fundamental limitation when applied to deeper CNN architectures with four or more convolutional layers [63]. The resulting last-layer heatmaps were diffuse, spatially unstable, and misaligned with relevant facial features [64]. On the other hand, SHAP provides more stable, pixel-level feature attribution, enabling consistent identification of facial regions; therefore, we utilize SHAP to prioritize stable and meaningful explanations.

Algorithm 1 demonstrates the general architecture of the configurable lightweight CNN. An input of  $48 \times 48$  a grayscale facial image is processed through  $B$  convolutional blocks, where  $B$  corresponds to the depth of the filter configuration  $\{F_1, F_2, \dots, F_B\}$ .  $F_b$  denotes the number of convolutional filters used in block  $b$ . Each block applies a  $3 \times 3$  convolution with  $F_b$  filters, followed by  $2 \times 2$  max pooling. These filter configurations were varied to compare different lightweight architectural designs for optimization. After feature extraction, the resulting feature maps are flattened and passed through two fully connected layers, with dropout applied for regularization. The final classification layer uses a softmax activation to predict the target emotions.

---

**Algorithm 1** Configurable lightweight CNN architecture with convolutional blocks for facial emotion recognition
 

---

**Require:**  $\mathcal{I} \in \mathbb{R}^{48 \times 48}$  ▷ Grayscale input facial image  
**Require:**  $\{F_1, F_2, \dots, F_B\}$  ▷ Filter configuration, where  $B \in \{3, 4\}$   
**Require:**  $k_{conv} = 3, k_{pool} = 2, p_{drop} = 0.2, C = 4$   
**Ensure:** Predicted emotion class  $c$  and probability vector  $\mathbf{p}$

```

1:  $\mathbf{x}^{(0)} \leftarrow \mathcal{I}$ 
2: for  $b = 1$  to  $B$  do
3:    $\mathbf{x}^{(b)} \leftarrow \text{Conv2D}(\mathbf{x}^{(b-1)}, F_b, k_{conv})$ 
4:    $\mathbf{x}^{(b)} \leftarrow \text{ReLU}(\mathbf{x}^{(b)})$ 
5:    $\mathbf{x}^{(b)} \leftarrow \text{MaxPool}(\mathbf{x}^{(b)}, k_{pool})$ 
6: end for
7:  $\mathbf{f} \leftarrow \text{Flatten}(\mathbf{x}^{(B)})$ 
8:  $\mathbf{h}_1 \leftarrow \text{ReLU}(\text{FC}(\mathbf{f}, 256))$ 
9:  $\mathbf{h}_1 \leftarrow \text{Dropout}(\mathbf{h}_1, p_{drop})$ 
10:  $\mathbf{h}_2 \leftarrow \text{ReLU}(\text{FC}(\mathbf{h}_1, 128))$ 
11:  $\mathbf{h}_2 \leftarrow \text{Dropout}(\mathbf{h}_2, p_{drop})$ 
12:  $\mathbf{z} \leftarrow \text{FC}(\mathbf{h}_2, C)$ 
13:  $\mathbf{p} \leftarrow \text{Softmax}(\mathbf{z})$ 
14:  $c \leftarrow \arg \max(\mathbf{p})$ 
15: return  $(c, \mathbf{p})$ 

```

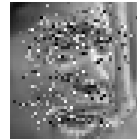
---

### 3.4. Quantitative Explainability Evaluation with SHAP

Explainability quality was evaluated using SHAP [8], which computes the contribution of each pixel to the model's prediction. SHAP explanations were generated using the DeepExplainer. All images were preprocessed to  $48 \times 48$  grayscale and normalized to the range  $[0, 1]$ . DeepExplainer approximates SHAP values using a connection between DeepLIFT and Shapley values, enabling efficient computation for deep networks [8,65]. A background set of 20 samples was used to approximate the expected model output while maintaining computational efficiency. Furthermore, insertion and deletion curves were computed over 20 progressive steps, each modifying approximately 5% of the total pixels, following the step-wise protocol [66]. To ensure fair and reliable evaluation of explainability, SHAP analysis was conducted on samples meeting two criteria: correct ground-truth classification and unanimous prediction agreement across all evaluated CNN architectures. The SHAP test set consisted of 775 anger, 800 happy, 798 sad, and 800 neutral samples. A target of 800 samples per emotion was set; when there were insufficient consensus samples, the maximum qualifying number was used. The samples were drawn from FER2013, RAF-DB, and ExpW as follows: anger (376, 86, 313), happy (588, 212), neutral (381, 406, 13), and sad (252, 159, 387), respectively.

SHAP calculates attribution values for each pixel in the input image. The top- $k\%$  highest-valued pixels correspond to the most significant features. The top- $k\%$  threshold was chosen to capture relevant features with minimal noise. Table 2 demonstrates the different  $k$  thresholds used to generate the SHAP attribution maps. As shown, the top 20% SHAP values had more noise due to the inclusion of background and irrelevant regions, while the top 5% were unable to capture all relevant features. The top 10% threshold ( $k = 0.1$ ) was found to provide a balance between capturing discriminative features while minimizing noise [64]. Therefore,  $k$  was set to 0.1 for subsequent analysis.

**Table 2.** Visualization of the top- $k\%$  threshold selection using SHAP attribution maps.

|       | Original                                                                          | Top-5%<br>( $k = 0.05$ )                                                           | Top-10%<br>( $k = 0.1$ )                                                            | Top-20%<br>( $k = 0.2$ )                                                            |
|-------|-----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| Anger |  |  |  |  |
| Sad   |  |  |  |  |

Two complementary sets of metrics were employed to evaluate the faithfulness and localization of identified important features. Perturbation-based metrics examine the change in model predictions as important features are removed from or added to the input sample. These metrics include Insertion Area Under the Curve (IAUC), Deletion Area Under the Curve (DAUC), and Average Drop in Deletion (ADD). IAUC measures the increase in model confidence as important pixels are gradually added, while DAUC measures the decrease in model confidence as the same pixels are gradually removed. ADD measures the drop in prediction accuracy after masking the important features. Feature localization metrics evaluate the Importance Focus (IF) and Spatial Focus (SF) of important features. IF quantifies how concentrated the explanations are within a few main features, and Spatial Focus (SF) measures the spatial clustering of important regions.

To identify the best-performing model, all models were trained and tested under identical conditions, as explained in the previous section, and were subsequently evaluated based on classification accuracy and cross-dataset generalization performance. The leading candidates were assessed using the quantitative explanation metrics. Finally, qualitative evaluation through human surveys examined the interpretability and trustworthiness of explanations by collecting perception ratings on the relevance and clarity of SHAP visualizations. Detailed formulations of the utilized quantitative metrics are discussed as follows.

#### 3.4.1. Insertion Area Under Curve

To evaluate the faithfulness of SHAP explanations, the insertion analysis measures how the model's confidence changes as pixels are progressively revealed in order of decreasing importance [66,67]. This assesses how effectively the explanation highlights regions that positively contribute to the model's decision.

Let the input grayscale image be  $X \in \mathbb{R}^{H \times W}$  and its SHAP attribution map be  $A \in \mathbb{R}^{H \times W}$ , normalized to  $[0, 1]$ . The attribution map is flattened and sorted in descending order. For each step  $t = 1, 2, \dots, T_{\text{steps}}$ , we compute  $\rho_t = \frac{t}{T_{\text{steps}}}$  which represents the portion of pixels to be revealed in the current step.  $k_t = \lfloor \rho_t \cdot N \rfloor$ , where  $N$  is the total number of pixels. Then the image pixels are updated as follows [66]:

$$X_t^{(\text{ins})}(i, j) = \begin{cases} X(i, j), & \text{if pixel rank} \leq k_t \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Starting from a blank (zero-initialized) image, pixels are progressively revealed in order of decreasing importance rank. The prediction probability  $p_t^{(\text{ins})}$  for class  $c$  is computed for the updated input image at each step.

$$p_t^{(\text{ins})} = P_c(X_t^{(\text{ins})}) \quad (3)$$

With each step, more pixels are revealed until the full image is reconstructed. Then IAUC by approximating the area under the confidence curve:

$$\text{IAUC} = \int_0^1 p_t^{(\text{ins})} d\rho_t \quad (4)$$

The trapezoidal rule is used to approximate these integrals. High IAUC values represent a faithful explanation, characterized by a rapid increase in confidence as important pixels are revealed.

### 3.4.2. Deletion Area Under Curve

Deletion analysis measures the decrease in model confidence as pixels are gradually removed in order from most to least important [67]. The image pixels are updated as follows [66]:

$$X_t^{(\text{del})}(i, j) = \begin{cases} 0, & \text{if pixel rank} \leq k_t \\ X(i, j), & \text{otherwise} \end{cases} \quad (5)$$

At each step, a portion of the pixels is removed, and the probability for class  $c$  is computed as:

$$p_t^{(\text{del})} = P_c(X_t^{(\text{del})}) \quad (6)$$

Similarly, DAUC is calculated as the area under the corresponding confidence curve:

$$\text{DAUC} = \int_0^1 p_t^{(\text{del})} d\rho_t \quad (7)$$

Low DAUC values represent a faithful explanation, characterized by a rapid decrease in confidence as important pixels are removed.

### 3.4.3. Average Drop in Deletion

ADD measures the drop in classification accuracy when the important features are removed [64]. An explanation is faithful if it identifies the important regions that, when removed, cause the model to make incorrect predictions. Therefore, ADD is formulated as [64]:

$$\text{ADD} = \frac{1}{N_{\text{correct}}} \sum_{i=1}^{N_{\text{correct}}} \mathbf{1}_{(R_i^{\text{pred}} \neq T_i)} \quad (8)$$

$\mathbf{1}(\cdot)$  is the indicator function,  $R_i^{\text{pred}}$  is the predicted class after removing the most important pixels, and  $T_i$  is the true class label for the same sample  $i$ . The summation is averaged over  $N_{\text{correct}}$ , which denotes the number of samples originally classified correctly. A high drop in performance, which results in high ADD values, indicates that the model has successfully identified important regions that are critical for correct predictions.

Masked pixels were replaced with the mean intensity  $\mu$  of the input image to preserve image smoothness, ensuring that any performance drop is linked to the removal of important pixels rather than artifacts caused by abrupt changes in pixel values.

#### 3.4.4. Importance Focus

IF measures the concentration of feature importance within the identified important feature [7]. IF is defined as:

$$IF = \frac{\sum_{p \in P} |A(p)|}{\sum_{p \in \Omega} |A(p)|} \quad (9)$$

$A(p)$  represents the SHAP importance value of pixel  $p$ .  $\Omega$  is the set that contains all the pixels in the input image  $X$ , and  $P$  is the set of important pixels identified by SHAP. A large IF value means that the attribution scores are concentrated on a smaller subset of pixels, yielding a more focused and interpretable explanation.

#### 3.4.5. Spatial Focus

SF measures the spatial clustering of the selected top  $k\%$  pixels and is defined as:

$$SF = 1 - \frac{\overline{d(P)}}{D_{\max}} \quad (10)$$

where  $D_{\max}$  is the maximum distance in  $X$  and  $\overline{d(P)}$  is the average pairwise distance among the important set of pixels  $P$ . Higher SF values indicate that important pixels are spatially clustered [7], making the explanation more focused and interpretable.

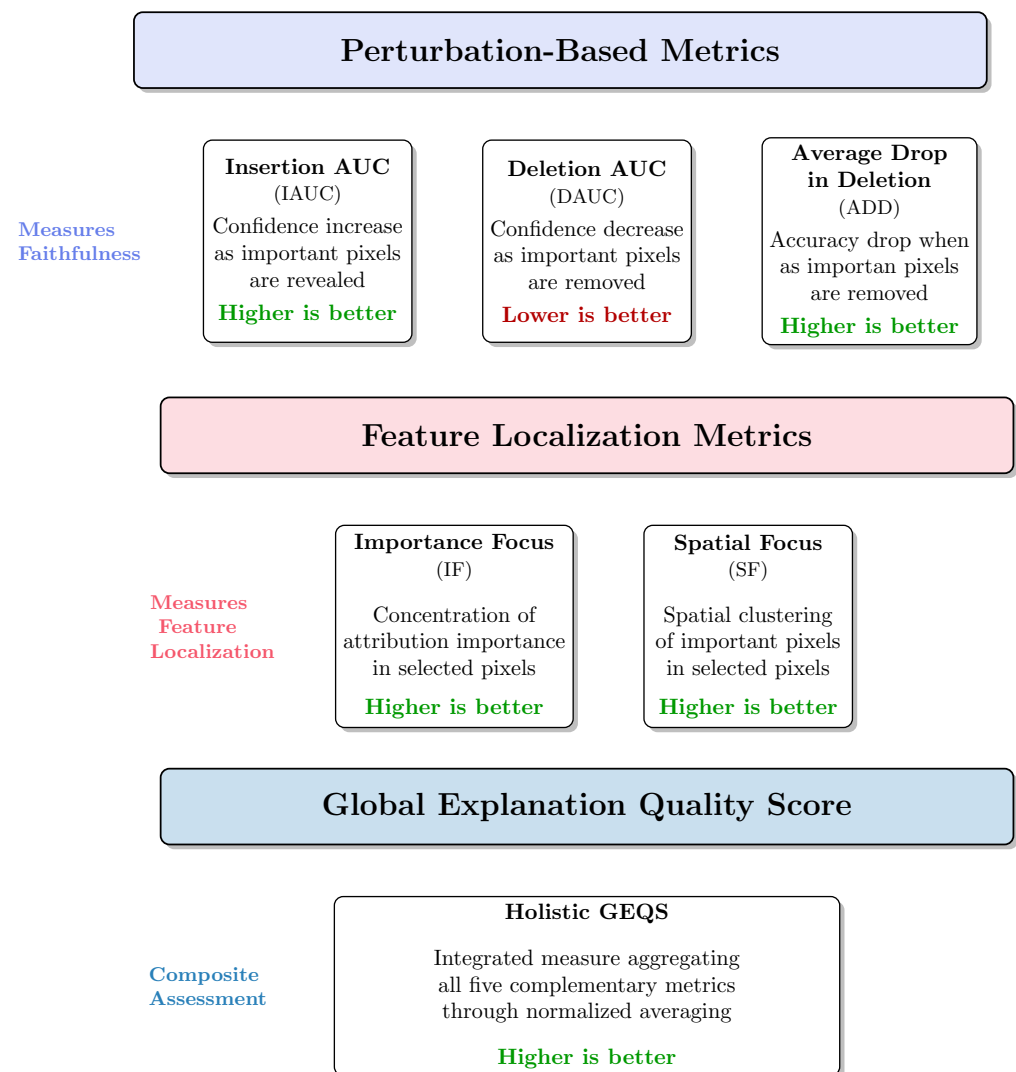
#### 3.4.6. Global Explanation Quality Score

To obtain a comprehensive assessment of explanation quality, the Global Explanation Quality Score (GEQS) was computed as the mean of all five metrics:

$$GEQS = \text{Mean}(\text{IAUC}, 1 - \text{DAUC}, \text{ADD}, \text{IF}, \text{SF}) \quad (11)$$

All metrics are normalized to have the range of  $[0, 1]$ . Each metric captures a distinct yet complementary aspect of explanation quality; therefore, the arithmetic mean was chosen to ensure that all metrics contribute equally to the final assessment. For IAUC, ADD, IF, and SF, higher values indicate better explanation quality. However, lower DAUC values indicate better explanations. To ensure consistent aggregation,  $(1 - \text{DAUC})$  is used. GEQS provides an integrated measure of explanation fidelity and interpretability. Higher values of GEQS indicate better explanation quality. Figure 2 presents a summary of the quantitative XAI evaluation framework employed in this study.

GEQS evaluates faithfulness using IAUC and DAUC, as proposed in [66], and further incorporates ADD, which measures the drop in classification accuracy after feature deletion. However, unlike [66], GEQS goes beyond the faithfulness dimension by incorporating feature localization metrics (IF and SF) to determine whether explanations focus on spatially coherent regions. This multidimensional design ensures that explanations are both faithful and interpretable, an important consideration in clinically oriented applications where explanations should focus on emotion-relevant facial features rather than diffuse, scattered pixels. GEQS uses equal weighting for all metrics to provide a balanced evaluation of faithfulness and spatial interpretability. Therefore, GEQS provides a more comprehensive assessment of explanation quality by balancing complementary aspects such as faithfulness and spatial interpretability.



**Figure 2.** The framework integrates perturbation-based faithfulness metrics (IAUC, DAUC, ADD) and feature localization metrics (IF, SF) into a holistic Global Explanation Quality Score for systematic comparison.

### 3.5. Qualitative Explainability Evaluation Through Human Assessment

Complementing the quantitative explainability evaluation, a qualitative human evaluation was carried out to assess the reliability of SHAP explanations and the relevance of highlighted regions in aligning with meaningful facial cues as perceived by humans. The study involved 35 independent adult participants, including both males and females, to evaluate the explainability of four CNN architectures. The survey comprised 12 tasks per emotion, totaling 36 evaluation tasks, in which image samples were selected from the test set with an equal gender distribution.

For each task, the same image sample, along with the corresponding SHAP explanations, was presented to all participants. Participants rated the explanation quality for each model using a five-point Likert scale (1 = poor, 5 = excellent) based on whether the highlighted regions aligned with facial features relevant to the expressed emotion. Participants were then asked to choose the most trustworthy model for each sample based on their confidence in the model's prediction.

Survey responses were aggregated to calculate evaluation metrics for each model across emotions, with an overall mean Likert score representing perceived quality. Rating consistency was assessed for each model using the coefficient of variation (CV), which quan-

tifies rating stability by evaluating whether samples received similar scores or exhibited substantial variability. The confidence selection frequency tracked how often each model was chosen as most trustworthy. Additionally, a reliability analysis defined explanations as reliable if they received a rating of four or higher. The proportion of reliable explanations was calculated for each emotion and model, providing insight into how frequently each model generated explanations perceived as clearly informative and trustworthy.

## 4. Experimental Results

### 4.1. CNN Architecture Selection for Explainability Evaluation

Table 3 reports the performance of the various CNN architectures across FER2013, RAF-DB, and ExpW. Among the three-layer architectures, (64-128-256) and (128-256-512) were the top performers and were selected for explainability evaluation due to their strong cross-dataset generalization. Similarly, the (64-128-256-512) and (128-128-256-512) configurations were identified as the top-performing four-layer models and were included to enable comparison of explainability across architectures with different depths.

**Table 3.** Performance comparison of 3-layer and 4-layer CNN architectures across FER2013, RAF-DB, and ExpW datasets.

| Architecture | Configuration   | FER2013 (Mean [95% CI])     |                             | RAF-DB (Mean [95% CI])      |                             | ExpW (Mean [95% CI])        |                             |
|--------------|-----------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|
|              |                 | Acc                         | F1                          | Acc                         | F1                          | Acc                         | F1                          |
| 3-Layer      | 16-32-64        | 0.682 [0.675, 0.689]        | 0.679 [0.670, 0.688]        | 0.789 [0.783, 0.795]        | 0.787 [0.782, 0.792]        | 0.664 [0.660, 0.668]        | 0.657 [0.654, 0.660]        |
|              | 32-64-128       | 0.710 [0.704, 0.716]        | 0.709 [0.703, 0.715]        | 0.802 [0.797, 0.807]        | 0.800 [0.795, 0.805]        | 0.684 [0.680, 0.688]        | 0.675 [0.672, 0.678]        |
|              | 64-128-256      | 0.714 [0.710, 0.718]        | 0.714 [0.710, 0.718]        | 0.817 [0.812, 0.822]        | 0.815 [0.809, 0.821]        | <b>0.691 [0.688, 0.694]</b> | <b>0.682 [0.679, 0.685]</b> |
|              | 128-128-256     | 0.714 [0.710, 0.718]        | 0.714 [0.710, 0.718]        | 0.816 [0.813, 0.819]        | 0.815 [0.812, 0.818]        | 0.688 [0.685, 0.691]        | 0.679 [0.677, 0.681]        |
|              | 128-256-512     | <b>0.717 [0.713, 0.721]</b> | <b>0.716 [0.712, 0.720]</b> | <b>0.822 [0.819, 0.825]</b> | <b>0.820 [0.817, 0.823]</b> | 0.689 [0.686, 0.692]        | 0.681 [0.678, 0.684]        |
| 4-Layer      | 16-32-64-128    | 0.670 [0.667, 0.673]        | 0.669 [0.666, 0.672]        | 0.762 [0.758, 0.766]        | 0.762 [0.758, 0.766]        | 0.647 [0.642, 0.652]        | 0.646 [0.641, 0.651]        |
|              | 32-64-128-256   | 0.696 [0.687, 0.705]        | 0.695 [0.687, 0.703]        | 0.789 [0.786, 0.792]        | 0.787 [0.784, 0.790]        | 0.670 [0.667, 0.673]        | 0.666 [0.664, 0.668]        |
|              | 64-128-256-512  | <b>0.717 [0.709, 0.725]</b> | <b>0.716 [0.709, 0.723]</b> | <b>0.803 [0.796, 0.810]</b> | <b>0.801 [0.793, 0.809]</b> | <b>0.686 [0.683, 0.689]</b> | <b>0.678 [0.675, 0.681]</b> |
|              | 128-128-256-512 | 0.710 [0.706, 0.714]        | 0.709 [0.705, 0.713]        | 0.802 [0.795, 0.809]        | 0.800 [0.793, 0.807]        | 0.681 [0.676, 0.686]        | 0.675 [0.670, 0.680]        |

Note: Acc = Accuracy; F1 = F1-Score. Values are reported as mean [95% CI] across five independent runs. Best performance per architecture type is bolded.

### 4.2. Quantitative Explainability Analysis of CNN Architectures

Tables 4 and 5 present sample SHAP explanations for each emotion across the selected CNN architectures. These visualizations illustrate how different models highlight facial regions for emotion classification, providing a visual foundation for understanding the quantitative metrics evaluated below. As shown in Table 10, the (128-256-512) architecture achieves the highest GEQS scores for anger (0.635) and happy (0.635). For sad, the (64-128-256-512) architecture achieves the highest GEQS of 0.607, while for neutral, the (128-128-256-512) model achieves the highest GEQS of 0.555.

### 4.3. Human-Centered Explainability Analysis of Selected CNN Architectures

To assess the alignment between quantitative explainability evaluation and human perception, 35 participants evaluated explanations generated by the selected CNN architectures across happy, anger and sad. Participants were shown facial images along with the corresponding explanations produced by each model.

Models were anonymized throughout the evaluation process to avoid bias resulting from architectural variations or model complexity. For each sample, participants rated the explanation quality of all four models and then selected the model whose explanation inspired the greatest confidence.

**Table 4.** SHAP visualizations for different CNN architectures across anger and happy emotion classes, highlighting the most important features for model predictions.

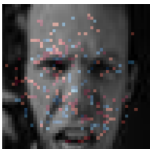
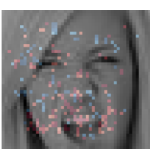
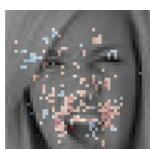
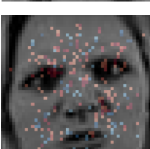
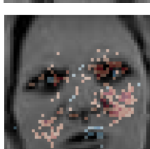
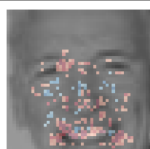
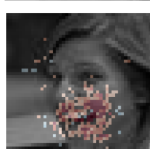
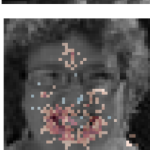
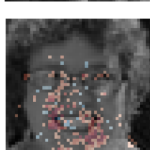
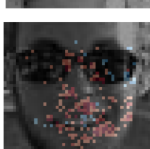
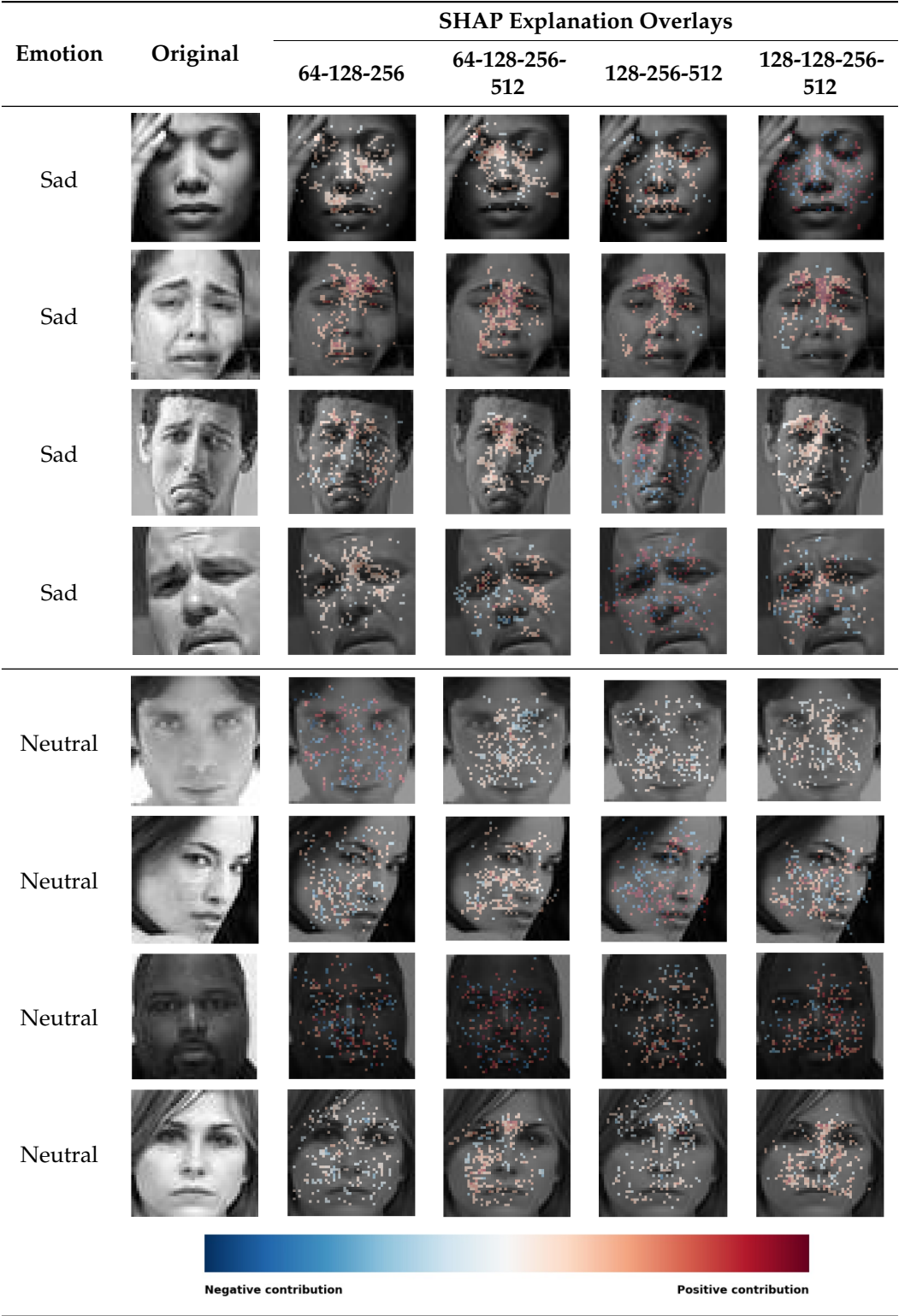
| Emotion | Original                                                                            | SHAP Explanation Overlays                                                            |                                                                                      |                                                                                       |                                                                                       |
|---------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
|         |                                                                                     | 64-128-256                                                                           | 64-128-256-512                                                                       | 128-256-512                                                                           | 128-128-256-512                                                                       |
| Anger   |    |     |    |    |    |
|         |    |     |    |    |    |
|         |    |     |    |    |    |
|         |   |    |   |   |   |
| Happy   |  |   |  |  |  |
|         |  |   |  |  |  |
|         |  |   |  |  |  |
|         |  |   |  |  |  |
|         |                                                                                     |  |                                                                                      |                                                                                       |                                                                                       |
|         |                                                                                     | Negative contribution <span style="float: right;">Positive contribution</span>       |                                                                                      |                                                                                       |                                                                                       |

Table 6 summarizes the results of the rating-based explainability evaluation across the emotions. The (128-256-512) model achieved the highest mean ratings for anger, with 3.21, and happy, with 3.47, while for sad, the (128-128-256-512) configuration achieved a slightly higher mean of 3.28.

**Table 5.** SHAP visualizations for different CNN architectures across sad and neutral emotion classes, highlighting the most important features for model predictions.



To provide an overall evaluation, the ratings were averaged across all emotions as presented in Table 7 along with the corresponding CV values. Table 7 shows that the (128-256-512) architecture had the highest mean rating of 3.3 while maintaining the lowest CV value among all models.

**Table 6.** Rating-based explainability analysis per emotion with 95% Confidence Intervals ( $N = 210$  per configuration).  $\uparrow$ : higher is better;  $\downarrow$ : lower is better.

| Emotion   | CNN Configuration | Mean $\uparrow$ | 95% CI       | CV $\downarrow$ |
|-----------|-------------------|-----------------|--------------|-----------------|
| Anger     | 64-128-256        | 3.03            | [2.85, 3.20] | 0.42            |
|           | 64-128-256-512    | 3.10            | [2.95, 3.26] | 0.38            |
|           | 128-256-512       | 3.21            | [3.06, 3.37] | <b>0.35</b>     |
|           | 128-128-256-512   | 3.00            | [2.83, 3.17] | 0.40            |
| Sadness   | 64-128-256        | 2.99            | [2.83, 3.15] | 0.40            |
|           | 64-128-256-512    | 3.20            | [3.05, 3.35] | <b>0.34</b>     |
|           | 128-256-512       | 3.23            | [3.08, 3.38] | <b>0.34</b>     |
|           | 128-128-256-512   | <b>3.28</b>     | [3.11, 3.44] | 0.37            |
| Happiness | 64-128-256        | 3.12            | [2.97, 3.28] | 0.36            |
|           | 64-128-256-512    | 3.38            | [3.23, 3.53] | 0.33            |
|           | 128-256-512       | <b>3.47</b>     | [3.31, 3.62] | <b>0.33</b>     |
|           | 128-128-256-512   | 3.37            | [3.21, 3.53] | 0.35            |

**Table 7.** Overall rating-based explainability across all emotions with 95% Confidence Intervals ( $N = 630$  per configuration).  $\uparrow$ : higher is better;  $\downarrow$ : lower is better.

| CNN Configuration | Mean $\uparrow$ | 95% CI       | CV $\downarrow$ |
|-------------------|-----------------|--------------|-----------------|
| 64-128-256        | 3.05            | [2.95, 3.14] | 0.40            |
| 64-128-256-512    | 3.23            | [3.14, 3.32] | 0.35            |
| 128-256-512       | 3.30            | [3.22, 3.39] | 0.34            |
| 128-128-256-512   | 3.21            | [3.12, 3.31] | 0.38            |

Table 8 reports the percentage of explanations rated as reliable, defined as receiving a score of four or higher. Consistent with the rating-based results, the (128-256-512) architecture achieves the highest overall reliability.

**Table 8.** Overall reliability across all emotions with 95% Confidence Intervals ( $N = 630$  per configuration).

| CNN Configuration | Reliable (%) | 95% CI         |
|-------------------|--------------|----------------|
| 64-128-256        | 39.84        | [36.09, 43.71] |
| 64-128-256-512    | 40.79        | [37.02, 44.67] |
| 128-256-512       | 47.14        | [43.27, 51.04] |
| 128-128-256-512   | 45.24        | [41.39, 49.14] |

The results of the user confidence evaluation are shown in Table 9. The (64-128-256-512) architecture received the highest number of confidence selections (179, 28.41%), followed closely by the (128-256-512) model (176, 27.94%).

**Table 9.** Confidence-based model selection with 95% Confidence Intervals ( $N = 630$  total votes).

| CNN Configuration | Votes | Percentage (%) | 95% CI         |
|-------------------|-------|----------------|----------------|
| 64-128-256        | 137   | 21.75          | [18.70, 25.13] |
| 64-128-256-512    | 179   | 28.41          | [25.03, 32.06] |
| 128-256-512       | 176   | 27.94          | [24.57, 31.57] |
| 128-128-256-512   | 138   | 21.90          | [18.85, 25.30] |

## 5. Discussion

### 5.1. CNN Architecture Selection for Explainability Evaluation

Table 3 reveal that three-layer architectures consistently outperformed all four-layer configurations across all datasets. The results indicate that the (128-256-512) configuration achieved the highest performance on both FER2013 and RAF-DB, with accuracy and F1-scores of 0.717 and 0.716 on FER2013 and 0.822 and 0.820 on RAF-DB, respectively. On ExpW, the (64-128-256) configuration achieved the best performance. Therefore, the (64-128-256) and (128-256-512) three-layer models were chosen for explainability evaluation. Among the four-layer architectures, the results in Table 3 demonstrate that the (64-128-256-512) configuration achieved the strongest performance across all datasets, while the (128-128-256-512) model is the second best. Accordingly, these two configurations were also selected for explainability analysis. The aim is to evaluate the impact of predictive performance and architectural depth on explanation quality.

### 5.2. Analysis of Quantitative Explainability

As per the results in Table 10, the (128-256-512) architecture exhibits superior explainability for both happy and angry emotions, achieving the highest GEQS scores of 0.635. Although the (64-128-256-512) architecture achieved the highest GEQS score of 0.607 for sad, the improvement is marginal compared to the (128-256-512) configuration.

Explainability of the selected architectures was evaluated using the proposed quantitative metrics. IAUC, DAUC, ADD, IF, SF, and GEQS were computed for each model on the same consensus samples, with results reported as averages across N samples per emotion along with 95% confidence intervals (Table 10).

All architectures showed lower scores across metrics for neutral. As visually confirmed in Table 5, this trend is consistent with previous findings that neutral expressions lack strongly discriminative features [18], leading to more diffuse explanations [63], which is reflected in the lower quantitative metrics. Accordingly, as per the quantitative explainability analysis, the (128-256-512) architecture provides the most robust and balanced overall explainability quality.

### 5.3. Alignment Between Quantitative Metrics and Human Evaluation

The human evaluation results demonstrate strong alignment with the quantitative analysis. As shown in Table 6, the (128-256-512) model achieved the highest mean ratings for anger with 3.21 and happy with 3.47, with the lowest CV values across all emotions.

Although the (128-128-256-512) configuration achieved a slightly higher mean for sad with 3.28 compared to 3.23 for the (128-256-512), it does not achieve the lowest variability, and the overlapping confidence intervals suggest the difference is not statistically significant. The findings in Table 7 also show that the (128-256-512) model exhibits consistent and stable explanation quality across emotions, achieving the lowest CV of 0.34 and the highest overall mean rating of 3.30. This conclusion is supported by the reliability analysis shown in Table 8, demonstrating that the (128-256-512) architecture has the highest percentage of reliable explanations (47.14%).

**Table 10.** XAI evaluation metrics across emotions for different CNN architectures. Values are reported as mean with 95% confidence intervals computed over N samples per emotion: anger (N = 775), happiness (N = 800), sadness (N = 798), and neutral (N = 800). IAUC: Insertion Area Under Curve; DAUC: Deletion Area Under Curve; ADD: Average Drop in Deletion; IF: Importance Focus; SF: Spatial Focus; GEQS: Global Explanation Quality Score. Bold indicates the maximum value.

| Emotion | Model           | IAUC<br>Avg (95% CI)       | DAUC<br>Avg (95% CI)       | ADD<br>Avg (95% CI)        | IF<br>Avg (95% CI)         | SF<br>Avg (95% CI)         | GEQS<br>Avg (95% CI)       |
|---------|-----------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|
| Anger   | 64-128-256      | 0.343 (0.330–0.355)        | 0.216 (0.206–0.225)        | 0.481 (0.396–0.567)        | 0.441 (0.438–0.443)        | 0.745 (0.744–0.747)        | 0.559 (0.551–0.566)        |
|         | 64-128-256-512  | 0.439 (0.427–0.452)        | <b>0.166 (0.160–0.173)</b> | 0.597 (0.508–0.687)        | 0.460 (0.457–0.462)        | <b>0.786 (0.785–0.787)</b> | 0.623 (0.616–0.630)        |
|         | 128-256-512     | 0.459 (0.446–0.472)        | 0.170 (0.162–0.178)        | <b>0.668 (0.635–0.702)</b> | 0.453 (0.450–0.455)        | 0.763 (0.762–0.765)        | <b>0.635 (0.628–0.642)</b> |
|         | 128-128-256-512 | <b>0.561 (0.548–0.574)</b> | 0.256 (0.247–0.265)        | 0.564 (0.529–0.599)        | <b>0.473 (0.471–0.475)</b> | <b>0.786 (0.785–0.787)</b> | 0.625 (0.618–0.633)        |
| Happy   | 64-128-256      | 0.388 (0.376–0.400)        | <b>0.063 (0.060–0.066)</b> | 0.548 (0.513–0.582)        | 0.459 (0.456–0.462)        | 0.775 (0.773–0.777)        | 0.621 (0.613–0.629)        |
|         | 64-128-256-512  | 0.391 (0.380–0.402)        | 0.130 (0.126–0.133)        | <b>0.444 (0.409–0.478)</b> | 0.452 (0.450–0.454)        | <b>0.791 (0.790–0.792)</b> | 0.590 (0.582–0.597)        |
|         | 128-256-512     | <b>0.408 (0.396–0.420)</b> | 0.073 (0.070–0.076)        | 0.603 (0.569–0.636)        | 0.458 (0.455–0.461)        | 0.777 (0.776–0.779)        | <b>0.635 (0.627–0.642)</b> |
|         | 128-128-256-512 | 0.402 (0.391–0.413)        | 0.141 (0.138–0.145)        | 0.466 (0.432–0.501)        | <b>0.460 (0.458–0.463)</b> | 0.788 (0.786–0.789)        | 0.595 (0.588–0.602)        |
| Sad     | 64-128-256      | <b>0.392 (0.383–0.402)</b> | 0.176 (0.169–0.183)        | <b>0.465 (0.430–0.500)</b> | 0.428 (0.425–0.430)        | 0.748 (0.746–0.749)        | 0.571 (0.564–0.579)        |
|         | 64-128-256-512  | 0.344 (0.335–0.353)        | <b>0.095 (0.091–0.098)</b> | 0.556 (0.521–0.590)        | 0.445 (0.443–0.447)        | <b>0.785 (0.783–0.786)</b> | <b>0.607 (0.600–0.614)</b> |
|         | 128-256-512     | 0.333 (0.324–0.342)        | 0.139 (0.133–0.146)        | 0.626 (0.592–0.659)        | 0.427 (0.425–0.429)        | 0.756 (0.755–0.757)        | 0.601 (0.593–0.608)        |
|         | 128-128-256-512 | 0.348 (0.340–0.357)        | 0.205 (0.199–0.210)        | 0.601 (0.567–0.635)        | <b>0.455 (0.452–0.457)</b> | 0.782 (0.781–0.784)        | 0.596 (0.589–0.603)        |
| Neutral | 64-128-256      | 0.724 (0.714–0.733)        | 0.657 (0.647–0.667)        | <b>0.325 (0.292–0.358)</b> | 0.458 (0.455–0.461)        | 0.736 (0.735–0.737)        | 0.517 (0.510–0.524)        |
|         | 64-128-256-512  | 0.684 (0.675–0.692)        | 0.689 (0.683–0.696)        | 0.321 (0.289–0.354)        | 0.447 (0.445–0.449)        | 0.771 (0.770–0.772)        | 0.507 (0.500–0.513)        |
|         | 128-256-512     | <b>0.788 (0.779–0.796)</b> | <b>0.707 (0.698–0.716)</b> | 0.274 (0.243–0.305)        | <b>0.460 (0.457–0.462)</b> | 0.738 (0.737–0.739)        | 0.511 (0.504–0.517)        |
|         | 128-128-256-512 | 0.631 (0.621–0.640)        | 0.512 (0.502–0.521)        | 0.434 (0.399–0.468)        | 0.453 (0.451–0.456)        | <b>0.769 (0.768–0.770)</b> | <b>0.555 (0.548–0.562)</b> |

The (64-128-256-512) architecture received slightly more confidence selections (179, 28.41%) than the (128-256-512) model (176, 27.94%), according to the confidence-based evaluation in Table 9. Participants chose the (64-128-256-512) model more often in confidence-based tests, despite the fact that the (128-256-512) model produced better explanation quality and reliability as per the results. This implies that user confidence and explanation quality might not be the same cognitive construct. While the rating task required participants to explicitly evaluate how well explanations highlighted emotion-relevant facial features, the confidence selection task reflected more intuitive trust than analytical comparison. Similar differences between explanation quality and user trust have been documented in earlier research, where perceptual simplicity and cognitive ease had a greater impact on confidence judgments than quantitative explanation metrics [68–71]. However, the small difference of only three votes indicates that this does not reflect a meaningful difference in explanation quality.

The validity of the proposed quantitative evaluation framework is confirmed by the consistent alignment between GEQS scores and human judgments. This indicates that the (128-256-512) architecture is the most reliable for explainable FER. In addition to achieving the highest classification accuracy, this architecture also attained relatively high GEQS scores, the highest mean human rating (3.30/5.0), the lowest rating variability (CV = 0.34), and the greatest proportion of reliable explanations (47.14%).

#### 5.4. Per-Class Performance Analysis

To further analyze the predictive performance of the selected (128-256-512) architecture, the normalized confusion matrices are shown in Figure 3.

Table 1 shows a class imbalance, where anger and sad samples are minorities compared to happy and neutral. Despite this, the confusion matrix for FER2013 shows strong recognition across classes, achieving recall values of 0.83, 0.77, 0.72, and 0.56 for happy, neutral, anger, and sad, respectively. A similar performance is observed on RAF-DB, achieving recall values of 0.91, 0.83, 0.71, and 0.63 across the same classes. While misclassifications occur between sad and neutral samples, the mutual confusion suggests that this likely arises from the similarities in facial expressions between the two emotions [72,73].

On ExpW, sad and anger achieve recall values of 0.29 and 0.45, respectively, with many samples misclassified as neutral. This behavior may be influenced by ExpW's severe class imbalance, where neutral and happy dominate, resulting in a strong neutral recall value of 0.79. Such imbalance is a common challenge in large-scale FER datasets, where dominant classes often overshadow minority emotions [74]. However, this pattern also likely reflects the increased variability and subtle emotional expressions characteristic of natural datasets.

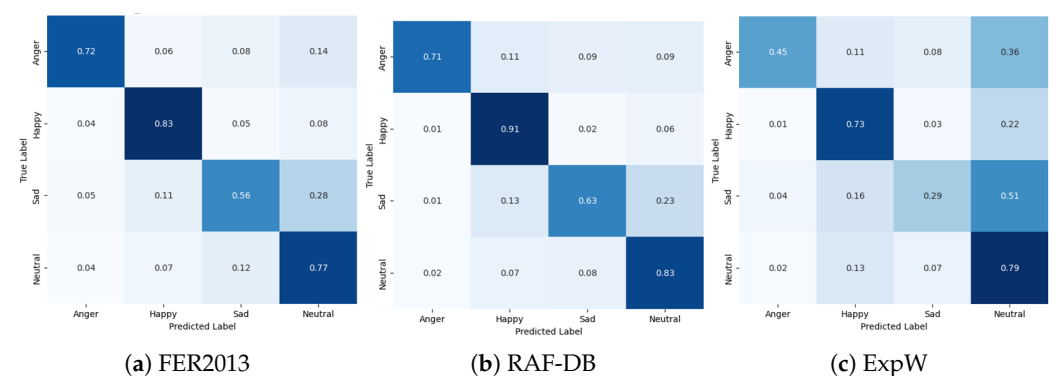


Figure 3. Normalized confusion matrices showing per-class performance across test datasets.

### 5.5. GEQS vs. Existing XAI Evaluation Frameworks

The XAI research landscape still lacks standardized, reliable evaluation protocols, limiting consistent comparisons across explanation methods [75,76]. Many studies rely on qualitative visual inspection or subjective assessment [7]. Nauta et al. [7] define multiple explanation-quality properties in their Co-12 framework but treat them independently, lacking a unified score for comparison or decision-making. In a systematic review of over 300 studies, they further reported that one-third relied solely on visual inspection and one-fifth on user studies. Recent analyses further highlight the inconsistency in XAI metrics and the need for standardized and structured evaluation frameworks [77].

Existing XAI evaluation frameworks further differ in the specific evaluation domain they emphasize. Perturbation-based approaches, such as the Randomized Input Sampling for Explanation (RISE) insertion–deletion framework [66], evaluate faithfulness using metrics such as IAUC and DAUC. While effective for evaluating prediction sensitivity, they do not explicitly capture spatial coherence, which is essential in visually grounded tasks such as facial attribution. Moreover, ranking-based perturbation metrics may overlook information about attribution magnitude when used in isolation [67]. Similarly, Mean Opinion Score (MOS) [78] focuses on faithfulness by evaluating explanations with perturbation-based metrics and analyzing their correlation with human judgment, but it does not aggregate metrics and primarily serves as a human-validation study rather than a composite framework. In contrast, Ref. [79] focus on the localization domain by evaluating spatial alignment between explanation maps and ground-truth regions using Intersection over Union (IoU) and Dice Similarity Coefficient; however, they remain focused on a single domain, limiting their ability to support holistic explanation evaluation. Quantus was proposed as a framework for quantitative evaluation of XAI methods [80] and organizes diverse evaluation metrics across multiple categories without aggregation, limiting direct comparison.

GEQS is designed to bridge these domain-specific gaps through a unified multi-domain evaluation framework. Specifically, it integrates perturbation-based fidelity metrics (IAUC, DAUC, ADD) with spatial localization metrics (IF and SF) within a composite score. Unlike frameworks that evaluate only one aspect of explanation quality, GEQS captures both faithfulness and spatial coherence, enabling explanations to be both decision-relevant and spatially meaningful for facial emotion analysis. This unified formulation supports direct and standardized comparison of explanation quality assessment.

## 6. Comparison with Cross-Dataset FER Systems

To evaluate the cross-dataset generalization of the proposed framework, we compare it against a recent cross-dataset FER study [81] using the same datasets, RAF-DB, ExpW, and FER2013. Table 11 demonstrates performance comparisons across large-scale (ResNet-50) and lightweight (MobileNet-V2) architectures, including domain adaptation methods such as Stepwise Adaptive Feature Norm (SAFN) [82], Conditional Adversarial Domain Adaptation (CADA) [83], and Enhanced Channel Attention Network (ECAN) [84], as well as attention-based approaches including Attention-Guided Residual Aggregation (AGRA) [85], Cross-Global Local Representation Learning (CGLRL) [86], and Label Dynamic Weight Matching (LDWM) [81]. The table shows the best reported accuracies across cross-dataset experimental setups as reported in [81] in addition to the performance of our proposed framework.

The baseline methods reported in [81] employ cross-dataset centralized training under a single-source domain adaptation protocol, where models are trained on one labeled dataset and leverage unlabeled samples from a target dataset prior to evaluation. RAF-DB and ExpW results correspond to models trained on FER2013 as the source domain and subsequently adapted to the respective target datasets, while the FER2013 results

correspond to models trained on RAF-DB as the source domain and adapted to FER2013. Mean accuracy is reported to provide an overall comparison across datasets. These baseline methods classify seven emotion classes. While the proposed framework focuses on four clinically relevant emotions and employs a lightweight federated architecture that facilitates generalized representation learning across heterogeneous datasets without the need for dataset-specific adaptation. The proposed federated approach trains across all datasets simultaneously, allowing collaborative knowledge transfer without centralized data sharing.

The proposed federated model achieves accuracies of 82.2% on RAF-DB, 71.7% on FER2013, and 68.9% on ExpW, yielding a mean accuracy of 74.3%. Compared with LDWM, which reports the best baseline performance, improvements of 6.5% on RAF-DB and 10.6% on FER2013 are observed, while LDWM achieves higher accuracy on ExpW. LDWM relies on centralized domain adaptation using pseudo-label generation and dynamic class weighting for target-specific adaptation. In contrast, the proposed model achieves cross-dataset generalization through collaborative federated training without domain adaptation.

**Table 11.** Cross-dataset FER accuracy performance comparison. Bold indicates the maximum value.

| Method                                          | RAF-DB      | ExpW        | FER2013     | Mean        |
|-------------------------------------------------|-------------|-------------|-------------|-------------|
| <i>Large-Scale Architectures (ResNet-50)</i>    |             |             |             |             |
| SAFN [82]                                       | 62.8        | 64.9        | 55.6        | 61.1        |
| CADA [83]                                       | 66.0        | 63.2        | 57.6        | 62.3        |
| ECAN [84]                                       | 53.4        | 47.4        | 56.5        | 52.4        |
| AGRA [85]                                       | 67.6        | 68.5        | 59.0        | 65.0        |
| CGLRL [86]                                      | 71.8        | 70.0        | 59.3        | 67.0        |
| LDWM [81]                                       | 75.7        | <b>73.3</b> | 61.1        | 70.0        |
| <i>Lightweight Architectures (MobileNet-V2)</i> |             |             |             |             |
| SAFN [82]                                       | 38.7        | 61.4        | 49.9        | 50.0        |
| CADA [83]                                       | 53.2        | 59.4        | 49.3        | 54.0        |
| ECAN [84]                                       | 42.3        | 45.1        | 45.8        | 44.4        |
| AGRA [85]                                       | 52.3        | 64.0        | 45.8        | 54.0        |
| CGLRL [86]                                      | 62.0        | 64.9        | 52.5        | 59.8        |
| LDWM [81]                                       | 64.8        | 64.9        | 53.0        | 60.9        |
| <b>Proposed Approach</b>                        | <b>82.2</b> | 68.9        | <b>71.7</b> | <b>74.3</b> |

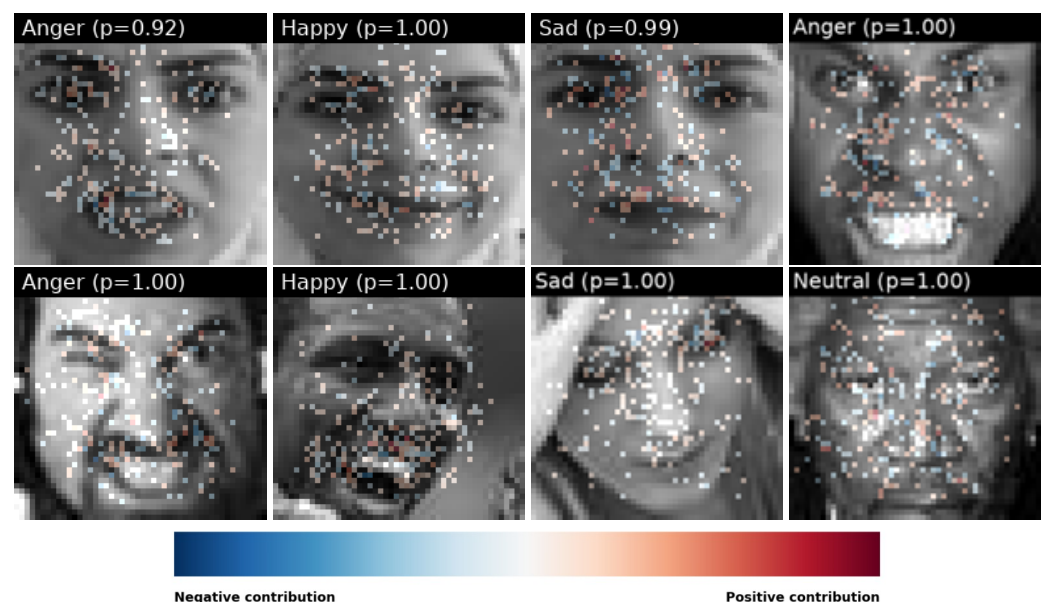
To ensure a fair comparison, LDWM was implemented under the same experimental setup as the proposed framework, using identical training and test splits, emotion classes, training epochs. LDWM utilizes ResNet-50 pretrained on ImageNet as the feature extractor. The weighting formulation proposed in [81] is applied to dynamically adjust class weights based on the predicted class distribution. Experiments were repeated across five random seeds (7, 42, 123, 999, and 2024). Performance was evaluated on FER2013, RAF-DB, and ExpW test sets. Statistical significance between the proposed method and LDWM was assessed using paired *t*-tests across repeated runs. Table 12 reports the paired test results, showing that the proposed framework significantly outperforms LDWM across all datasets ( $p < 0.001$ ). These results demonstrate that collaborative federated learning combined with a lightweight architecture provides an effective alternative to centralized domain adaptation pipelines for cross-dataset facial emotion recognition.

**Table 12.** Accuracy comparison between the proposed model and LDWM across five random seeds. Results are reported as mean [95% CI]. Statistical significance is computed using a paired *t*-test. Bold indicates the maximum value.

| Dataset | Proposed                    | LDWM [81]            | <i>p</i> -Value |
|---------|-----------------------------|----------------------|-----------------|
| FER2013 | <b>0.717 [0.713, 0.722]</b> | 0.680 [0.676, 0.685] | 0.00017         |
| RAF-DB  | <b>0.822 [0.819, 0.825]</b> | 0.750 [0.742, 0.757] | 0.00008         |
| ExpW    | <b>0.689 [0.687, 0.692]</b> | 0.637 [0.630, 0.643] | 0.00016         |

## 7. Deployment Viability of the Proposed FER System with SHAP on Raspberry Pi

Deployment on a Raspberry Pi 4 Model B (quad-core ARM Cortex-A72, 8 GB RAM, 1.5 GHz) with integrated SHAP explainability demonstrated on-device emotion inference, confirming the proposed FER's viability. The FER2013 test set was used to measure inference latency, resident memory (RSS), and CPU utilization. The model achieved an average inference time of 60.0 ms, 17 frames per second (FPS), while consuming 25.1% CPU and 166.9 MB of memory. SHAP was integrated into the deployed model on the Raspberry Pi for on-device explainability. Despite achieving real-time inference [87], the perturbation-based methodology of SHAP computation results in additional processing overhead. For large-scale continuous explainability generation, more powerful edge hardware may be required. Overall, this implementation demonstrates that privacy-preserving and explainable FER systems are feasible on resource-constrained edge devices, with explainability available to support clinical decisions without compromising real-time emotion recognition performance. Figure 4 presents representative test samples with predicted emotion confidence executed on the Raspberry Pi.



**Figure 4.** Facial emotion recognition outputs generated on the Raspberry Pi 4, showing recognized emotions, confidence scores, and SHAP-based pixel-level explanations.

## 8. Conclusions

This work presented a privacy-preserving, trustworthy, and explainable FER framework designed for continuous emotion monitoring in clinically relevant settings. Leveraging federated learning, the proposed system enables collaborative training of lightweight CNN architectures across heterogeneous datasets while ensuring strict client-side data

privacy. Explainability was integrated into the model optimization process using SHAP to evaluate prediction trustworthiness and performance. GEQS is a multidimensional quantitative explainability evaluation framework that assesses explainability in faithfulness and feature localization domains. In addition to quantitative assessment, a qualitative, human-centered user study was conducted to evaluate the perceived reliability and trustworthiness of the explanation from a human perspective. Findings showed that human assessments and quantitative GEQS scores were highly aligned, with both recommending the same lightweight CNN architecture. The chosen model exhibited strong generalization, achieving a mean cross-dataset accuracy of 74.3%. Additionally, qualitative analysis showed that highlighting relevant facial regions does not guarantee human trust, emphasizing the need for a deeper understanding of the factors that influence trust in explainable systems. Implementation on a Raspberry Pi 4 with integrated SHAP explainability demonstrated the proposed model's viability in potential clinical settings with limited resources, enabling real-time, privacy-aware, and explainable FER. To better support clinical interpretation and trust, future work will assess how well model explanations align with clinically relevant facial cues, with input from domain experts.

**Author Contributions:** Conceptualization, D.S. and A.H.; Methodology, D.S.; Software, D.S.; Validation, D.S., H.T., A.B. and A.H.; Formal analysis, A.H.; Resources, H.T., A.B. and A.H.; Writing—original draft, D.S.; Writing—review and editing, H.T., A.B. and A.H.; Supervision, H.T., A.B. and A.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Appendix A. Reproducibility Details

To ensure experimental transparency and reproducibility of the proposed work, this appendix provides details of the proposed methodology:

### *Appendix A.1. Datasets and Data Preprocessing*

RAF-DB [58], FER2013 [59], and ExpW [60] are the public natural datasets used for training and testing. Four classes, anger, happy, sad, and neutral, were chosen based on their relevance to the proposed application. The original dataset splits were preserved for FER2013 and RAF-DB, with an 80:20 split between training and testing, while ExpW was split 70:30 following common practice.

The dataset per class distribution is presented in Table 1. All samples were preprocessed as follows, normalized to the range  $[0, 1]$ , resized to  $48 \times 48$  pixels, and converted to grayscale. The dataset contains samples of pre-cropped faces; therefore, no additional face detection was required.

### *Appendix A.2. Model Architectures*

The various evaluated CNN architectures are shown in Table A1. The table presents their filter configurations and the parameter counts for each model.

**Table A1.** Evaluated lightweight CNN configurations and parameter counts. K denotes thousands of parameters, and M denotes millions of parameters.

| Filter Configuration | Parameters |
|----------------------|------------|
| (16-32-64)           | 319 K      |
| (32-64-128)          | 651 K      |
| (64-128-256)         | 1.5 M      |
| (128-128-256)        | 1.53 M     |
| (128-256-512)        | 3.60 M     |
| (16-32-64-128)       | 164 K      |
| (32-64-128-256)      | 487 K      |
| (64-128-256-512)     | 1.72 M     |
| (128-128-256-512)    | 1.8 M      |

### Appendix A.3. Federated Training Setup

To ensure a fair comparison, all CNN architectures were trained using the same framework. The models were trained from scratch and initialized with randomly assigned weights. Five random seeds (7, 42, 123, 999, 2024) were used during training to reduce variability and improve the reliability of the reported results.

**Data Partitioning:** Three clients participated in the federated learning process. Training samples were partitioned into non-overlapping sets, where each client received a set containing all labels. Prior to distribution, training samples were shuffled using a fixed random seed (seed = 42) to promote reproducibility.

**Aggregation Strategy:** Client models were aggregated using the standard Federated Averaging (FedAvg) algorithm [4]. At each communication round, all three clients participated in local training.

**Training Protocol:** The complete training and federated configuration, including the optimizer, learning rate, batch size, local epochs, communication rounds, aggregation strategy, and seeds, is summarized in Table A2.

**Table A2.** Federated training configuration.

| Parameter             | Setting                                    |
|-----------------------|--------------------------------------------|
| Number of clients     | 3                                          |
| Data partitioning     | Non-overlapping samples                    |
| Classes per client    | All 4 classes (Anger, Happy, Sad, Neutral) |
| Local training epochs | 15                                         |
| Communication rounds  | 20                                         |
| Optimizer             | Adam                                       |
| Learning rate         | 0.001                                      |
| Batch size            | 64                                         |
| Loss function         | Categorical cross-entropy                  |
| Aggregation strategy  | FedAvg                                     |
| Data split seed       | 42                                         |
| Training seeds        | 7, 42, 123, 999, 2024                      |

All experiments were conducted using Google Colab Pro with GPU acceleration. Models were implemented using TensorFlow 2.x and Keras API with CUDA-enabled GPU support. The federated learning framework was implemented in Python 3.12 using TensorFlow and Keras.

#### Appendix A.4. Explainability Evaluation

SHAP explanations were generated using the DeepExplainer [8] to obtain pixel-level attribution maps for the predicted emotion class. All images used for explainability were preprocessed identically to the training pipeline (Appendix A.1). A background set of 20 randomly sampled images was used for DeepExplainer to approximate the expected model output.

SHAP analysis: Explainability analysis was performed on the same samples that were correctly classified across all evaluated CNN models. For each explanation, the top 10% of pixels with the highest absolute SHAP values were selected as the important set for visualization purposes, following the threshold selection procedure described in the main methodology. Raw SHAP values were converted to absolute values to capture both positive and negative contributions, then normalized to be within [0, 1].

Evaluation Protocol: Quantitative explanation quality was assessed using perturbation-based metrics: Insertion AUC (IAUC) and Deletion AUC (DAUC) were computed on the full attribution maps. Insertion and deletion curves were computed over 20 progressive steps, with pixels perturbed in order of importance. Average Drop in Deletion (ADD), Importance Focus (IF), and Spatial Focus (SF) were computed using the top 10% most important pixels. The Global Explanation Quality Score (GEQS) was calculated by averaging these five normalized metrics.

Visualization: To ensure consistent color mapping across models, a global color scale was computed per image using the 99th percentile of absolute SHAP values.

#### Appendix A.5. Reproducibility Checklist

Table A3 maps each main result presented in the paper to the corresponding appendix section containing the implementation details and configuration required for reproduction.

**Table A3.** Reproducibility checklist mapping the main results to the corresponding appendix sections.

| Main Result    | Reproduced Using                                                                                           |
|----------------|------------------------------------------------------------------------------------------------------------|
| Table 3        | Model architectures from Appendix A.2, trained using the federated setup described in Appendix A.3         |
| Table 10       | SHAP evaluation metrics protocol detailed in Appendix A.4, applied to models from Appendix A.2             |
| Tables 4 and 5 | SHAP visualization protocol from Appendix A.4 using selected models from Appendix A.2                      |
| Tables 6–9     | Human evaluation of SHAP explanation visualizations generated using the protocol described in Appendix A.4 |
| Figure 3       | Model (128-256-512) from Appendix A.2, trained per Appendix A.3, evaluated on test sets from Appendix A.1  |

## References

1. Fei, Z.; Yang, E.; Li, D.D.U.; Butler, S.; Ijomah, W.; Li, X.; Zhou, H. Deep convolution network based emotion analysis towards mental health care. *Neurocomputing* **2020**, *388*, 212–227. [CrossRef]
2. Kaissis, G.A.; Makowski, M.R.; Rückert, D.; Braren, R.F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2020**, *2*, 305–311. [CrossRef]
3. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
4. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), Volume 54, Ft. Lauderdale, FL, USA, 20–22 April 2017*; PMLR: Brookline, MA, USA, 2017; pp. 1273–1282.

5. Zhu, H.; Xu, J.; Liu, S.; Jin, Y. Federated learning on non-IID data: A survey. *Neurocomputing* **2021**, *465*, 371–390. [[CrossRef](#)]
6. Lu, Z.; Pan, H.; Dai, Y.; Si, X.; Zhang, Y. Federated Learning with Non-IID Data: A Survey. *IEEE Internet Things J.* **2024**, *11*, 19188–19209. [[CrossRef](#)]
7. Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; Van Keulen, M.; Seifert, C. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Comput. Surv.* **2023**, *55*, 1–42. [[CrossRef](#)]
8. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. [[CrossRef](#)]
9. Minaee, S.; Minaei, M.; Abdolrashidi, A. Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors* **2021**, *21*, 3046. [[CrossRef](#)]
10. Hou, C.; Ai, J.; Lin, Y.; Guan, C.; Li, J.; Zhu, W. Evaluation of Online Teaching Quality Based on Facial Expression Recognition. *Future Internet* **2022**, *14*, 177. [[CrossRef](#)]
11. Shi, C.; Tan, C.; Wang, L. A facial expression recognition method based on a multibranch cross-connection convolutional neural network. *IEEE Access* **2021**, *9*, 39255–39274. [[CrossRef](#)]
12. Zhou, N.; Liang, R.; Shi, W. A lightweight convolutional neural network for real-time facial expression detection. *IEEE Access* **2020**, *9*, 5573–5584. [[CrossRef](#)]
13. Kim, J.; Kang, J.K.; Kim, Y. A Resource Efficient Integer-Arithmetic-Only FPGA-Based CNN Accelerator for Real-Time Facial Emotion Recognition. *IEEE Access* **2021**, *9*, 104367–104381.
14. Zhao, G.; Wei, W.; Xie, X.; Fan, S.; Sun, K. An FPGA-Based BNN Real-Time Facial Emotion Recognition Algorithm. In *2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, Dalian, China, 24–26 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 20–24.
15. Saurav, S.; Gidde, P.; Saini, R.; Singh, S. Dual integrated convolutional neural network for real-time facial expression recognition in the wild. *Vis. Comput.* **2022**, *38*, 1083–1096. [[CrossRef](#)]
16. Saurav, S.; Saini, R.; Singh, S. EmNet: A deep integrated convolutional neural network for facial emotion recognition in the wild. *Appl. Intell.* **2021**, *51*, 5543–5570. [[CrossRef](#)]
17. Zhao, G.; Yang, H.; Yu, M. Expression recognition method based on a lightweight convolutional neural network. *IEEE Access* **2020**, *8*, 38528–38537. [[CrossRef](#)]
18. Huang, Z.Y.; Chiang, C.C.; Chen, J.H.; Chen, Y.C.; Chung, H.L.; Cai, Y.P.; Hsu, H.C. A study on computer vision for facial emotion recognition. *Sci. Rep.* **2023**, *13*, 8425. [[CrossRef](#)]
19. Dias, W.; Andaló, F.; Padilha, R.; Bertocco, G.; Almeida, W.; Costa, P.; Rocha, A. Cross-dataset emotion recognition from facial expressions through convolutional neural networks. *J. Vis. Commun. Image Represent.* **2022**, *82*, 103395. [[CrossRef](#)]
20. Liang, L.; Lang, C.; Li, Y.; Feng, S.; Zhao, J. Fine-grained facial expression recognition in the wild. *IEEE Trans. Inf. Forensics Secur.* **2020**, *16*, 482–494. [[CrossRef](#)]
21. Reghunathan, R.K.; Ramankutty, V.K.; Kallingal, A.; Vinod, V. Facial Expression Recognition Using Pre-Trained Architectures. *Eng. Proc.* **2024**, *62*, 22.
22. Gupta, S.; Kumar, P.; Tekchandani, R.K. Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. *Multimed. Tools Appl.* **2023**, *82*, 11365–11394.
23. Ma, Y.; Shen, J.; Zhao, Z.; Liang, H.; Tan, Y.; Liu, Z.; Qian, K.; Yang, M.; Hu, B. What can facial movements reveal? Depression recognition and analysis based on optical flow using Bayesian networks. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 3459–3468.
24. Shangguan, Z.; Liu, Z.; Li, G.; Chen, Q.; Ding, Z.; Hu, B. Dual-stream multiple instance learning for depression detection with facial expression videos. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *31*, 554–563. [[CrossRef](#)]
25. Gue, J.X.; Chong, C.Y.; Lim, M.K. Facial Expression Recognition as markers of Depression. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*; IEEE: Piscataway, NJ, USA, 2023; pp. 674–680.
26. Ramis Guarinos, S.; Manresa Yee, C.; Buades Rubio, J.M.; Gaya-Morey, F.X. Explainable Facial Expression Recognition for People with Intellectual Disabilities. In *Interacción '23: Proceedings of the XXIII International Conference on Human Computer Interaction*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–7.
27. Rathod, M.; Dalvi, C.; Kaur, K.; Patil, S.; Gite, S.; Kamat, P.; Kotecha, K.; Abraham, A.; Gabralla, L.A. Kids' emotion recognition using various deep-learning models with explainable AI. *Sensors* **2022**, *22*, 8066. [[CrossRef](#)] [[PubMed](#)]
28. Ruangdit, T.; Sungkhin, T.; Phenglong, W.; Phaisangittisagul, E. Integration of Facial and Speech Expressions for Multimodal Emotional Recognition. In *TENCON 2023—2023 IEEE Region 10 Conference (TENCON)*, Chiang Mai, Thailand, 31 October–3 November 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 519–523.
29. Hettiarachchi, H.; Ekanayake, M.; Kaveendhya, G.; Koralage, O.; Samarasekara, P.; Kasthurirathna, D. Gender influence on emotional recognition using facial expressions and voice. In *2023 5th International Conference on Advancements in Computing (ICAC)*, Colombo, Sri Lanka, 7–8 December 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 721–726.

30. Cesarelli, M.; Martinelli, F.; Mercaldo, F.; Santone, A. Emotion recognition from facial expression using explainable deep learning. In *2022 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCCom/CyberSciTech)*, Falerna, Italy, 12–15 September 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
31. Kandeel, A.A.; Abbas, H.M.; Hassanein, H.S. Explainable model selection of a convolutional neural network for driver's facial emotion identification. In *Proceedings of the International Conference on Pattern Recognition*; Springer: Cham, Switzerland, 2021; pp. 699–713.
32. Deramgozin, M.; Jovanovic, S.; Rabah, H.; Ramzan, N. A hybrid explainable ai framework applied to global and local facial expression recognition. In *2021 IEEE International Conference on Imaging Systems and Techniques (IST)*, Kaohsiung, Taiwan, 24–26 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–5.
33. Zhu, H.; Yu, C.; Cangelosi, A. Explainable emotion recognition for trustworthy human–robot interaction. In *Proceedings of the Workshop Context-Awareness Human-Robot Interaction Approaches Challenges ACM/IEEE HRI*, Sapporo, Japan, 7–10 March 2022; IEEE Press: Piscataway, NJ, USA, 2022.
34. Borriero, A.; Milazzo, M.; Diano, M.; Orsenigo, D.; Villa, M.C.; DiFazio, C.; Tamietto, M.; Perotti, A. Explainable Emotion Decoding for Human and Computer Vision. In *Proceedings of the World Conference on Explainable Artificial Intelligence*; Springer: Cham, Switzerland, 2024; pp. 178–201.
35. Punuri, S.B.; Kuanar, S.K.; Mishra, T.K.; Rao, V.V.R.M.; Reddy, S.S. Decoding Human Facial Emotions: A Ranking Approach using Explainable AI. *IEEE Access* **2024**, *12*, 186229–186245. [[CrossRef](#)]
36. Manresa-Yee, C.; Ramis, S.; Gaya-Morey, F.X.; Buades, J.M. Impact of explanations for trustworthy and transparent artificial intelligence. In *Interacción '23: Proceedings of the XXIII International Conference on Human Computer Interaction*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 1–8.
37. Rajpal, A.; Sehra, K.; Bagri, R.; Sikka, P. Xai-fr: Explainable AI-based face recognition using deep neural networks. *Wirel. Pers. Commun.* **2023**, *129*, 663–680.
38. Gaya-Morey, F.X.; Ramis-Guarinos, S.; Manresa-Yee, C.; Buades-Rubio, J.M. Unveiling the human-like similarities of automatic facial expression recognition: An empirical exploration through explainable AI. *Multimed. Tools Appl.* **2024**, *83*, 85725–85753. [[CrossRef](#)]
39. Lorch, S.; Gebele, J.; Brune, P. Towards Trustworthy AI: Evaluating SHAP and LIME for Facial Emotion Recognition. In *Proceedings of the 58th Hawaii International Conference on System Sciences*; Available at University of Hawaii Scholar Space; University of Hawaii at Manoa: Honolulu, HI, USA, 2025; pp. 7532–7546.
40. Di Luzio, F.; Rosato, A.; Panella, M. An explainable fast deep neural network for emotion recognition. *Biomed. Signal Process. Control* **2025**, *100*, 107177. [[CrossRef](#)]
41. Jaswanth, M.; Narayana, N.; Rahul, S.; Amudha, J.; Milan K, A. Emotion and Advertising Effectiveness: A Novel Facial Expression Analysis Approach Using Federated Learning. In *2023 IEEE 20th India Council International Conference (INDICON)*, Hyderabad, India, 14–17 December 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 368–373.
42. Zhang, C.; Li, M.; Wu, D. Federated multidomain learning with graph ensemble autoencoder GMM for emotion recognition. *IEEE Trans. Intell. Transp. Syst.* **2022**, *24*, 7631–7641. [[CrossRef](#)]
43. Franco, D.; Oneto, L.; Navarin, N.; Anguita, D. Toward learning trustworthily from data combining privacy, fairness, and explainability: An application to face recognition. *Entropy* **2021**, *23*, 1047. [[CrossRef](#)]
44. Ghosh, T.; Banna, M.H.A.; Nahian, M.J.A.; Kaiser, M.S.; Mahmud, M.; Li, S.; Pillay, N. A privacy-preserving federated-mobilenet for facial expression detection from images. In *Proceedings of the International Conference on Applied Intelligence and Informatics*; Springer: Cham, Switzerland, 2022; pp. 277–292.
45. Simić, N.; Suzić, S.; Milošević, N.; Stanojev, V.; Nosek, T.; Popović, B.; Bajović, D. Enhancing Emotion Recognition through Federated Learning: A Multimodal Approach with Convolutional Neural Networks. *Appl. Sci.* **2024**, *14*, 1325. [[CrossRef](#)]
46. Schmidt, P.; Reiss, A.; Duerichen, R.; Marberger, C.; Van Laerhoven, K. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *ICMI '18: 20th ACM International Conference on Multimodal Interaction*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 400–408.
47. Wijasena, H.Z.; Ferdiana, R.; Wibirama, S. A survey of emotion recognition using physiological signal in wearable devices. In *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, Bandung, Indonesia, 28–30 April 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
48. Treadway, M.T.; Zald, D.H. Reconsidering anhedonia in depression: Lessons from translational neuroscience. *Neurosci. Biobehav. Rev.* **2011**, *35*, 537–555. [[CrossRef](#)]
49. Pizzagalli, D.A. Depression, stress, and anhedonia: Toward a synthesis and integrated model. *Annu. Rev. Clin. Psychol.* **2014**, *10*, 393–423. [[CrossRef](#)]
50. Calvo, M.G.; Nummenmaa, L. Detection of emotional faces: Salient physical features guide effective visual search. *J. Exp. Psychol. Gen.* **2008**, *137*, 471. [[CrossRef](#)]

51. Suess, F.; Rabovsky, M.; Abdel Rahman, R. Perceiving emotions in neutral faces: Expression processing is biased by affective person knowledge. *Soc. Cogn. Affect. Neurosci.* **2015**, *10*, 531–536.
52. Rizvi, S.J.; Pizzagalli, D.A.; Sproule, B.A.; Kennedy, S.H. Assessing anhedonia in depression: Potentials and pitfalls. *Neurosci. Biobehav. Rev.* **2016**, *65*, 21–35. [[CrossRef](#)] [[PubMed](#)]
53. Rottenberg, J. Mood and emotion in major depression. *Curr. Dir. Psychol. Sci.* **2005**, *14*, 167–170. [[CrossRef](#)]
54. Bylsma, L.M.; Morris, B.H.; Rottenberg, J. A meta-analysis of emotional reactivity in major depressive disorder. *Clin. Psychol. Rev.* **2008**, *28*, 676–691. [[CrossRef](#)] [[PubMed](#)]
55. Fredrickson, B.L.; Losada, M.F. Positive affect and the complex dynamics of human flourishing. *Am. Psychol.* **2005**, *60*, 678–686. [[CrossRef](#)] [[PubMed](#)]
56. Martin, L.A.; Neighbors, H.W.; Griffith, D.M. The experience of symptoms of depression in men vs women: Analysis of the National Comorbidity Survey Replication. *JAMA Psychiatry* **2013**, *70*, 1100–1106. [[CrossRef](#)]
57. Busch, F.N. Anger and depression. *Adv. Psychiatr. Treat.* **2009**, *15*, 271–278. [[CrossRef](#)]
58. Li, S.; Zhang, M.; Liu, Z.; Chen, W.; Yang, Y.; Jiang, Z. RAF-DB: A Real-world Affective Faces Database. In *Proceedings of the 2017 International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017*; IEEE: Piscataway, NJ, USA, 2017; pp. 51–57.
59. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*; Springer: Cham, Switzerland, 2013; pp. 117–124.
60. Zhang, Z.; Luo, P.; Loy, C.C.; Tang, X. From facial expression recognition to interpersonal relation prediction. *Int. J. Comput. Vis.* **2018**, *126*, 550–569.
61. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2018; pp. 4510–4520. [[CrossRef](#)]
62. Zhang, X.; Zhou, X.; Qiao, Y. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2018; pp. 6848–6856. [[CrossRef](#)]
63. Shehada, D.; Tawfik, H.; Bouridane, A.; Hussain, A. An Explainable Framework for Mental Health Monitoring Using Lightweight and Privacy-Preserving Federated Facial Emotion Recognition. *Sensors* **2025**, *25*, 7320. [[CrossRef](#)]
64. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 839–847.
65. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning*; PMIR: Brookline, MA, USA, 2017; pp. 3145–3153.
66. Petsiuk, V.; Das, A.; Saenko, K. Rise: Randomized input sampling for explanation of black-box models. *arXiv* **2018**, arXiv:1806.07421. [[CrossRef](#)]
67. Gomez, T.; Fréour, T.; Mouchère, H. Metrics for saliency map evaluation of deep learning explanation methods. In *Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence*; Springer: Cham, Switzerland, 2022; pp. 84–95.
68. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for explainable AI: Challenges and prospects. *arXiv* **2018**, arXiv:1812.04608.
69. Liao, Q.V.; Gruen, D.; Miller, S. Questioning the AI: Informing design practices for explainable AI user experiences. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–15.
70. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608. [[CrossRef](#)]
71. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [[CrossRef](#)]
72. Barsoum, E.; Zhang, C.; Ferrer, C.C.; Zhang, Z. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ICMI '16: Proceedings of the 18th ACM International Conference on Multimodal Interaction*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 279–283.
73. Li, S.; Deng, W. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**, *13*, 1195–1215. [[CrossRef](#)]
74. Mollahosseini, A.; Hasani, B.; Mahoor, M.H. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **2017**, *10*, 18–31. [[CrossRef](#)]
75. Chaddad, A.; Peng, J.; Xu, J.; Bouridane, A. Survey of explainable AI techniques in healthcare. *Sensors* **2023**, *23*, 634. [[CrossRef](#)] [[PubMed](#)]
76. Saarela, M.; Podgorelec, V. Recent applications of Explainable AI (XAI): A systematic literature review. *Appl. Sci.* **2024**, *14*, 8884. [[CrossRef](#)]

77. Pawlicki, M.; Pawlicka, A.; Uccello, F.; Szelest, S.; D'Antonio, S.; Kozik, R.; Choraś, M. Evaluating the necessity of the multiple metrics for assessing explainable AI: A critical examination. *Neurocomputing* **2024**, *602*, 128282. [\[CrossRef\]](#)
78. Yu, H.; Benois-Pineau, J.; Bourqui, R.; Giot, R.; Zhukov, A. Mean Opinion Score as a New Metric for User-Evaluation of XAI Methods. In *Proceedings of the International Conference on Pattern Recognition*; Springer: Cham, Switzerland, 2024; pp. 443–457.
79. Kondaveeti, H.K.; Simhadri, C.G. Evaluation of deep learning models using explainable AI with qualitative and quantitative analysis for rice leaf disease detection. *Sci. Rep.* **2025**, *15*, 31850. [\[CrossRef\]](#)
80. Hedström, A.; Weber, L.; Bareeva, D.; Krakowczyk, D.; Motzkus, F.; Samek, W.; Lapuschkin, S.; Höhne, M.M.C. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *arXiv* **2023**, arXiv:2202.06861. [\[CrossRef\]](#)
81. Gao, Y.; Cai, Y.; Bi, X.; Li, B.; Li, S.; Zheng, W. Cross-domain facial expression recognition through reliable global–local representation learning and dynamic label weighting. *Electronics* **2023**, *12*, 4553. [\[CrossRef\]](#)
82. Xu, R.; Li, G.; Yang, J.; Lin, L. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2019; pp. 1426–1435.
83. Long, M.; Cao, Z.; Wang, J.; Jordan, M.I. Conditional adversarial domain adaptation. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1647–1657.
84. Li, S.; Deng, W. A deeper look at facial expression dataset bias. *IEEE Trans. Affect. Comput.* **2020**, *13*, 881–893. [\[CrossRef\]](#)
85. Chen, T.; Pu, T.; Wu, H.; Xie, Y.; Liu, L.; Lin, L. Cross-Domain Facial Expression Recognition: A Unified Evaluation Benchmark and Adversarial Graph Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 9887–9903. [\[CrossRef\]](#)
86. Xie, Y.; Gao, Y.; Lin, J.; Chen, T. Learning consistent global-local representation for cross-domain facial expression recognition. In *Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022*; IEEE: Piscataway, NJ, USA, 2022; pp. 2489–2495.
87. Yang, J.; Qian, T.; Zhang, F.; Khan, S.U. Real-Time Facial Expression Recognition Based on Edge Computing. *IEEE Access* **2021**, *9*, 76178–76190. [\[CrossRef\]](#)

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.